End User Documentation **CUSTOMER**

Document Version: 1.0 - 2014-12

SAP InfiniteInsight® 7.0 SP1

Guide utilisateur

Table of Contents

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

1 Bienvenue dans ce guide

1.1 A propos de ce document

1.1.1 A qui s'adresse ce document

Ce document s'adresse aux personnes qui souhaitent **évaluer** ou **utiliser** SAP InfiniteInsight®.

1.1.2 Prérequis à la lecture de ce document

La lecture de ce guide ne nécessite aucune connaissance préalable, y compris en statistiques ou en bases de données.

Les fonctionnalités SAP InfiniteInsight® reposent sur des technologies pointues et utilisent des techniques statistiques complexes et novatrices. En même temps, elles sont simples et rapides à utiliser : elles mettent de puissantes techniques de Data Mining à la portée de tout "utilisateur métier".

Pour obtenir des informations plus techniques sur SAP InfiniteInsight®, consultez nos White Papers.

1.1.3 Objet de ce document

Ce document est le guide de prise de main des deux fonctionnalités SAP InfiniteInsight® décrites dans le tableau suivant.

La fonctionnalité...	Vous permet de...	Exemple...
InfiniteInsight® Modeler / Régression ou Classement	comprendre et prédire un phénomène	Vous travaillez pour un constructeur automobile et souhaitez envoyer un courrier publicitaire à vos prospects. InfiniteInsight® Modeler / Régression ou Classement vous permet de : ▪ comprendre les raisons pour lesquelles d'anciens prospects ont déjà répondu à un tel courrier, ▪ prédire le taux de réponses à un tel courrier envoyé à de vos nouveaux prospects.
InfiniteInsight® Modeler / Segmentation	décrire un jeu de données, en le décomposant en groupes de données homogènes, ou segments	Votre société commercialise deux produits A et B. InfiniteInsight® Modeler / Segmentation vous permet de : ▪ regrouper vos clients en plusieurs groupes homogènes, ▪ connaître le comportement de chacun de ces groupes par rapport aux produits A et B.

Ce document vous présente les notions fondamentales relatives à SAP InfiniteInsight®, ainsi que les principales fonctionnalités des composants InfiniteInsight® Modeler / Régression ou Classement et InfiniteInsight® Modeler / Segmentation. Grâce à deux scénarios d'utilisation, il vous permet de prendre rapidement en main les fonctionnalités SAP InfiniteInsight® présentées et de créer vos premiers modèles avec la plus grande facilité.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

1.1.4 Comment utiliser ce document

Organisation de ce document

Ce document se subdivise en six chapitres.

Le présent chapitre, **Bienvenue dans ce guide**, fait fonction d'introduction au reste du guide. Vous y trouvez des informations concernant la lecture de ce guide et des informations vous permettant de nous contacter.

Le chapitre 2, **SAP InfiniteInsight®**, donne une vue d'ensemble de la plate-forme analytique, de son architecture et de son fonctionnement. Il présente également deux prérequis méthodologiques indispensables à l'utilisation des fonctionnalités de SAP InfiniteInsight®.

Le chapitre 3, **Notions fondamentales**, présente les notions fondamentales relatives à la modélisation de données avec SAP InfiniteInsight®.

Le bref chapitre 4, **Présentation générale des scénarios**, donne un résumé des scénarios d'utilisation des fonctionnalités InfiniteInsight® Modeler / Régression ou Classement et InfiniteInsight® Modeler / Segmentation. Il présente également l'interface d'utilisation et les fichiers de données utilisés pour ces scénarios.

Les chapitre 5 et 6, **Générer des modèles explicatifs et prédictifs avec InfiniteInsight® Modeler / Régression ou Classement** et **Générer des modèles descriptifs avec InfiniteInsight® Modeler / Segmentation**, présentent respectivement les fonctionnalités de régression / classement et et segmentation. Ces deux chapitres sont organisés de la même manière, en deux parties :

- la première partie présente un scénario d'utilisation détaillée de la fonctionnalité,
- la deuxième partie présente l'utilisation proprement dite de la fonctionnalité, sur la base du scénario d'utilisation correspondant.

Un sommaire et une table des matières détaillée situés au début de guide et un système de renvois vous permettent de trouver rapidement l'information que vous cherchez.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Que devez-vous lire ?

En fonction de votre profil et de vos besoins, vous pouvez souhaiter lire la totalité de ce guide ou seulement certaines parties. Dans tous les cas, il est essentiel que vous lisiez la partie sur les indicateurs de performance (voir à la page 40) SAP InfiniteInsight®. Ces indicateurs constituent l'une des notions les plus importantes de SAP InfiniteInsight®. Ils permettent d'évaluer la qualité et la robustesse des modèles générés à partir de SAP InfiniteInsight®.

Le tableau suivant donne quelques repères visant à faciliter votre utilisation de ce guide.

Quel est votre profil ?	**Comment pouvez-vous utiliser au mieux ce guide ?**
Vous souhaitez **évaluer** SAP InfiniteInsight® et votre temps est compté	Vous pouvez vous contenter de : 1. Lire le scénario de la fonctionalité qui vous intéresse (ou tout du moins le résumé de ce scénario) : ▪ **Scénario d'utilisation de InfiniteInsight® Modeler / Régression ou Classement** (voir "**Scénario d'utilisation : Gagnez en efficacité et maîtrisez votre budget grâce à la modélisation**" à la page 54) ▪ **Scénario d'utilisation de InfiniteInsight® Modeler / Segmentation** (voir "**Scénario d'utilisation : Personnalisez votre communication grâce à la modélisation de données**" à la page 186) 2. Passer directement à la partie "Utiliser la fonctionnalité" correspondante : ▪ **Utiliser la fonctionnalité InfiniteInsight® Modeler / Régression ou Classement** (voir à la page 66) ▪ **Utiliser la fonctionnalité InfiniteInsight® Modeler / Segmentation** (voir à la page 194)
▪ Vous souhaitez être **guidé pas à pas** au travers de la découverte de SAP InfiniteInsight® ▪ Vous n'avez qu'une **expérience légère** en modélisation de données	Lisez au moins une fois ce guide de manière linéaire, c'est-à-dire en lisant les chapitres dans l'ordre dans lequel ils vous sont présentés. Dans tous les cas, assurez-vous que vous possédez une bonne connaissance des notions fondamentales relatives à l'utilisation de SAP InfiniteInsight® en consultant le **chapitre 3, Notions fondamentales. Ces notions sont essentielles autant pour l'utilisation des fonctionnalités SAP InfiniteInsight® que pour l'analyse des résultats obtenus.** (voir à la page 17)
Vous avez une **bonne expérience** en modélisation de données	Vous pouvez vous contenter de : 1. Vérifier que la terminologie utilisée par SAP InfiniteInsight® vous est familière, par exemple en consultant le contenu du **chapitre Notions fondamentales, dans la table des matières détaillées.** (voir à la page 17) 2. Lire le résumé du scénario de la fonctionnalité qui vous intéresse. ▪ **Résumé du scénario d'utilisation de InfiniteInsight® Modeler / Régression ou Classement** à la page 54 ▪ **Résumé du scénario d'utilisation de InfiniteInsight® Modeler / Segmentation** à la page 186 3. Passer directement à la partie Utiliser la fonctionnalité. ▪ **Utiliser la fonctionnalité InfiniteInsight® Modeler / Régression ou Classement** (voir à la page 66) ▪ **Utiliser la fonctionnalité InfiniteInsight® Modeler / Segmentation** (voir à la page 194)
▪ Vous **avez déjà suivi une formation** à SAP InfiniteInsight® ▪ Vous êtes **déjà utilisateur** de SAP InfiniteInsight®	Vous pouvez : ▪ suivre les scénarios d'utilisation pour une "reprise" en main des fonctionnalités qui vous intéressent. • **Scénario d'utilisation de InfiniteInsight® Modeler / Régression ou Classement** (voir "**Scénario d'utilisation : Gagnez en efficacité et maîtrisez votre budget grâce à la modélisation**" à la page 54) • **Scénario d'utilisation de InfiniteInsight® Modeler / Segmentation** (voir "**Scénario d'utilisation : Personnalisez votre communication grâce à la modélisation de données**" à la page 186) ▪ utiliser ce document comme un guide de référence, en le consultant de manière ponctuelle. Dans ce cas, la table des matières détaillées et l'index vous seront d'une aide précieuse pour trouver l'information que vous cherchez.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

1.2 Avant de commencer

1.2.1 Fichiers et documentations livrés avec ce guide

Fichiers de données exemples

SAP InfiniteInsight® est livré avec des fichiers de données exemples. Ces fichiers vous permettent d'évaluer et de faire vos premiers pas avec les différents fonctionnalités de SAP InfiniteInsight®.

Lors de l'installation de SAP InfiniteInsight®, ces fichiers sont enregistrés dans les sous-répertoires du répertoire suivant : `C:\Program Files\SAP InfiniteInsight\InfiniteInsightVx.y.z\\Samples\`.

Le tableau suivant décrit ces fichiers.

Nom du fichier	Description	Quand l'utilisez-vous ?
Census01.csv	Fichiers de données	Ce fichier est utilisé pour les scénarios d'utilisation de **InfiniteInsight® Modeler / Régression ou Classement** (modèles explicatifs et prédictifs) et de **InfiniteInsight® Modeler / Segmentation** (modèles descriptifs)
desc_census01.csv	Fichier de description du fichier `Census01.csv`	Ce fichier est utilisé pour les scénarios d'utilisation de **InfiniteInsight® Modeler / Régression ou Classement** (modèles explicatifs et prédictifs) et de **InfiniteInsight® Modeler / Segmentation** (modèles descriptifs)

Pour obtenir une description détaillée du fichier `Census01.csv`, voir Présentation des fichiers exemples (voir à la page 63).

Documentation

Documentation complète

Une documentation complète est fournie avec SAP InfiniteInsight®. Cette documentation porte sur :

- l'utilisation fonctionnelle des modules SAP InfiniteInsight®,
- l'architecture et l'intégration de l'API SAP InfiniteInsight®,
- l'interface utilisateur graphique Java *KxJWizard* et l'interpréteur de commandes *KxShell*, livrés en code source.

☑ Pour accéder à la documentation complète

1. Sélectionnez *Démarrer* > *Programmes* > *SAP Business Intelligence* > *SAP InfiniteInsight®* > *Documentation*. La page *Welcome to SAP InfiniteInsight®* apparaît.
2. Sur cette page, cliquez sur la documentation qui vous intéresse.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Aide contextuelle

Chaque panneau de l'assistant de modélisation est accompagné d'une aide contextuelle, décrivant les options présentées et les concepts nécessaires à leur utilisation.

☑ Pour accéder à l'aide contextuelle de l'assistant de modélisation

Sur le panneau pour lequel vous avez besoin d'aide, cliquez sur le bouton *Aide*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 SAP InfiniteInsight®

DANS CE CHAPITRE

2.1 Présentation

SAP InfiniteInsight® est la solution de Data Mining idéale pour modéliser vos données en toute simplicité et avec la plus grande rapidité, tout en obtenant des résultats pertinents et facilement interprétables. Grâce à SAP InfiniteInsight®, vous transformez rapidement vos données en connaissance et prenez les bonnes décisions stratégiques et opérationnelles au bon moment.

SAP InfiniteInsight® met les dernières techniques de Data Mining à la portée de n'importe quel utilisateur métier. SAP InfiniteInsight® vous permet d'accéder facilement à de nombreux formats de sources de données et de générer de manière semi-automatique et extrêmement rapide des modèles explicatifs et prédictifs et des modèles descriptifs.

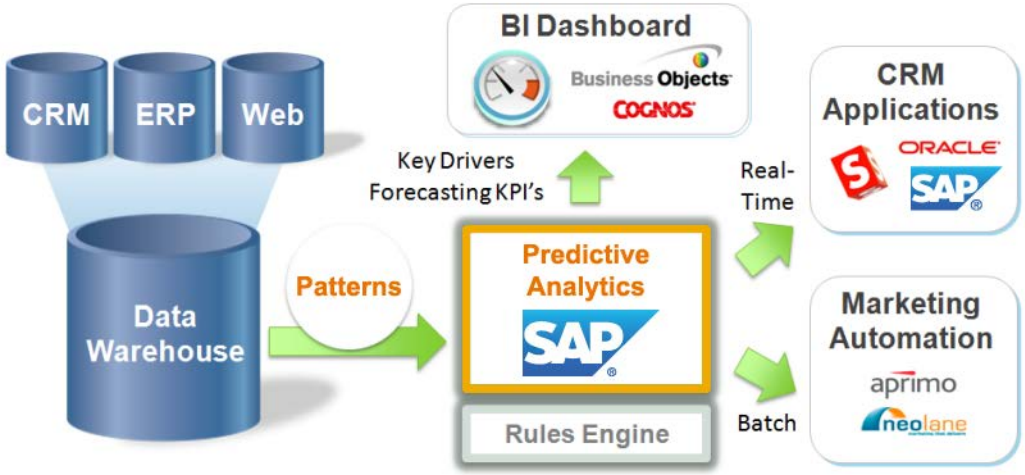
Avec SAP InfiniteInsight®, vous pouvez vous concentrer sur les activités à forte valeur ajoutée que constituent l'analyse des résultats de la modélisation de vos données et la prise de décision.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2.2 Architecture et fonctionnement

En se basant sur un schéma d'architecture général présenté ci-dessous, cette section présente :

- les différents types d'**interfaces** vous permettant d'utiliser SAP InfiniteInsight®.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2.2.1 Interfaces d'utilisation

Les trois types d'interface d'utilisation

Trois types d'interfaces vous permettent d'utiliser les fonctionnalités de SAP InfiniteInsight® :

- **une interface graphique utilisateur**,
- **un interpréteur de commandes**,
- **des API de contrôle** (*Application Programming Interface*).

L'interface graphique

L'interface **KxJWizard** s'adresse principalement aux "utilisateurs finaux" ou "métier". Elle donne accès à des assistants de modélisation, qui vous permettent d'utiliser les fonctionnalités SAP InfiniteInsight® et de modéliser vos données avec la plus grande facilité. En même temps, elle propose un ensemble de graphiques facilitant la visualisation et l'interprétation des résultats de la modélisation.

Cette interface, fournie à titre d'exemple, est développée en **Java** sur la base de l'API CORBA et fonctionne sur n'importe quelle plate-forme (Windows, UNIX, etc.). Grâce aux API proposées avec SAP InfiniteInsight®, vous pouvez développer vos propres interfaces graphiques.

L'interpréteur de commande KxShell

L'interpréteur **KxShell** vous permet d'utiliser SAP InfiniteInsight® à l'aide de commandes. Un script KxShell transmet les commandes d'une modélisation aux différentes fonctionnalités.

L'interpréteur de commandes est un exemple de développement basé sur l'API C++. Comme une API, il peut être utilisé pour intégrer SAP InfiniteInsight® à d'autres applications ou progiciels.

Les API de contrôle

Les **API de contrôle** (*Application Programming Interface*) s'adressent principalement aux développeurs, ou aux utilisateurs ayant une pratique de la programmation. Ces API donnent accès à tout l'éventail des fonctionnalités et aux paramétrages les plus fins des fonctionnalités SAP InfiniteInsight®. En même temps, elles permettent d'intégrer les fonctionnalités SAP InfiniteInsight® de manière personnalisée à d'autres applications ou progiciels.

Trois API sont livrées avec SAP InfiniteInsight® :

- une API **COM/DCOM**, utilisable sur les plates-formes Microsoft,
- une API **CORBA**, utilisable sur toute plate-forme en mode Client/Serveur,
- une API **C++**, utilisable sur toute plate-forme en mode standalone.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2.2.2 Fonctionnement

Le fonctionnement de SAP InfiniteInsight® peut être subdivisé en quatre phases :

- Phase 1 - **Accès aux données**
- Phase 2 - **Manipulation et préparation des données**
- Phase 3 - **Modélisation des données**
- Phase 4 - **Présentation et déploiement des modèles**

Phase 1 : Accès aux données

SAP InfiniteInsight® accède à divers types de sources de données :

- des fichiers "à plat", tels que les fichiers .csv, les fichiers tabulés et autres fichiers de type texte.
- des sources compatibles ODBC, telles que les bases de données Oracle, SQL Server ou IBM DB2.

L' **API C Data Access** permet de connecter des sources au format propriétaire, telles que des senseurs industriels.

Dans la majorité des cas, et notamment si vous utilisez les fonctionnalités SAP InfiniteInsight® via une interface graphique, vous n'avez pas à vous préoccuper des processus d'accès aux données. L'accès aux données est réalisé de manière quasi-transparente : dans l'interface utilisateur graphique, il vous suffit de sélectionner le format de source de données à utiliser (fichiers "à plat" ou sources de données compatibles ODBC) et de spécifier la localisation du fichier de données. L' API C Data Access est utile pour les développeurs qui souhaitent écrire des accès à des bases de données au format propriétaire.

La fonctionnalité InfiniteInsight® Access

La fonctionnalité I**nfiniteInsight® Access** (KAA) permet la lecture des données SAS et l'écriture dans une table SAS des scores obtenus par un modèle SAP InfiniteInsight®.

Actuellement, les formats gérés sont les suivants :

- fichiers SAS version 6 sous Windows & Unix
- SAS 7/8 sous Windows & Unix
- Fichiers SAS Transport

L'accès à une table SAS se fait directement grâce à l'interface SAP InfiniteInsight® en indiquant simplement le type du format du fichier à analyser. La génération d'une table SAS contenant les résultats de l'application d'un modèle SAP InfiniteInsight® (scores, probabilités, numéro du segment, valeur prévue...) se fera de la même en façon, en indiquant le format de la table en sortie grâce à l'interface SAP InfiniteInsight®. La table ainsi générée au format SAS est intégrée automatiquement dans le système d'information SAS.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Phase 2 : Manipulation et préparation des données

InfiniteInsight® Explorer / Codeur de séquences (KSC) et InfiniteInsight® Explorer / Codeur des journaux d'événements (KEL) sont des fonctionnalités de préparation et de manipulation de données. L'utilisation de ces fonctionnalités est simple pour l'utilisateur final et les traitements sont effectués de manière automatique.

InfiniteInsight® Explorer / Codeur des journaux d'événements (KEL) rassemble des évènements par période de temps. Il permet d'intégrer des données transactionnelles aux données démographiques des consommateurs. Il est utilisé dans le cas où les données brutes contiennent des informations statiques telles que l'âge, le sexe ou la profession d'une personne, et des variables dynamiques, telles que les habitudes de consommation ou les transactions de cartes bancaires.Les données sont automatiquement regroupées dans la période définie par l'utilisateur sans avoir à programmer en SQL ou à modifier les diagrammes de bases de données. InfiniteInsight® Explorer / Codeur des journaux d'événements combine et compresse ces données pour les rendre utilisables par les autres composants de SAP InfiniteInsight®.

InfiniteInsight® Explorer / Codeur de séquences (KSC) regroupe des évènements en une succession de transitions. Par exemple, le parcours d'un internaute dans un site web lors d'une session peut être transformé en un ensemble de données. Chaque colonne représente une transition particulière d'une page vers une autre. Comme pour InfiniteInsight® Explorer / Codeur des journaux d'événements, ces nouvelles colonnes de données peuvent être ajoutées aux données existantes d'un consommateur et sont rendues exploitables pour les autres composants de SAP InfiniteInsight®.

InfiniteInsight® Modeler / Codeur analytique (K2C) prépare et transforme automatiquement les données en un format approprié à l'utilisation de SAP InfiniteInsight®. InfiniteInsight® Modeler / Codeur analytique traduit les variables nominales et ordinales, remplit automatiquement les valeurs manquantes et détecte les données aberrantes. De plus, cette fonctionnalité contribue de façon significative à la robustesse des modèles générés par SAP InfiniteInsight® en créant un codage robuste des données.

Phase 3 : Modélisation des données

Les fonctionnalités InfiniteInsight® Modeler / Régression ou Classement et InfiniteInsight® Modeler / Segmentation, grâce aux techniques statistiques et aux technologies informatiques sur lesquelles elles reposent, permettent de générer des modèles d'analyse pertinents et robustes.

InfiniteInsight® Modeler / Régression ou Classement permet de générer des modèles explicatifs et prédictifs. Les modèles générés par InfiniteInsight® Modeler / Régression ou Classement permettent d'expliquer et de prédire un phénomène, ou variable cible, en fonction de données contenues dans le jeu de données analysé, ou variables explicatives. Les modèles générés par InfiniteInsight® Modeler / Régression ou Classement sont calculés grâce à un algorithme de régression et de classification. Cette régression polynomiale est un algorithme propriétaire développé et implémenté par KXEN où les calculs des paramètres se base sur le principe des SRM de Vapnik

InfiniteInsight® Modeler / Segmentation permet de générer des modèles descriptifs, c'est-à-dire de segmenter un jeu de données en un nombre de segments (ou groupes). InfiniteInsight® Modeler / Segmentation permet en outre de réaliser des segmentations supervisées grâce à l'introduction d'une variable cible prise en compte dans le codage des données. Une segmentation supervisée permet la constitution de groupes homogènes qui se distinguent entre eux par leur comportement vis à vis de la variable cible. Cette segmentation utilise une méthode optimisée et robustifiée de nuées dynamiques basée (K-means) sur les théories de Vapnik.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Phase 4 : Présentation et déploiement du modèle

Une fois les modèles générés, des indicateurs de performance des modèles, des graphiques et des rapports d'analyse au format HTML facilitent la visualisation et l'interprétation des résultats de la modélisation des données.

Une fois les modèles validés, vous pouvez les appliquer sur :

- une ou plusieurs observations spécifiques issues de votre base de données (mode *Simulation)*,
- une nouveau jeu de données complet, ou jeu de données d'application (mode *Application*).

Pour faciliter le déploiement et l'intégration des modèles, le code correspondant à chaque modèle peut également être généré dans différents langages de programmation. La fonctionnalité InfiniteInsight® Scorer, responsable de cette génération de code, est décrite ci-dessous.

La fonctionnalité InfiniteInsight® Scorer

La fonctionnalité **InfiniteInsight® Scorer** permet de générer le code correspondant à un modèle généré avec SAP InfiniteInsight® dans les langages suivants : C, XML, AWK, HTML, SQL, PMML2, SAS, or JAVA.

Sous cette forme, le modèle peut être intégré dans une application supportant les langages cités ci-dessus.

Les codes générés permettent d'intégrer les modèles SAP InfiniteInsight® au sein d'applications ou progiciels, ou de les appliquer sur des données sans nécessiter la présence de SAP InfiniteInsight®. Ils permettent notamment d'utiliser les modèles sur des plateformes techniques différentes de celle sur laquelle ils ont été générés.

Attention

La génération de code n'est disponible que pour des modèles générés par les fonctionnalités suivantes : **InfiniteInsight® Modeler / Codeur analytique**, **InfiniteInsight® Modeler / Régression ou Classement**, **InfiniteInsight® Modeler / Segmentation**.

2.3 Prérequis méthodologiques

Avant de modéliser vos données avec SAP InfiniteInsight®, vous devez :

- avoir défini une problématique à laquelle vous souhaitez répondre,
- posséder un jeu de données exposant cette problématique sous la forme d'un ensemble d'observations.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2.3.1 Vos données sont-elles exploitables

Une fois votre problématique identifiée et formulée, vous avez besoin de posséder des données qui permettent d'y répondre. Nous ne nous étendrons pas ici sur la notion de valeur informative associée aux données. Celle-ci dépend de vos processus et outils de collecte et d'extraction de données, et non des fonctionnalités SAP InfiniteInsight®. En revanche, pour que vos données soient exploitables par SAP InfiniteInsight®, les cinq conditions suivantes doivent être remplies :

- vous devez posséder un **volume de données suffisamment important** pour pouvoir construire un modèle valide, c'est-à-dire à la fois pertinent et robuste. Un modèle d'analyse qui serait généré à partir d'un jeu de données de 50 lignes aurait une capacité de généralisation faible, ainsi qu'une valeur informative faible, voire dangereuse. Nous pouvons vous conseiller sur les problématiques de volume de données.
- votre jeu de données doit contenir une **variable cible**, qui permette d'exprimer votre problématique au sein de SAP InfiniteInsight®.
- **pour chaque observation du jeu de données** d'apprentissage, **la variable cible doit être renseignée**. Autrement formulé, aucune valeur de la variable cible ne doit manquer sur la totalité du jeu de données d'apprentissage,
- le **format de votre source de données** doit être supporté par SAP InfiniteInsight®,
- vos données doivent être présentées sous la forme d'**une table de données unique**, sauf dans les cas où vous utilisez les fonctionnalités InfiniteInsight® Explorer / Codeur des journaux d'événements ou InfiniteInsight® Explorer / Codeur de séquences.

2.3.2 Quelle est votre problématique

Les fonctionnalités SAP InfiniteInsight® répondent tous à une même philosophie : ils permettent de faire de l'**analyse de données supervisée**. Le terme "supervisé" signifie que l'analyse de données ne se déroule pas dans l'absolu, mais toujours en fonction d'une problématique : votre problématique !

Pensez à la base de données comportant des informations sur vos clients. Une analyse qui aurait regroupé vos clients en groupes homogènes dans l'absolu n'a pas forcément un intérêt évident. En revanche, une analyse qui les aurait regroupé en fonction d'une problématique telle que le "chiffre d'affaire moyen qu'ils vous rapportent chaque année" prendrait toute sa valeur. Vous connaîtriez alors les profils caractéristiques des clients qui vous rapportent le plus d'argent.

Vous l'avez compris, l'étape préalable à l'utilisation SAP InfiniteInsight® consiste à **identifier et formuler votre problématique**.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Notions fondamentales

Cette section présente les notions fondamentales relatives à l'utilisation de SAP InfiniteInsight®.

Toutes ces notions sont présentées et mises en gras dans la section **Vue d'ensemble de SAP InfiniteInsight®**, qui décrit de manière générale le processus de génération d'un modèle à l'aide de SAP InfiniteInsight®.

DANS CE CHAPITRE

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.1 Fonctionnement de SAP InfiniteInsight® : Vue d'ensemble

SAP InfiniteInsight® vous permet de faire du Data Mining supervisé, c'est-à-dire de transformer vos données en connaissances, puis en action, en fonction d'une problématique métier.

SAP InfiniteInsight® supporte différents formats de source données (fichiers "à plat", sources compatibles ODBC, ...). Pour être exploitables par les fonctionnalités SAP InfiniteInsight®, les jeux de données à analyser doivent être présentés sous la forme d'une table de données (voir à la page 26) unique, sauf dans les cas où vous utilisez les fonctionnalités InfiniteInsight® Explorer / Codeur des journaux d'événements ou InfiniteInsight® Explorer / Codeur de séquences.

Pour utiliser les fonctionnalités SAP InfiniteInsight®, vous devez obligatoirement posséder **un jeu de données d'apprentissage**, contenant une variable cible dont toutes les valeurs sont renseignées. Vous pouvez ensuite appliquer le modèle généré à partir du jeu de données d'apprentissage sur **un ou plusieurs jeux de données d'application**.

Le jeu de données d'apprentissage est découpé en trois sous-jeux de données d'estimation, de validation et de test, grâce à une stratégie de découpage (voir à la page 20).

Les différents types de variables (voir à la page 27) continues, ordinales et nominales sont ensuite codés par l'encodeur analytique d'SAP InfiniteInsight®, et les fonctionnalités InfiniteInsight® Explorer / Codeur de séquences et InfiniteInsight® Explorer / Codeur des journaux d'événements dans le cas de données dynamiques. Avant de générer le modèle, vous devez :

- décrire les données. Un utilitaire intégré à SAP InfiniteInsight® permet de générer automatiquement une description du jeu de données à analyser. Vous devez valider cette description, en vérifiant si le type et le format de stockage de chaque variable a été correctement identifié.
- définir le rôle des variables contenues dans le jeu de données à analyser. Vous sélectionnez au moins une variable Y comme variable cible, ou variable qui correspond à votre problématique. Les autres variables de la table de données sont dîtes variables explicatives : elles permettent de calculer la valeur de la variable cible dans un contexte donné. Elles peuvent également être utilisées comme variables de poids.

Pour plus d'informations sur le rôle des fonctionnalités, rendez-vous dans la section Fonctionnement à la page 13.

Vous générez ensuite des modèles (voir à la page 35), capables soit d'expliquer et de prédire un phénomène, soit de décrire un jeu de données, dans les deux cas en fonction de la variable cible précédemment définie. Cette phase est appelée phase d'apprentissage.

Une fois les modèles générés, vous pouvez visualiser et interpréter leur pertinence et leur robustesse grâce :

- aux indicateurs de performance (voir à la page 40) : la capacité prédictive et la reproductibilité,
- différents graphiques, dont le graphique de la courbe de profit.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.2 Sources de données supportées

En standard, les fonctionnalités SAP InfiniteInsight® supportent les sources de données suivantes :

- les **fichiers "plats"** (*flat files*) **dont les données sont séparées par un élément séparateur**, tels que les fichiers au format .csv (voir à la page 70) ou les fichiers .txt tabulés. Par exemple, le fichier exemple *Census01.csv*, utilisé pour les scénarios d'utilisation de InfiniteInsight® Modeler / Régression ou Classement et de InfiniteInsight® Modeler / Segmentation, est un fichier .csv.
- les **sources de données compatibles ODBC**.

Selon votre licence, vous pouvez également utiliser des **fichiers SAS**.

Une API permet également d'interfacer les fonctionnalités SAP InfiniteInsight® avec n'importe quelle application (SPSS, Microsoft Excel, etc.), et ainsi d'accéder à n'importe quelle source de données. Une .dll spécifique doit être développée pour chaque nouvelle source.

> **Remarque**
>
> Pour des informations sur le formatage des données, et notamment pour connaître la liste exacte des sources compatibles ODBC supportées, voir le document **Data Modeling Specification**.

3.3 Jeu de données

Pour utiliser les fonctionnalités SAP InfiniteInsight®, vous devez obligatoirement posséder **un jeu de données d'apprentissage**, contenant une variable cible dont toutes les valeurs sont renseignées. Vous pouvez ensuite appliquer le modèle généré à partir du jeu de données d'apprentissage sur **un ou plusieurs jeux de données d'application**.

3.3.1 Jeu de données d'apprentissage

Un **jeu de données d'apprentissage** est un jeu de données utilisé pour la génération d'un modèle. Dans ce jeu, les valeurs de la variable cible (voir à la page 32) - ou variable correspondant à votre problématique - sont connues. En analysant le jeu de données d'apprentissage, les fonctionnalités SAP InfiniteInsight® génèrent un modèle qui permet d'expliquer la variable cible, grâce aux variables explicatives.

Pour permettre la validation du modèle généré, le jeu de données d'apprentissage est découpé en trois sous-jeux grâce à une stratégie de découpage (voir à la page 20).

Le jeu de données d'apprentissage peut correspondre soit à une partie exhaustive de votre base de données, soit à un échantillon extrait de celle-ci. Le choix dépend du type d'étude à réaliser, des outils utilisés et du budget alloué à l'étude.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.3.2 Jeu de données d'application

Un **jeu de données d'application** est un jeu de données sur lequel vous appliquez un modèle. Ce jeu de données contient une variable cible dont vous souhaitez connaître la valeur.

Le modèle appliqué sur un jeu de données d'application a été préalablement généré à partir d'un jeu de données d'apprentissage. Le jeu de données d'application doit contenir exactement les mêmes informations que le jeu de données d'apprentissage correspondant, c'est-à-dire :

- le même nombre de variables,
- les mêmes types de variables,
- le même ordre de présentation pour ces variables.

Attention

Le jeu de données d'application doit contenir une variable cible correspondant à celle du jeu de données d'apprentissage. Cette remarque est valable dans tous les cas, même si les valeurs de cette variable cible ne sont pas renseignées. Quand ces valeurs sont renseignées, elles peuvent servir à détecter d'éventuelles observations déviantes *(outliers)*.

3.4 Stratégies de découpage

3.4.1 Définition

Une **stratégie de découpage** est une technique qui permet de décomposer un jeu de données d'apprentissage en trois sous-jeux distincts :

- un **sous-jeu d'estimation**,
- un **sous-jeu de validation**,
- un **sous-jeu de test**.

Ce découpage permet une **validation croisée** des modèles générés.

Il existe neuf types de stratégies de découpage.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.4.2 Rôles des trois sous-jeux

Le tableau suivant définit le rôle des trois sous-jeux de données obtenus à l'aide des stratégies de découpages.

L'ensemble de données	Est utilisé pour...
estimation	générer différents modèles. Les modèles générés à ce stade sont hypothétiques
validation	sélectionner le meilleur modèle parmi ceux générés à partir du sous-jeu d'estimation, c'est-à-dire celui qui constitue le meilleur compromis entre un modèle ayant une qualité parfaite et un modèle ayant une robustesse parfaite.
test	vérifier la performance du modèle sélectionné sur un nouveau jeu de données.

Pour comprendre le rôle des stratégies de découpage dans le processus de génération d'un modèle, voir le schéma Génération d'un modèle.

3.4.3 Les types de stratégies de découpage

Pour générer vos modèles, vous pouvez utiliser deux types stratégies de découpage :

- la **stratégie de découpage personnalisée**,
- les **stratégies de découpage automatiques**.

La stratégie de découpage personnalisée

Définition

La **stratégie de découpage personnalisée** vous permet de définir vos propres sous-jeux de données. Pour l'utiliser, vous devez préparer au préalable (avant de lancer les fonctionnalités SAP InfiniteInsight®) trois sous-jeux correspondant aux sous-jeux d'estimation, de validation et de test.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comment l'utiliser

Avant de démarrer SAP InfiniteInsight®, découpez votre fichier de données initial en trois fichiers de la taille de votre choix. Par exemple :

- le premier fichier peut contenir les 1500 premières observations ou lignes de votre fichier de données initial,
- le deuxième fichier, ses observations 1501 à 3000,
- le troisième fichier, ses observations 3001 à 5000.

Avertissement

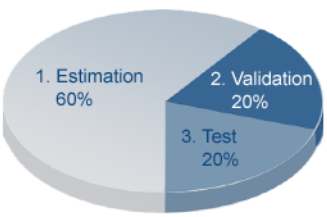
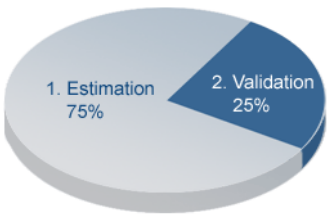
La stratégie de découpage personnalisée peut s'avérer risquée dans le cas d'un fichier initial dans lequel les données sont triées. En effet, les premières lignes ne sont alors plus représentatives de l'ensemble des données contenues dans le fichier initial. Pour éviter ce genre de biais, n'oubliez pas de brasser vos données préalablement à leur analyse.

Les stratégies de découpage automatique

Généralités

A l'exception de la stratégie de découpage personnalisée, les stratégies de découpage sont automatiques. Les stratégies de découpage automatiques travaillent sur un fichier de données unique, que constitue votre jeu de données initial.

Les stratégies de découpage automatiques découpent toujours le jeu de données initial dans les mêmes proportions. Le tableau ci-dessous détaille les proportions attribuées à chaque sous-jeu de données, selon la présence ou non d'un sous-jeu de test.

Stratégies de découpage automatiques avec test	**Stratégies de découpage automatiques sans test**
▪ 3/5 des données sont utilisées dans le sous-jeu d'estimation ▪ 1/5 des données sont utilisées dans le sous-jeu de validation ▪ 1/5 des données sont utilisées dans le sous-jeu de test 	▪ 3/4 des données sont utilisées dans le sous-jeu d'estimation, ▪ 1/4 des données sont utilisées dans le sous-jeu de validation 

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Stratégie de découpage aléatoire

Cette stratégie distribue les données du jeu de données initial de manière aléatoire dans les trois sous-jeux d'estimation, de validation et de test.

Aléatoire avec test à la fin

Cette stratégie distribue :

- les 4/5 du jeu de données initial de manière aléatoire dans les 2 sous-jeux d'estimation et de validation. Cette distribution respecte les proportions habituelles : 3/5 de ces 4/5 sont distribués dans le sous-jeu d'estimation, et 1/5 dans le sous-jeu de validation.
- le dernier 1/5 du jeu de données initial en une fois dans le sous-jeu de test.

Cette stratégie est utile dans le cas où :

- l'alimentation de votre base de données répond à une évolution bien définie, qui détermine un **ordonnancement chronologique des données** dans la base,
- vous souhaitez prendre en compte cet ordonnancement pour la génération du modèle.

Par exemple, imaginez que :

- de nouveaux clients sont référencés tous les mois dans votre base de données,
- vous savez que les jeux de données sur lesquels vous appliquerez le modèle, une fois généré, auront de fortes chances de ressembler à la partie la plus récente de votre base de données, c'est-à-dire celle contenant les derniers clients référencés.

Grâce à la stratégie de découpage aléatoire avec test à la fin, vous testez alors le modèle généré sur la partie de votre base de données qui a le plus de chances de ressembler à l'état de vos futurs jeux de données d'applications.

Aléatoire sans test (stratégie par défaut)

Cette stratégie est la stratégie de découpage proposée par défaut. Elle distribue l'intégralité des données initiales de façon aléatoire entre les sous-jeux d'estimation et de validation.

- 3/4 du jeu de données initial sont attribués au sous-jeu de données d'estimation,
- 1/4 du jeu de données initial est attribué au sous-jeu de données de validation.

Etant donné qu'aucun sous-jeu de données de test n'est utilisé, toutes les données de votre jeu de données d'apprentissage peuvent être utilisées pour les sous-jeux d'estimation et de validation. Ce qui peut permettre d'augmenter la qualité et la robustesse du modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Périodique

Cette stratégie suit le cycle de distribution suivant :

1. Trois lignes du jeu de données initial sont distribuées dans le sous-jeu d'estimation.
2. Une ligne est distribuée dans le sous-jeu de validation.
3. Une ligne est distribuée dans le sous-jeu de test.
4. La distribution reprend à l'étape *1*.

Périodique avec test à la fin

Cette stratégie distribue :

- les 4/5 du jeu de données initial de manière périodique dans les 2 sous-jeux d'estimation et de validation. Cette distribution respecte les proportions habituelles. 3/5 de ces 4/5 sont distribués dans le sous-jeu d'estimation et 1/5 dans le sous-jeu de validation.
- le dernier 1/5 du jeu de données initial d'un bloc dans le sous-jeu de test.

En d'autres mots, la stratégie suit le cycle de distribution suivant :

1. Trois lignes des premiers 4/5 du jeu de données initial sont distribuées dans le sous-jeu d'estimation.
2. Une ligne des premiers 4/5 du jeu de données initial est distribuée dans le sous-jeu de validation.
3. a. Si la totalité des premiers 4/5 du jeu de données initial ne sont pas encore distribués, la distribution reprend à l'étape *1*.
   b. Si la totalité des premiers 4/5 du jeu de données initial sont distribués, la distribution passe à l'étape *4*.
4. Le dernier 1/5 du jeu de données initial est distribué d'un bloc dans le sous-jeu de test.

Périodique sans test

Cette stratégie de découpage distribue l'intégralité du jeu de données initial de façon périodique entre les sous-jeux de données d'estimation et de validation :

- 3/4 du jeu de données initial sont attribués au sous-jeu d'estimation,
- 1/4 du jeu de données initial est attribué au sous-jeu de validation.

En d'autres mots, la stratégie suit le cycle de distribution suivant :

1. Trois lignes du jeu de données initial sont distribuées dans le sous-jeu d'estimation.
2. Une ligne est distribuée dans le sous-jeu de validation.
3. La distribution reprend à l'étape *1*.

Etant donné qu'aucun sous-jeu de données de test n'est utilisé, toutes les données de votre jeu de données d'apprentissage peuvent être utilisées pour les sous-jeux d'estimation et de validation. Ce qui peut permettre d'augmenter la qualité et la robustesse du modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Séquentielle

Cette stratégie découpe le jeu de données initial en trois blocs, correspondant aux proportions de découpage habituelles :

- les lignes correspondant aux premiers 3/5 du jeu de données initial sont distribuées d'un bloc dans le jeu de données d'estimation,
- les lignes correspondant aux 1/5 suivant du jeu de données initial sont distribuées d'un bloc dans le jeu de données de validation,
- les lignes correspondant aux derniers 1/5 du jeu de données initial sont distribuées d'un bloc dans le jeu de données de test.

Séquentielle sans test

Cette stratégie découpe le jeu de données initial en deux blocs, correspondant aux proportions de découpage habituelles lorsqu'il n'y a pas de sous-jeu de test :

- les lignes correspondant aux premiers 3/4 du jeu de données initial sont distribuées d'un bloc dans le jeu de données d'estimation,
- les lignes correspondant au dernier 1/4 du jeu de données initial sont distribuées d'un bloc dans le jeu de données de validation.

Etant donné qu'aucun sous-jeu de données de test n'est utilisé, toutes les données de votre jeu de données d'apprentissage peuvent être utilisées pour les sous-jeux d'estimation et de validation. Ce qui peut permettre d'augmenter la qualité et la robustesse du modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.5 Table de données

3.5.1 Définition

Une **table de données** est un ensemble de données présentées sous la forme d'un tableau à deux dimensions.

Dans cette table :

- chaque **ligne** représente une **observation** à traiter, soit dans le fichier exemple *Census01.csv* un américain.
- chaque **colonne** représente une **variable** qui décrit les observations, soit dans notre exemple "l'âge" ou le "sexe" des individus américains.
- chaque **cellule**, soit l'intersection d'une colonne et d'une ligne, représente la valeur de la variable en colonne pour l'observation en ligne.

Le tableau suivant donne un exemple de table de données.

Observations	Variable 1	Variable 2	Variable 3
Observation a	Valeur a1	Valeur a2	Valeur a3
Observation b	Valeur b1	Valeur b2	Valeur b3
...	...	...	...
Observation n	Valeur n1	Valeur n2	Valeur n3

3.5.2 Synonymes de "observations" et "variables"

Selon votre profil et votre domaine d'expertise, vous pouvez être habitué à employer d'autres termes pour référer aux observations (en lignes) et variables (en colonnes) des tables de données.

Le tableau suivant présente ces termes. Ils sont tous synonymes.

Termes équivalents au terme "Observation"	Termes équivalents au terme "Variable"
Ligne	Colonne
Enregistrement	Attribut
Table	Champ
Evénement	Propriété
Cas	-
Exemple	-

3.5.3 Formatage des données

Quelle que soit la source de données utilisée, les deux contraintes suivantes doivent être respectées :

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

- les données doivent être représentées sous la forme d'**une table**, unique sauf dans les cas où vous utilisez les fonctionnalités InfiniteInsight® Explorer / Codeur des journaux d'événements ou InfiniteInsight® Explorer / Codeur de séquences. .
- la variable cible doit être renseignée pour chaque observation de la table. Dans le fichier exemple *Census01.csv*, la variable "class" a été renseignée pour chaque individu.

Remarque

Pour des informations sur le formatage des données, et notamment pour connaître la liste exacte des sources compatibles ODBC supportées, voir le document **Data Modeling Specification**.

3.6 Variables

3.6.1 Définition générique

Une **variable** correspond à un attribut qui décrit les observations stockées dans votre base de données. Dans les fonctionnalités SAP InfiniteInsight®, une variable est définie par :

- un type,
- un format de stockage,
- un rôle.

Exemple

Dans une base de données contenant des informations sur vos clients, le "nom" et "l'adresse" de ces clients, par exemple, sont des variables.

3.6.2 Types de variables

Il existe trois types de variables :

- les **variables continues**,
- les **variables ordinales**,
- les **variables nominales**.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variables continues

Définition

Les **variables continues** sont des variables dont les valeurs sont **numériques continues** et **ordonnées**. Des opérations arithmétiques peuvent être effectuées sur ces valeurs, telles que la somme ou la moyenne.

Exemple

La variable "Salaire" est une variable numérique. Elle peut prendre les valeurs suivantes : "1200 Euros", "2000 Euros", ou "2035 Euros". Par exemple, la moyenne de ces valeurs peut être calculée.

Variables continues et modélisation

Lors d'une modélisation, une variable continue peut être découpée en tranches significatives.

Variables ordinales

Définition

Les **variables ordinales** sont des variables dont les valeurs sont **discrètes**, c'est-à-dire appartenant à des catégories, et **ordonnées**. Les variables ordinales peuvent être :

- **numériques**, c'est-à-dire avoir pour valeurs des nombres (*number*).. Elles sont alors ordonnées selon l'ordre numérique naturel (0, 1, 2, etc.).
- **textuelles**, c'est-à-dire avoir pour valeurs des chaîne de caractères (*string*). Elles sont alors ordonnées de manière alphabétique.

Exemple

La variable "note scolaire" est une **variable ordinale**. L'ensemble des valeurs que cette variable peut prendre constituent bien des catégories distinctes et ordonnées. Cette variable peut être :

- numérique, si elle prend des valeurs comprises entre "0" et "20",
- textuelle, si elle prend les valeurs A, B, C, D, E et F.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Attention

Une variable "appréciation" ayant pour valeurs "un peu", "beaucoup" et "passionnément" ne peut pas être traitée directement par les fonctionnalités SAP InfiniteInsight® comme si elle était une variable ordinale. L'ordre obtenu serait en effet l'ordre alphabétique ("beaucoup", "passionnément", puis "un peu"), et ne serait plus en phase avec les différents degrés d'appréciation correspondant aux valeurs de cette variable. Quand l'ordre des valeurs d'une variable nominale est important, la variable doit donc être codée, soit en lettres soit en chiffres, avant de pouvoir être utilisée par SAP InfiniteInsight®.

Variables nominales

Définition

Les **variables nominales** sont des variables dont les valeurs sont **discrètes**, c'est-à-dire appartenant à des catégories, et **non ordonnées**.

Les variables nominales peuvent être :

- **numériques**, c'est-à-dire avoir pour valeurs des nombres (*number*).
- **textuelles**, c'est-à-dire avoir pour valeurs des chaînes de caractères (*string*).

Attention

Les **variables binaires** sont considérées comme des variables nominales.

Exemple

La variable "Code postal" est une **variable nominale**. L'ensemble des valeurs que cette variable peut prendre ("36000", "75000", "93000", etc.) constituent bien des catégories distinctes non ordonnées et représentées par des nombres.

La variable "Couleur des yeux" est une **variable nominale.** L'ensemble des valeurs que cette variable peut prendre ("bleu", "marron", "noir", etc.) constituent bien des catégories distinctes non ordonnées et représentées par des chaînes de caractères.

Variables nominales et modélisation

Lors d'une modélisation, les valeurs des variables catégoriques sont regroupées en catégories homogènes. Les catégories sont ensuite ordonnées en fonction de l'importance de leur contribution par rapport aux valeurs de la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.6.3 Formats de stockage

Pour décrire les données, SAP InfiniteInsight® utilise plusieurs types de formats de stockage :

- date,
- datetime (date et horaire),
- number (nombre),
- integer (entier),
- string (chaine de caractères).

Le tableau suivant décrit ces formats de stockages.

Le format de stockage...	Est utilisé pour décrire les variables dont les valeurs correspondent à...	Par exemple...
date	des dates exprimées dans les formats suivants : ▪ AAAA-MM-JJ ▪ AAAA/MM/JJ	▪ "2001-11-30" ▪ "1999/04/28"
datetime	des dates et heures exprimées dans les formats suivants : ▪ AAAA-MM-JJ HH:MN:SS ▪ AAAA/MM/JJ HH:MN:SS	▪ "2001-11-30 14:08:17" ▪ "1999/04/28 07:21:58"
number	des chiffres, ou valeurs numériques, sur lesquelles peuvent être effectuées des opérations	▪ la variable "salaire", en Euros : "1000.00", "1593" et "2000.54"
integer	des chiffres, ou valeurs numériques entiers, sur lesquelles peuvent être effectuées des opérations	▪ la variable "âge", en années : "21", "34" et "99"
string	des chaînes de caractères alphanumériques	▪ la variable "nom de famille : "Dupond", "Martin" et "Dumoulin" ▪ la variable "profession" : "professeur", "ingénieur" et "traducteur" ▪ la variable "téléphone" : "01 41 44 88 44" et "01 41 44 94 79"

Remarque

Une variable ayant pour valeurs des chiffres ne doit pas nécessairement être décrite par le format de stockage number. Par exemple, les variables "téléphone" et "code postal" doivent être décrites avec le format de stockage string, car aucune opération arithmétique n'ayant de sens ne peut être effectuée sur leurs valeurs. De même, une variable qui servirait d'identifiant pour les observations d'une table et qui dépasserait le format de nombre supporté pourrait être décrite par le format de stockage string.

Attention

Pour le format de stockage number, le séparateur de valeurs décimales utilisé doit être un point, et non une virgule. Ainsi, la valeur "6.5" peut être traitée mais non la valeur "6,5".

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variables de date : les variables générées automatiquement

Lorsque votre jeu de données contient des variables de type date ou date et horaire la fonctionnalité de codage des dates extrait automatiquement des informations de date de ces variables. KDC extrait les informations temporelles suivantes.

Pour les variables de type date ou date et horaire :

Information temporelle	Valeurs	Nom de la variable générée
Jour de la semaine	selon la norme ISO : lundi=0 et dimanche=6	`<NomDeLaVariable>_DoW`
Jour du mois	de 1 à 31	`<NomDeLaVariable>_DoM`
Jour de l'année	de 1 à 366	`<NomDeLaVariable>_DoY`
Mois du trimestre	▪ janvier, avril, juillet et octobre = 1 ▪ février, mai, août et novembre = 2 ▪ mars, juin, septembre et décembre = 3	`<NomDeLaVariable>_MoQ`
Mois de l'année	de 1 à 12	`<NomDeLaVariable>_M`
Année	l'année en quatre chiffre	`<NomDeLaVariable>_Y`
Trimestre	▪ janvier à mars = 1 ▪ avril à juin = 2 ▪ juillet à septembre = 3 ▪ octobre à décembre = 4	`<NomDeLaVariable>_Q`

Pour les variables de type date et horaire :

Information temporelle	Valeurs	Nom de la variable générée
Heure	l'heure	`<NomDeLaVariable>_H`
Minute	la minute	`<NomDeLaVariable>_Mi`
Seconde	la seconde	`<NomDeLaVariable>_S`
µ seconde	la micro-seconde	`<NomDeLaVariable>_mu`

Les variables générées apparaîtront dans les résultats du modèle qui listent les variables, tels que la *Contributions des variables*, les *Détails des variables*, les *rapports de modélisation*, ainsi que dans la fonction de sélection automatique des variables.

3.6.4 Rôles des variables

Dans la modélisation de données, les variables peuvent avoir trois rôles. Elles peuvent être :

- variables cibles,
- variables explicatives,
- variables de poids.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variable cible

Définition

Une **variable cible** est une variable que vous cherchez à expliquer ou dont vous souhaitez prédire les valeurs dans un jeu de données d'application. Elle correspond à votre problématique métier.

Quand la variable cible est une variable binaire, SAP InfiniteInsight® considère que la valeur cible, ou **catégorie cible**, de cette variable (c'est-à-dire la valeur qui fait l'objet de l'analyse) est la valeur la moins fréquente dans le jeu de données d'apprentissage. Imaginons un jeu de données d'apprentissage contenant des informations sur les clients d'une entreprise et contenant la variable cible "a répondu à mon mailing". Cette variable cible a pour valeurs "Oui" ou "Non". Si la valeur "Oui" est la valeur la moins représentée (par exemple, si 40% des clients référencés ont répondu au mailing), SAP InfiniteInsight® considère cette valeur comme catégorie cible de la variable cible.

Synonymes

Selon votre profil et votre domaine d'expertise, vous pouvez être habitué à employer l'un des termes suivants pour référer aux variables cibles :

- variables à expliquer,
- variables dépendantes,
- variables de sortie.

Ces termes sont synonymes.

Exemple

Votre entreprise commercialise deux produits A et B.

Vous possédez une base de données dans laquelle sont référencés :

- 1500 de vos clients. Vous savez quel produit, produit A ou produit B, a acheté chaque client.
- 10000 prospects. Vous souhaitez savoir quel produit est susceptible d'acheter chaque prospect.

La variable "produit acheté" est votre variable cible : elle correspond à votre problématique. Elle est :

- connue sur le jeu de données d'apprentissage (dans notre exemple, les clients),
- inconnue sur le jeu de données d'application (dans notre exemple, les prospects).

Les fonctionnalités SAP InfiniteInsight® vous permettent de modéliser cette variable cible, et donc de prédire quel produit est susceptible d'acheter chacun de vos prospects.

La table suivante représente votre base de données.

Nom	Age	Lieu d'habitation	Catégorie socioprofessionnelle	Produit acheté
Charles	34	Marseille	cadre	Produit A
Jean	37	Paris	cadre	Produit A
Maryline	31	Melun	fonctionnaire	Produit B

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Prospect 1	34	Lille	cadre	?
Prospect 2	24	Paris	fonctionnaire	?
...	...	...	...	...
Prospect n	35	Bordeaux	ouvrier spécialisé	?

Contraintes d'utilisation

Une variable cible présente les contraintes d'utilisation suivantes :

- dans un jeu de données d'apprentissage, **toutes les valeurs de la variable cible doivent être connues**.
- seules les variables **binaires** ou **continues** peuvent être utilisées comme variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variable explicative

Définition

Une **variable explicative** est une variable qui décrit vos données et qui sert à expliquer une variable cible.

Synonymes

Selon votre profil et votre domaine d'expertise, vous pouvez être habitué à employer l'un des termes suivants pour référer aux variables explicatives :

- variables causales,
- variables indépendantes,
- variables d'entrée.

Ces termes sont synonymes.

Exemple

Votre entreprise commercialise deux produits A et B.

Vous possédez une base de données dans laquelle sont référencés :

- 1500 de vos clients. Vous savez quel produit, produit A ou produit B, a acheté chaque client.
- 10000 prospects. Vous souhaitez savoir quel produit est susceptible d'acheter chaque prospect.

Les variables "Nom", "Âge", "Adresse" et "catégorie socioprofessionnelle" sont vos variables explicatives : elles permettent de générer un modèle capable d'expliquer et de prédire les valeurs de variable cible "Produit acheté".

La table suivante représente votre base de données.

Nom	Age	Adresse	Catégorie socioprofessionnelle	Produit acheté
Charles	34	Marseille	cadre	Produit A
Jean	37	Paris	cadre	Produit A
Marilyne	31	Melun	fonctionnaire	Produit B
Prospect 1	34	Lille	cadre	?
Prospect 2	24	Paris	fonctionnaire	?
...	...	...	...	...
Prospect n	35	Bordeaux	ouvrier spécialisé	?

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variable de poids

Définition

Une **variable de poids** permet d'attribuer un poids relatif à chacune des observations qu'elle décrit, et d'orienter le processus d'apprentissage en conséquence. Déclarer une variable comme variable de poids revient à faire un nombre de copies pour chacune des observations du jeu de données qui soit proportionnel à la valeur qu'elles possèdent pour cette variable.

Exemple

Imaginons un jeu de données dans lequel les observations correspondent à des personnes. Ces observations sont entre autres décrites par une variable "Age". Définir la variable "Age" comme variable de poids signifie que pour la génération du modèle, les individus ayant un âge plus élevé auront un poids plus fort que les individus ayant un âge moins élevé.

Contrainte d'utilisation

Seules les **variables continues positives** peuvent être utilisées comme variables de poids.

3.7 Modèles

Le terme "modèle" est fréquemment utilisé et son sens dépend de son champ d'application. En Data Mining, un modèle permet de prédire et d'expliquer des phénomènes, ou de les décrire.

3.7.1 Définition générique

Le terme "modèle" a de nombreuses significations différentes selon le domaine d'application dans lequel il est utilisé. En Data Mining, un modèle décrit et explique les relations qui existent entre des données d'entrée (variables explicatives) et des données de sortie (une ou plusieurs variables cibles). Il permet de prédire et d'expliquer un phénomène, ou de le décrire.

D'après George E.P. Box "*Tous les modèles sont mauvais, mais certains peuvent être utiles*".

Note

Citation de "Robustness is the Strategy of Scientific Model Building" in *Robustness in Statistics. eds.*, R.L. Launer and G.N. Wilkinson, 1979, Academic Press.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.7.2 Performance d'un modèle

Un modèle performant possède à la fois :

- un **bon pouvoir explicatif**, c'est-à-dire une bonne capacité à expliquer la variable cible. Ce pouvoir explicatif est indiqué par l'indicateur de qualité KI.
- une **bonne robustesse**, c'est-à-dire une bonne capacité à conserver les mêmes performances sur de nouveaux jeux de données contenant des observations de la même nature que ceux du jeu de données d'apprentissage. Ce pouvoir explicatif est indiqué par l'indicateur de robustesse KR.

3.7.3 Types de modèles

En Data Mining, il existe deux types de modèles :

- les **modèles prédictifs et explicatifs**, qui permettent de prédire et d'expliquer des phénomènes,
- les **modèles descriptifs**, qui permettent de décrire des jeux de données.

3.7.4 Génération d'un modèle

Le modèle est généré pendant une phase dite "d'apprentissage". Un modèle est généré sur la base d'un jeu de données d'apprentissage.

Selon le cas, ce jeu de données doit être découpé en trois sous-jeux :

- un sous-jeu d'**estimation**,
- un sous-jeu de **validation**,
- un sous-jeu de **test**.

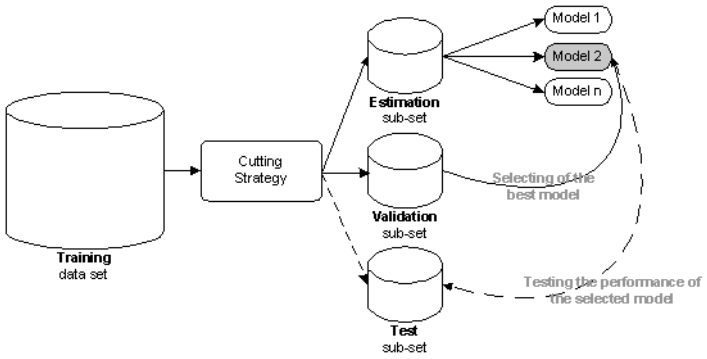
Une stratégie de découpage détermine la manière dont les données du jeu d'apprentissage sont distribuées dans les sous-jeux.

Remarque

Les sous-jeux de données sont virtuels : ils ne sont à aucun moment stockés en mémoire. Le fichier correspondant au jeu de données initial reste intact à tout moment.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Le schéma suivant illustre le processus de génération d'un modèle, également appelé "phase d'apprentissage".



3.7.5 Représentation d'un modèle

Un modèle peut être représenté entre autres sous la forme :

- d'un arbre de décision,
- d'un réseau de neurones,
- d'une fonction mathématique.

Dans SAP InfiniteInsight®, les modèles sont représentés sous la forme de fonctions mathématiques, et plus exactement de **polynômes**.

Description du polynôme

Un polynôme peut être de degré 1, 2, 3 ou plus. En définissant l'**ordre du polynôme**, vous définissez le **degré de complexité du modèle.**

Exemples de polynômes

Un polynôme d'ordre 1 est de la forme :

```
f(X1, X2, ..., Xn) = w0 + w1.X1 + w2.X2 + ... + wn.Xn
```

Un polynôme d'ordre 2 est de la forme :

```
f(X1, X2, ..., Xn) = w0 + w1.X1 + w2.X2 + ... + wn.Xn + w11X1.X1 + w12.X1.X2 + w13.X1.X3
+ ... + wij.Xi.Xj
```


© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Méthodologie

Dans la grande majorité des cas, un degré 1 est suffisant pour générer un modèle pertinent et robuste.

Un ordre de polynôme élevé ne garantit pas toujours l'obtention de résultats meilleurs que ceux obtenus avec un polynôme d'ordre 1. De plus, plus vous sélectionnez un ordre de polynôme élevé et plus :

- le temps nécessaire pour générer le modèle correspondant est important,
- le temps nécessaire pour appliquer le modèle à de nouveaux jeux de données est important,
- les résultats de la modélisation sont difficiles à interpréter.

Le choix de tel ou tel ordre pour le polynôme dépend de la nature des données à analyser. La méthodologie conseillée est de :

- générer en premier lieu un modèle ayant un degré d'ordre 1. Dans la grande majorité des cas, ce degré est suffisant pour garantir un modèle pertinent et robuste.
- tester les résultats obtenus avec des modèles de degré supérieur, si les performances du modèle de degré 1 semblent insuffisantes.

3.7.6 Validation d'un modèle

Une fois le modèle généré, vous devez vérifier sa validité en observant les indicateurs de performance :

- la **capacité prédictive** vous permet de connaître le **pouvoir explicatif** du modèle, c'est-à-dire sa capacité à expliquer les valeurs de la variable cible sur le jeu de données d'apprentissage. Un modèle parfait possède une capacité prédictive égale à 1 et un modèle purement aléatoire possède une capacité prédictive égale à 0.
- la **reproductibilité** vous permet de connaître le **degré de robustesse** du modèle, c'est-à-dire sa capacité à conserver le même pouvoir explicatif sur un nouveau jeu de données. En d'autres mots, le degré de robustesse correspond à la capacité prédictive du modèle sur un jeu de données d'application.

Pour savoir comment sont calculés la capacité prédictive et la reproductibilité, voir Capacité prédictive, reproductibilité et courbes de profit à la page 232.

Remarque

La validation du modèle est une phase primordiale dans le processus global de Data Mining. Accordez toujours une importance majeure aux valeurs obtenues pour la capacité prédictive et la reproductibilité d'un modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.7.7 Dans quels cas un modèle est-il acceptable

Reproductibilité : indicateur de robustesse acceptable

Un modèle possédant une reproductibilité inférieure à 0.95 doit être considéré avec précaution. Les performances d'un tel modèle ont de fortes chances de varier entre le jeu de données d'apprentissage et les jeux de données d'application.

Capacité prédictive : indicateur de qualité acceptable

Aucun seuil minimum n'est requis pour le pouvoir prédictif d'un modèle. Tout dépend de votre contexte métier, c'est-à-dire de votre domaine d'application, de la nature de vos données et de votre problématique.

Dans certains cas, un modèle possédant une capacité prédictive de seulement 0,1 peut permettre de réaliser un profit équivalent à plusieurs milliers d'euros. Dans tous les cas, une capacité prédictive positive indique que le modèle généré est plus performant qu'un modèle de type aléatoire, et permet donc de réaliser un profit.

3.7.8 Comment obtenir un meilleur modèle

Obtenir un meilleur modèle consiste :

- soit à améliorer la reproductibilité du modèle,
- soit à améliorer la capacité prédictive du modèle,
- soit à améliorer à la fois la capacité prédictive et la reproductibilité du modèle.

Plusieurs techniques permettent d'améliorer ces indicateurs :

- vous pouvez augmenter le degré de complexité du modèle (ordre du polynôme).
- le tableau suivant présente d'autres techniques.

Pour améliorer...	Vous pouvez...
la *capacité prédictive* d'un modèle	▪ ajouter des variables dans le jeu de données d'apprentissage ▪ effectuer des combinaisons de variables explicatives qui vous semblent pertinentes
la *reproductibilité* d'un modèle	ajouter des observations dans le jeu de données d'apprentissage

> **Remarque**
>
> Pour plus d'informations sur l'amélioration de la capacité prédictive et de la reproductibilité, consultez l'aide contextuelle de SAP InfiniteInsight®.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.8 Indicateurs de performance

3.8.1 Indicateurs spécifiques à SAP InfiniteInsight®

Deux indicateurs vous permettent de connaître la performance d'un modèle.

- la **capacité prédictive (KI)**, qui est l'indicateur de qualité,
- la **reproductibilité (KR)**, qui est l'indicateur de robustesse.

La capacité prédictive : indicateur de qualité

Définition

La capacité prédictive est l'indicateur de qualité des modèles générés par SAP InfiniteInsight®. Cet indicateur correspond au taux d'information contenu dans la variable cible que les variables explicatives permettent d'expliquer.

Exemple

Un modèle possédant une capacité prédictive égale à :

- "0,79" est capable d'expliquer 79% de l'information contenue dans la variable cible grâce aux variables explicatives contenues dans le jeu de données analysé.
- "1" est un hypothétique modèle parfait, capable d'expliquer 100% de la variable cible grâce aux variables explicatives contenues dans le jeu de données analysé. Dans la réalité, une telle capacité prédictive indique généralement qu'une variable 100% corrélée à la variable cible n'a pas été exclue du jeu de données analysé.
- "0" est un modèle purement aléatoire.

Améliorer la capacité prédictive d'un modèle

Pour améliorer la capacité prédictive d'un modèle, de nouvelles variables peuvent être ajoutées au jeu de données d'apprentissage. Des combinaisons de variables explicatives peuvent également être effectuées.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

La reproductibilité : indicateur de robustesse

Définition

La reproductibilité est l'indicateur de robustesse des modèles générés par SAP InfiniteInsight®. Elle indique la capacité d'un modèle à conserver les mêmes performances dans le cas où il est appliqué à un nouveau jeu de données présentant les mêmes attributs que le jeu de données d'apprentissage.

Exemple

Un modèle possédant une reproductibilité:

- égale à "0,98" est très robuste. Il possède une forte capacité de généralisation.
- inférieure à "0,95" devrait être considéré avec précaution. Son application sur un nouveau jeu de données présenterait le risque de générer des résultats douteux.

Améliorer la reproductibilité d'un modèle

Pour améliorer la reproductibilité d'un modèle, des lignes d'observations peuvent être ajoutées au jeu de données d'apprentissage.

Capacité prédictive, reproductibilité et courbe de profit

Sur le graphique des courbes de profit :

- du jeu de données d'estimation (graphique par défaut), la capacité prédictive correspond au rapport entre "la surface se trouvant entre la courbe du **modèle généré** et celle du **modèle aléatoire"** et "la surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire"**. Ainsi plus la courbe du modèle généré se rapproche de la courbe du modèle parfait, plus la capacité prédictive se rapproche de 1.
- des jeux de données d'estimation, de validation et de test (sélectionnez l'option correspondante dans la liste *Jeu de données*, située sous le graphique), la reproductibilité correspond à 1 moins le rapport entre la "surface se trouvant entre la courbe du **jeu d'estimation** et celle du **jeu de validation**" et la "surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire**".

Pour plus d'informations sur les courbes de profit, voir Les courbes de profit.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Utilisation avancée : la capacité prédictive pour des cibles continues

A partir de SAP InfiniteInsight® 7.0, dans le cas d'une variable cible continue, la régression utilise la capacité prédictive (KI) pour le choix du modèle.

En supposant que nous voulons calculer la capacité prédictive d'un score d'une variable rr_T tout en prenant en compte sa cible T d'un jeu de données Validation, nous allons donc nous référer aux catégories cibles avec $T_j \; pour \; j = 1 \dots B_T$. Ainsi, nous notons :

$$\mu_j = mean(T_j) \text{ pour } j = 1 \dots B_T$$

$$f_j = \; frequency(T_j) \text{ pour } j = 1 \dots B_T$$

Les catégories cibles sont données par ordre décroissant, de manière à ce que $(\mu_1 > \mu_2 > \cdots > \mu_{B_T})$. Considérons μ comme la moyenne globale de la cible T dans le jeu de données Validation.

Nous nous référons également aux catégories des scores selon $S_j, \; j = 1 \dots B_S$ et nous notons donc :

$$m_j = target\ mean(S_j) \text{ pour } j = 1 \dots B_S$$

$$F_j = frequency(S_j) \text{ pour } j = 1 \dots B_S$$

La courbe du Wizard est calculée à partir des profits cumulatifs de la cible en fonction des fréquences cumulatives, la courbe est définie par les points suivants :

$$\left(\sum_{j=1}^{j=b} f_j \; , \sum_{j=1}^{j=b} f_j \left(\mu_j - \; \mu \right) \right) \; pour \; b = 1 \dots B_T$$

La courbe est normalisée de manière à ce que sa valeur maximale soit égale à 1.

De plus, la courbe Validation est calculée à partir des profits cumulatifs des scores en fonction des fréquences cumulatives :

$$\left(\sum_{j=1}^{j=b} F_j \; , \sum_{j=1}^{j=b} F_j \left(m_j - \; \mu \right) \right) \text{ pour } b = 1 \dots B_S$$

La valeur de la capacité prédictive est calculée à partir du Wizard et des zones de la courbe Validation. Par exemple, les zones peuvent être calculées en utilisant la méthode des trapèzes.

> **Note**
>
> Le cas d'une cible nominal peut être considéré comme un cas spécial où la notion de profit correspond au taux positif (qui est équivalent à la moyenne de la cible binaire dans ce cas).

La capacité prédictive repose alors sur l'ordre des valeurs estimées et cet ordre est comparé aux réelles cibles continues. En conséquence, elle est plus robuste que les indicateurs L1 (l'erreur absolue moyenne) ou L2 (l'erreur quadratique moyenne, l'erreur racine carrée de l'erreur quadratique moyenne ou le coefficient de Pearson) souvent utilisés pour la régression, car une cible erronée ne peut pas diminuer la capacité prédictive globale (mais constitue une des causes principales pour l'instabilité de tous les autres indicateurs). De l'autre côté, la capacité prédictive ne prend pas en compte les valeurs estimées par rapport aux valeurs cible. C'est-à-dire qu'un modèle avec des valeurs estimées dans l'intervalle [-2;2] peut obtenir une très bonne capacité prédictive, même si les cibles réelles se trouvent dans l'intervalle [0;100], pourvu que le modèle ait trouvé l'ordre correct entre les valeurs estimées et les cibles réelles. La technologie InfiniteInsight® limite cet effet en offrant une recalibration linéaire par morceau des valeurs estimées vers les cibles réelles basées sur les statistiques du jeu de données de validation. Ainsi vous n'obtenez pas seulement de bonnes estimations de l'ordre mais également de bonnes estimations de l'intervalle.

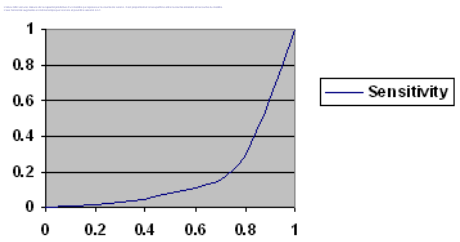
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.8.2 Autres indicateurs

Trois autres indicateurs, communément utilisé en Data Mining, sont fournis pour évaluer la qualité d'un modèle SAP InfiniteInsight®:

- le GINI index,
- le K-S,
- le AUC.

GINI index



La formule correspondante est :

$$GINI = 2(1-p_G)\left(\frac{1}{2}-(1-AUC)\right) = (1-p_G)(2AUC-1)$$

K-S

Le K-S est le critère de Kolmogorov-Smirnov appliqué comme mesure de la déviation par rapport aux taux de réponse uniformes pour les catégories d'une variable. K-S est un test d'ajustement non paramétrique qui repose sur la déviation maximale entre les fonctions de distribution cumulative et empirique.

Dans le cas d'un classement binaire, ce qui intéresse les utilisateurs c'est la différence entre la courbe de Lorenz pour les cas positifs '1-α' (voir à la page 49), et la courbe de Lorenz pour les cas négatifs 'β' (voir à la page 49) lorsqu'on sélectionne une proportion croissante de la population. Ces courbes évoluent en même temps entre 0 et 1, et le K-S est la déviation maximale entre ces deux courbes. Lorsqu'on a un système parfait, le K-S est égal à 1, et lorsque le système est aléatoire le K-S est égal à 0, à cause de l'égalité entre les deux courbes.

Conseil

Le K-S est utilisé pour calculer la différence entre deux distributions afin d'avoir une meilleur idée de la qualité du jeu de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.8.3 Indicateurs d'erreurs

Quelques précisions préalables :

- Cible (valeur de réponse) : y_i
- Prédicteur (prédicteur des valeurs de résponse) : $\hat{y}_i$
- Résidu : $r_i = y_i - \hat{y}_i$
- Erreur : $u_i = | y_i - \hat{y}_i | = | r_i |$
- Poids des observations testées : w_i
- Poids total de la population : $W = \sum_{i=1}^{n} w_i$
- Cible moyenne : $\bar{y} = \frac{1}{W} \sum_{i=1}^{N} w_i y_i$
- Prédicteur moyen : $\bar{\hat{y}} = \frac{1}{W} \sum_{i=1}^{N} w_i \hat{y}_i$

Erreur absolue moyenne (L1)

Définition : moyenne arithmétique des valeurs absolues des écarts (distance Manhattan ou City block)

Formule :

$$L1 = \frac{1}{W} \sum_{i=1}^{N} w_i u_i$$

Erreur quadratique moyenne (L2)

Définition : racine carré de la moyenne arithmétique des carrés des écarts (l'importance des grosses erreurs est majorée) (distance Euclidienne)

Formule :

$$SSE_w = \sum_{i=1}^{N} w_i {u_i}^2$$

$$MSE = \frac{SSE_w}{W} = \frac{1}{W} \sum_{i=1}^{N} w_i {u_i}^2$$

$$L2 = \sqrt{MSE}$$

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Erreur maximale (LInf)

Définition : écart maximum (distance de Chebyshev)

Formule :

$$L\infty = \max_i u_i$$

Erreur moyenne (ErrorMean)

Définition : moyenne arithmétique des écarts

Formule :

- Erreur moyenne en pourcentage (MPE) :

$$MPE = \frac{1}{W}\sum_{i=1}^{N} w_i \frac{y_i - \hat{y}_i}{y_i}$$

- Erreur moyenne absolue en pourcentage (MAPE) :

$$MAPE = \frac{1}{W}\sum_{i=1}^{N} w_i \frac{|\, y_i - \hat{y}_i \,|}{y_i}$$

Ecart-type de l'erreur (ErrorStdDev)

Définition : dispersion des erreurs autour du résultat réel

Formule :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - \overline{u})^2}$$

où

$$\overline{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$$

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Taux de classement (ClassificationRate)

Définition : rapport entre le nombre d'enregistrements classifiés correctement et le nombre total d'enregistrements

Formule :

$$CR(t) = \frac{G\alpha(t)+B\beta(t)}{G+B} = p_G\alpha(t) + (1-p_G)\beta(t)$$

Coefficient de détermination (R2)

Définition : rapport entre la variabilité des prédictions (somme des carrés expliqués) et la variabilité des données (somme des carrés totaux)

Formule :

$$SSR = \sum_{i=1}^{N} w_i(\hat{y}_i - \bar{y})^2$$

$$SST = \sum_{i=1}^{N} w_i(y_i - \bar{y})^2$$

$$R2 = \frac{SSR}{SST}$$

3.9 Types de profit

3.9.1 Définition

Un **type de profit** permet de calculer le profit réalisable grâce à l'utilisation d'un modèle. De manière générale, un bénéfice est associé aux valeurs souhaitées (ou attendues) de la variable cible et un coût est associé à ses valeurs non souhaitées (ou non attendues). Par exemple, dans le cadre d'une campagne d'envois publicitaires, une personne se voit associée à :

- un **bénéfice** si elle répond à l'envoi publicitaire,
- un **coût** si elle ne répond pas l'envoi publicitaire.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.9.2 Les quatre types de profit

Pour visualiser le profit réalisable grâce à un modèle généré avec SAP InfiniteInsight®, vous pouvez utiliser les quatre types de profit suivants :

- **Détecté**,
- **Lift**,
- **Normalisé**,
- **Personnalisé**.

Le profit détecté

Le **profit détecté** est le type de profit proposé par défaut. Il permet de visualiser le pourcentage d'observations appartenant à la catégorie cible de la variable cible, c'est-à-dire la catégorie la moins fréquente, en fonction du taux d'observations sélectionné sur la totalité du jeu de données. Avec ce profit :

- la valeur "0" est affectée aux observations n'appartenant pas à la catégorie cible de la variable cible,
- la valeur "1/ (fréquence de la catégorie cible de la variable cible dans le jeu de données)" est affectée aux observations appartenant à la cible.

Le profit Lift

Le **profit Lift** permet de visualiser la différence entre un modèle parfait (Wizard) et un modèle aléatoire et entre le modèle généré et un modèle aléatoire. Le modèle aléatoire sert de référence et est toujours égal à 1.

Le profit normalisé

Le **profit normalisé** permet de visualiser l'apport du modèle généré par les fonctionnalités SAP InfiniteInsight® par rapport à un modèle de type aléatoire, c'est-à-dire un modèle qui vous permettrait de sélectionner uniquement au hasard des observations dans votre base de données.

Ce profit est utilisé pour les graphiques de détail des variables, qui présentent l'importance de chacune des catégories d'une variable donnée par rapport à la variable cible.

Le profit personnalisé

Le **profit personnalisé** vous permet de définir vos propres valeurs de profit, c'est-à-dire d'associer à chaque valeur de la variable cible un coût et un bénéfice. Par exemple, vous pouvez définir le coût d'envoi d'un mailing et le gain apporté par la réponse à ce mailing.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.10 Courbes avancées

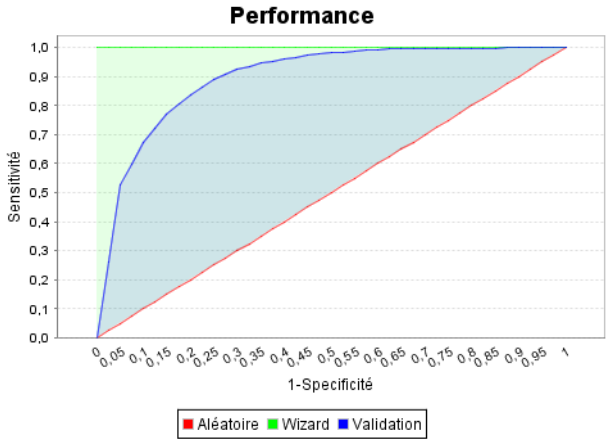
En plus des courbes de profit décrites dans la section précédente, un ensemble de courbes avancées est proposé par SAP InfiniteInsight®.

3.10.1 ROC

La courbe *ROC* (*Receiver Operating Characteristic*) est dérivée de la théorie de détection du signal. Elle permet d'étudier les variations de la spécificité et de la sensitivité d'un test pour différentes valeurs du seuil de discrimination.

La *Sensitivité*, qui apparaît sur l'axe des ordonnées, est la proportion de signaux trouvés qui ont été **correctement** identifiés (également appelés *vrais positifs*).

[1- la Spécificité], qui apparaît sur l'axe des abscisses, est la proportion de signaux **incorrectement** identifiés comme positifs (autrement dit les *faux positifs*)

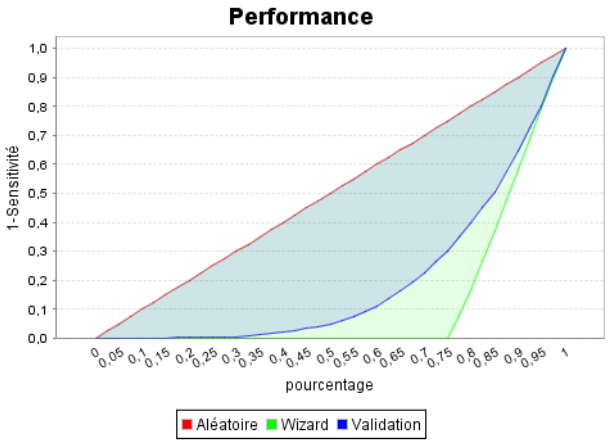


© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.10.2 Courbes de Lorenz

Lorenz "Bon"

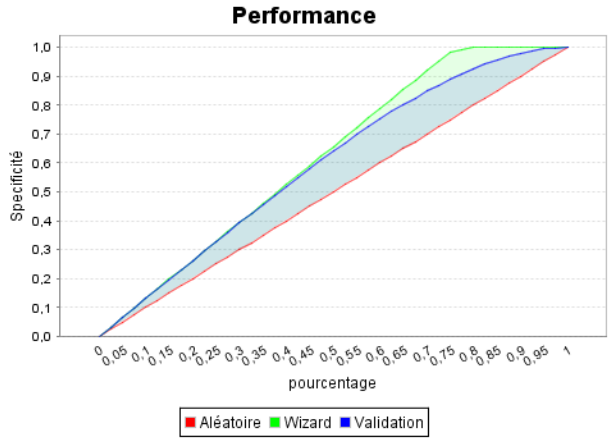
La courbe *Lorenz "Bon"* affiche la proportion cumulative des signaux mal devinés (faux négatifs) trouvés dans les n% de scores les plus bas.



L'axe des ordonnées mesure *[1- Sensitivité]*, c'est-à-dire `[1 - proportion de vrais positifs]`, ce qui équivaut à la proportion des signaux manqués ou des opportunités perdues. Les données étant ordonnées de gauche à droite, des enregistrements les moins susceptibles d'être des signaux ceux les plus susceptibles de l'être, plus la courbe montre lentement, plus le modèle est sensible en terme de détection des signaux. La courbe du modèle parfait (en vert) augmente à partir du point de l'axe des abscisses correspondant à la proportion de non-signaux dans le jeu de données de validation.

Lorenz "Mauvais"

La courbe de *Lorenz 'Mauvais'* affiche la proportion cumulée de vrais négatifs (specificité) représentés par les x% scores les plus bas du modèle. Plus la courbe augmente rapidement, plus la fréquence de détection erronée est faible.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3.10.3 Courbes de densité

Les courbes de densité affiche la fonction de la densité de la variable *Score* dans l'ensemble des signaux (*Courbe de densité 'Bon'*) et dans l'ensemble des non-signaux (*Courbe de densité 'Mauvais'*). Ces courbes peuvent aussi être vues comme la dérivée de la courbe de Lorenz.

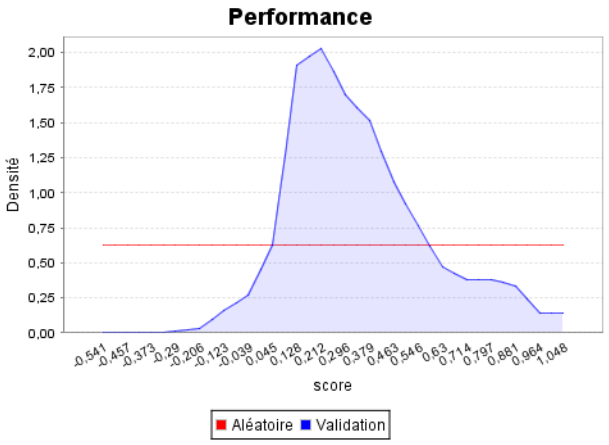
La fonction estimée de la densité dans un groupe ou intervalle est égale à:

```
(nombre de signaux dans l'intervalle / nombre total de signaux) / longueur de l'intervalle
```

La longueur d'un intervalle est par définition sa borne supérieure moins sa borne inférieure.

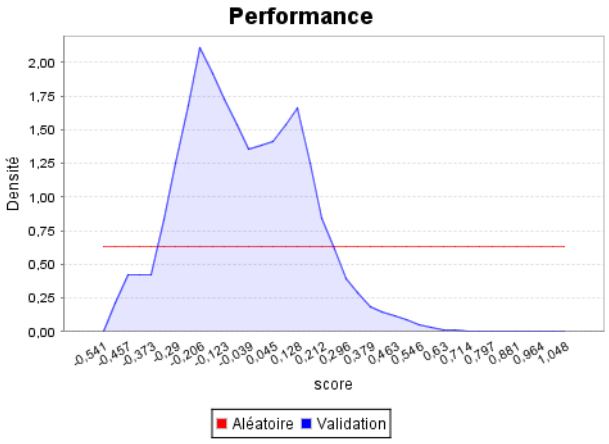
Courbe de densité "Bon"

La *courbe de densité "Bon"* représente la distribution des scores du modèles pour les réponses positives.



Courbe de densité "Mauvais"

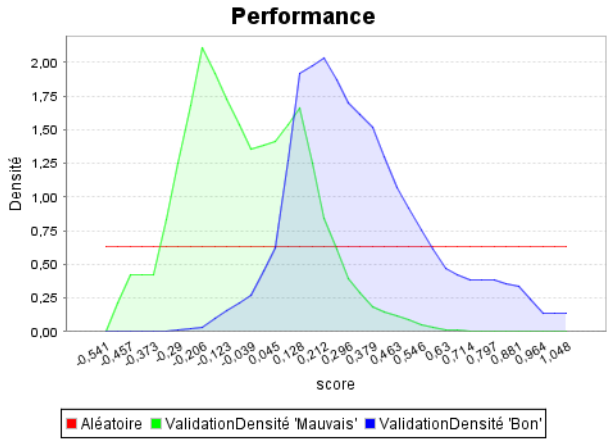
La *courbe de densité "Mauvais"* représente la distribution des scores du modèle pour les réponses négatives.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Courbes avancées > Courbe de densité "Tous"

La *courbe de densité "Tous"* affiche à la fois les courbes de densité *"Bon"* et *"Mauvais"*, vous permettant ainsi de comparer les deux distributions.

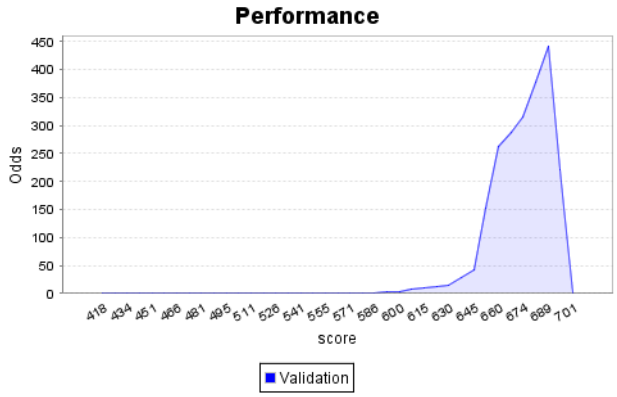


3.10.4 Courbes de "Risque"

Good/Bad Odds

L'axe des abscisses représente le risque et l'axe des ordonnées la valeur du rapport bon/mauvais.

Le rapport bons/mauvais est égal à `(1-p)/p` , où `p` est défini comme étant la probabilité du risque.

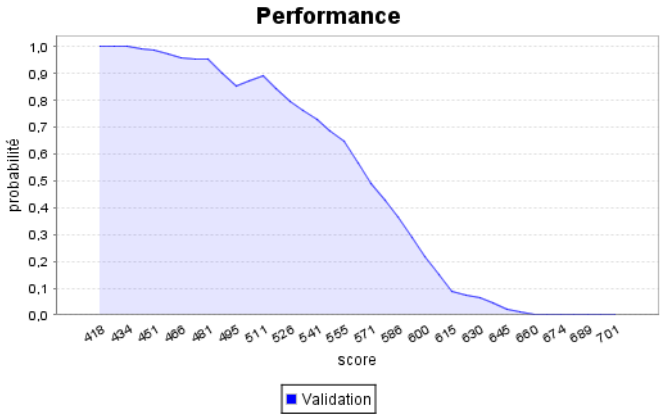


© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Probabilité du risque

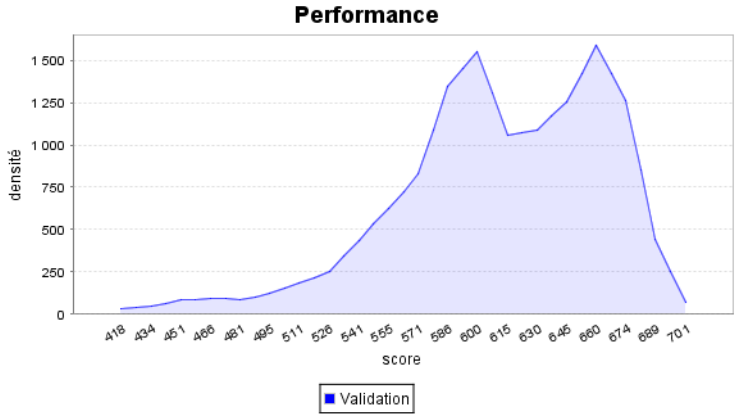
L'axe des abscisses représente le risque et l'axe des y la valeur de probabilité de risque.

La probabilité du risque p est calculée pour chaque regroupement de score de risque comme suit : le nombre de "mauvais" divisé par le nombre d'enregistrements dans un regroupement.



Densité de la population

La densité de la population est calculée en se basant sur le nombre d'enregistrements de score de risque dans chaque regroupement de score de risque (20 par défaut).



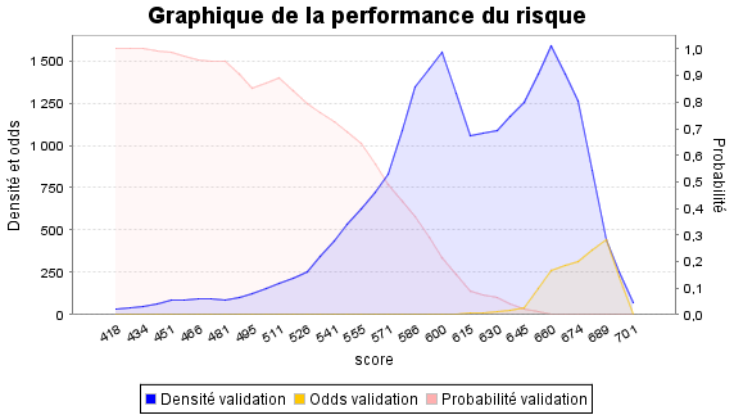
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Risque 'tout'

Les courbes représentant le risque sont affichées sur un même graphe (à l'exception de la courbe **Log(Good/Bad Odds)**).

Note

L'axe des ordonnées pour la courbe de probabilité se trouve sur la droite et que la courbe de population de densité et du Bon/Mauvais partagent le même axe des ordonnées à gauche.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4 Scénario d'utilisation : Gagnez en efficacité et maîtrisez votre budget grâce à la modélisation

4.1 Présentation

Dans ce scénario, vous êtes le Directeur Marketing d'une grande banque de détail. Cette banque souhaite proposer un nouveau produit financier à ses clients. Votre projet consiste à lancer une campagne de marketing direct visant à promouvoir ce nouveau produit. Vous avez à disposition une importante base de données de prospects et un budget restreint et fortement contrôlé, et vous êtes soumis à des contraintes de temps importantes. Afin de maximiser les bénéfices associés à votre campagne, votre problématique consiste à :

- **contacter les prospects les plus susceptibles d'être intéressés** par le nouveau produit financier,
- **identifier le nombre idéal de prospects à contacter** sur l'ensemble de votre base de données.

Grâce au composant **InfiniteInsight® Modeler / Régression ou Classement** (K2R), vous construisez un modèle explicatif et prédictif dans les meilleurs délais et à moindre coût. Ce modèle vous permet de répondre à votre problématique et de remplir vos objectifs.

4.2 Votre objectif

Imaginons le cas suivant.

Vous êtes le Directeur Marketing d'une grande banque de détail. Cette banque a décidé de proposer à ses clients **un nouveau produit d'épargne haut de gamme**. Elle s'apprête à **lancer une campagne de marketing direct** d'envergure pour promouvoir cette nouvelle offre auprès de ses prospects et de ses clients.

La banque connaît un contexte concurrentiel fort et la Direction Générale, consciente de l'enjeu que représente le lancement de ce nouveau produit financier, souhaite que la campagne marketing soit accomplie **dans les meilleurs délais**.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4.3 Vos moyens

4.3.1 Un budget restreint et fortement contrôlé

Le Contrôle de Gestion de la banque est très rigoureux, et le **budget** qui vous est alloué pour cette campagne marketing :

- ne vous permet pas de contacter l'ensemble des prospects de la banque,
- ne peut pas être dépassé.

4.3.2 L'information à votre disposition

Le Département Marketing dispose pour cette campagne d'**une base de données dans laquelle sont référencés 1 000 000 prospects**, identifiés par leurs caractéristiques principales :

- Age,
- Sexe,
- Catégorie socioprofessionnelle,
- Nationalité,
- Profession,
- Diplôme,
- Nombre d'heures travaillées par semaine,
- Etc.

Vous constatez rapidement que la base de données que vous avez à disposition n'est pas optimale. Cette base de données contient en effet :

- des données disparates,
- des données redondantes,
- des données manquantes.

Des données disparates

La base de données contient aussi bien des informations alphanumériques (telles que "profession" et "nationalité") que des informations numériques (telles que "âge" et "montants des encours sur les comptes").

Des données manquantes

Dans la base de données, certaines informations sont manquantes. Pour gérer cette absence d'information, la Direction Informatique a utilisé la norme suivante :

- le symbole "?" signifie qu'une valeur alphanumérique (telle que la "profession") est manquante,
- la valeur "99999" signifie qu'une valeur numérique (telle que "l'âge") est manquante.

Vous n'avez malheureusement ni le temps ni les ressources nécessaires pour :

- lancer une enquête visant à compléter les informations manquantes,
- mettre en forme la base de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Editer les options

☑ Editer les options de l'assistant de modélisation

1 Dans le menu *Fichier*, cliquez sur *Préférences...*.
Une fenêtre *Editer les options...* s'ouvre.

Les options suivantes peuvent être modifiées :

Catégorie	Options
Général	▪ Pays ▪ Langage ▪ Niveau de message ▪ Taille maximum du fichier log ▪ Niveau de message pour les valeurs aberrantes ▪ Afficher l'arbre des paramètres ▪ Taille de l'historique des répertoires ▪ Toujours quitter sans confirmer ▪ Inclure test dans la stratégie de découpage par défaut
Emplacements	▪ Emplacement par défaut pour les données d'application en entrée ▪ Emplacement par défaut pour les données d'application en sortie ▪ Emplacement par défaut pour l'enregistrement des modèles
Entrepôt de métadonnées	▪ Activer un espace de stockage unique pour les métadonnées ▪ Editer le contenu de la bibliothèque de variables
Graphique	▪ Nombre de points de la courbe de performance ▪ Nombre de barres affichées ▪ Désactiver le Look and feel SAP InfiniteInsight® ▪ Afficher les diagrammes en 3D ▪ Désactiver le double tampon ▪ Optimiser pour les affichages distants ▪ Se souvenir de la position et de la taille en quittant
Rapport	▪ Nombre de variables intéressantes ▪ Feuille de style active ▪ Personnalisez vos feuilles style
Géolocalisation	▪ Protocol du système d'information géographique

Personnaliser les feuilles de style

SAP InfiniteInsight® vous offre la possibilité de personnaliser les rapports. La feuille de style par défaut, appelée *Feuille de style SAP InfiniteInsight® (par défaut)*, ne peut être modifiée. Vous devez créer vos propres feuilles de styles pour changer la configuration.

Note

Pour créer, charger et enregistrer une feuille de style, vous devez préciser le répertoire des feuilles de style dans le panneau *Editer les options...* avant d'ouvrir la fenêtre *Editeur de feuilles de style SAP InfiniteInsight®*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Créer une nouvelle feuille de style

1. Dans le champ *Répertoire*, cliquez sur le bouton (*Parcourir*).
2. Sélectionnez un dossier qui contiendra vos feuilles de style.
3. Cliquez sur le bouton (*Ajouter*).
   Une nouvelle feuille de style a été créée.
4. Cliquez sur le bouton .
   La fenêtre *Editeur de feuilles de style* s'ouvre.
5. Dans le champ *Nom de la feuille de style*, entrez un nom pour la nouvelle feuille de style.
   L'extension `.krs` est automatiquement ajoutée.

> **Note**
>
> Vous pouvez dupliquer une feuille de style en modifiant le nom de votre feuille. La feuille de style précédente n'est pas supprimée.

☑ Supprimer une feuille de style

1. Sélectionnez une des feuilles de styles proposées.
2. Cliquez sur le bouton (*Retirer*).

> **Note**
>
> La feuille de style n'est pas seulement supprimée de la liste, mais également du répertoire.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Modifier la configuration générale

Configuration...	Options...	Note...
Couleur de fond	▪ choisir la couleur ▪ rendre transparent	Uniquement les formats PDF et HTML peuvent afficher une couleur de fond.
Editer la configuration	▪ taille des polices ▪ style ▪ couleurs de fond ▪ configuration de tableaux	Cochez l'option *Rendre dynamiquement les changements* ou cliquez sur *Appliquer* pour visualiser les modifications.

Les options sélectionnées s'appliquent à l'assistant de modélisation et aux rapports générés.

☑ Modifier les paramètres des graphiques

Configuration...	Options...	Note...
Couleurs des graphiques	▪ modifier les couleurs	
Histogrammes	▪ horizontal ▪ vertical	Il est possible de choisir une orientation différente que celle définie par défaut pour une section spécifique.

☑ Modifier des sections de rapport

1 Sélectionnez les propriétés de votre choix.

2 Cliquez sur *Enregistrer*.
Une fenêtre s'ouvre, indiquant que votre feuille de style a bien été sauvegardée.

3 Cliquez sur *OK*.

Configuration...	Options...	Note...
Type de vue	choisissez entre tabulaire, HTML et graphique. La dernière option n'est disponible que si la section peut être affichée comme graphique.	
Type de graphique	choisissez un des types proposés.	Cette option n'est disponible que pour le sections du type *graphique*.
Basculer l'orientation	cette option vous permet de choisir une orientation différente que celle définie par défaut pour une section de rapport	
Trier	vous pouvez choisir la colonne à utiliser pour le tri et l'ordre de tri	
Visibilité	vous pouvez cacher une colonne d'une section ou même toute une section de rapport	Au moins une colonne d'une section do rester visible.

☑ Appliquer la nouvelle feuille de style aux rapports générés

4 Dans la fenêtre *Rapport*, sélectionnez la feuille de style que vous souhaitez appliquer à vos rapports.

5 Cliquez sur *OK*.
Une fenêtre s'ouvre, indiquant que vous devez redémarrer l'assistant de modélisation pour prendre en compte les options modifiées.

6 Cliquez sur *OK*.
Lorsque vous exécutez un modèle, tous les rapports générés (rapport de modélisation, rapport excel et rapport statistique) sont personnalisés selon votre feuille de style.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Définir un entrepôt de métadonnées

L'entrepôt de métadonnées vous permet de spécifier l'emplacement où les métadonnées doivent être enregistrées.

☑ Pour définir un entrepôt de métadonnées

1. Choisissez de placer les métadonnées au même endroit que les données ou dans un endroit spécifique en cochant l'option de votre choix.
2. Dans la liste *Type de données*, sélectionnez le type de données auxquelles vous souhaitez accéder. L'accès à certains types de données nécessitent une licence spécifique.
3. Utilisez le bouton *Parcourir* correspondant au champ *Répertoire* pour sélectionner le répertoire ou la base de données contenant les données désirées. Si la base de données est protégée, saisissez le nom d'utilisateur et le mot de passe dans les champs *Identifiant* et *Mot de passe*.
4. Cliquez sur le bouton *Editer le contenu de la bibliothèque de variables* pour éditer les descriptions des variables stockées dans la bibliothèque de variables.
5. Cliquez *OK* pour valider.

Environnement technique

La base de données mise à votre disposition est stockée dans un SGBD/R (système de gestion de bases de données relationnelles) sur un serveur UNIX, géré par la Direction Informatique de la banque. Cet environnement informatique constitue des contraintes techniques pour le choix d'un éventuel outil d'analyse de données.

4.4 Votre approche

En raison de l'enjeu important de la campagne à mener, de votre budget limité et du manque de visibilité sur le nouveau produit, vous avez décidé de minimiser les risques en divisant le projet en deux étapes :

1. **Test de la campagne marketing sur un échantillon** de 50 000 personnes issues de la base de prospects de 1 000 000 de personnes.
2. **Lancement global de la campagne marketing sur la totalité** de la base de prospects.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4.4.1 La phase de test de votre campagne marketing

La phase de test de votre campagne marketing vous a permis de collecter un échantillon de 50 000 personnes dont vous connaissez le comportement par rapport au nouveau produit :

- 25% des prospects se sont montrés clairement intéressés. Ils ont décidé d'accepter un rendez-vous avec un des opérateurs de vos canaux de vente,
- 75% des prospects ont décliné votre invitation.

Votre problématique consiste à comprendre les résultats de ce test, en identifiant les raisons pour lesquelles certaines personnes ont répondu favorablement à votre offre et pourquoi d'autres, au contraire, ont répondu négativement. Vous pourrez alors vous servir du modèle d'analyse obtenu pour prédire le comportement de chacun des 1 000 000 prospects de votre base de données. Vous optimiserez ainsi votre campagne marketing en ne proposant cette offre qu'à des personnes susceptibles d'être intéressées.

Le fichier contenant le jeu de données utilisé pour le test vous a été remis par la Direction Informatique de la banque sous la forme d'un fichier plat (.csv). Ce fichier correspond au fichier exemple *Census01.csv*, livré avec SAP InfiniteInsight® et décrit dans la section Présentation des fichiers exemples (voir à la page 63).

4.5 Votre problématique

Suite à la phase de test votre campagne, vous possédez dans votre base de données marketing :

- une liste de 1 000 000 prospects,
- une liste de 50 000 prospects, sélectionnés de manière aléatoire lors de cette phase de test, et dont vous connaissez maintenant la réponse vis à vis de votre campagne. Cet échantillon, issu de votre base de données initiale, comporte également des valeurs manquantes et des variables corrélées.

Votre problématique consiste à utiliser en l'état ce jeu de données, en tant que jeu de données d'apprentissage, pour :

- **créer rapidement un modèle explicatif et prédictif**,
- **appliquer ensuite ce modèle sur la totalité de votre base de données**.

Grâce au modèle généré, vous serez en mesure de savoir :

- **A combien d'individus référencés dans votre base de prospects vous devez envoyer votre courrier**, afin de maximiser le profit/retour sur investissement de votre campagne ?
- Comment **classer l'ensemble des individus de votre base de prospects selon leur « appétence »** (probabilité d'achat) pour ce nouveau produit. Cette appétence se traduit par une probabilité, ou "score", qu'un prospect réponde favorablement à la campagne.
- **Quels sont ces individus et quel est leur profil** ? Valider quels sont les critères (âge, catégorie socioprofessionnelle, diplôme) qui expliquent qu'une personne se montre intéressée ou pas par le nouveau produit financier.
- Comment **simuler en temps réel la capacité d'un individu isolé à répondre favorablement à la nouvelle offre**, notamment pour permettre au "Call Center" de votre banque ou à un chargé de clientèle de connaître immédiatement l'appétence d'un nouveau client pour ce produit financier (Simulation).
- Comment enregistrer ce Score dans votre base de données de prospects, afin de pouvoir sélectionner simplement ultérieurement des sous-ensembles de population pour de nouvelles campagnes.
- Comment **mesurer la qualité et la fiabilité** (capacité à traiter des nouveaux individus) **de votre modèle**.

Afin de vous permettre de répondre au mieux à ces questions, plusieurs solutions s'offrent à vous.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4.6 Vos solutions

Pour sélectionner les individus à qui envoyer un courrier, vous avez plusieurs solutions. Vous pouvez utiliser :

- une **méthode globale**,
- une **méthode intuitive**,
- une **méthode statistique classique** (réseaux de neurones, réseaux bayésiens, modèles logistiques, arbres de décisions, etc.),
- la **méthode InfiniteInsight**.

4.6.1 Méthode globale

Cette méthode consiste à n'effectuer aucune sélection sur votre base de données et envoyer massivement un courrier à la totalité des personnes référencées dans votre base de données. Cette solution vous garantit que toutes les personnes susceptibles d'acheter votre produit seront bien contactées.

En revanche, elle engendre un coût exorbitant, qui dépasse de loin de votre budget et est dans tous les cas rarement adoptée dans la réalité. De plus, elle risque de saturer les prospects de la banque avec des offres inadaptées *(spamming)*.

4.6.2 Méthode intuitive

Cette méthode consiste à effectuer une sélection selon votre connaissance métier, c'est-à-dire à envoyer vos courriers à des individus sélectionnés de manière intuitive dans votre base de données. Cette solution vous permet de diminuer significativement le coût de votre campagne marketing pour qu'elle rentre dans votre budget.

Cette méthode n'est pas optimale, car elle ne permet pas de :

- **maîtriser le coût réel et de retour sur investissement** de votre opération marketing.
- **baser la sélection des prospects à contacter sur un retour réel**. En effet, il est probable que vous ayez une connaissance relativement bonne des individus ayant de bonnes chances de devenir un jour vos clients, mais optimiser votre campagne consiste à pouvoir identifier les clients ayant toutes les chances de devenir clients suite à la campagne marketing en cours.
- **découvrir de nouvelles niches** de prospects, que votre connaissance du marché ne vous pas encore permis d'identifier.
- **sélectionner un nombre prédéfini d'individus**. Imaginez qu'une contrainte de votre campagne consiste à contacter 5000 prospects. Votre intuition peut vous aider à en sélectionner 2400, par exemple. En revanche, comment sélectionnez-vous ensuite les 2600 autres prospects à contacter ? Une sélection purement aléatoire, et donc totalement non optimisée, constitue alors votre seule solution.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4.6.3 Méthode statistique classique

Vous pouvez décider d'utiliser une méthode statistique "classique" pour mieux contrôler l'efficacité de votre campagne, et ainsi de votre budget.

Sur la base des informations que vous possédez, des experts en Data Mining peuvent en effet construire des modèles prédictifs. En d'autres termes, vous allez demander à un expert statisticien de créer un modèle mathématique qui vous permette de prévoir la probabilité que chaque individu a de répondre à votre campagne marketing, en fonction de son profil.

Afin de mettre en place cette méthode le statisticien doit :

- analyser en détails les résultats de votre campagne de test,
- préparer minutieusement votre base de données, notamment en encodant les différents types de données de manière à ce qu'ils soient exploitables par les outils d'analyse à utiliser,
- tester différents types d'algorithmes (réseaux de neurones, réseaux bayésiens, modèles logistiques, arbres de décisions, etc.) et sélectionner le plus adapté à votre problématique.

Après quelques semaines, l'expert-statisticien est en mesure de fournir pour chacun individu de votre base de données une probabilité d'être ou non intéressé par votre campagne marketing.

Cette méthode présente des contraintes importantes. Vous devez :

- vous assurer que l'expert statisticien, externe au Département Marketing, est disponible selon le planning fixé,
- vous assurer que le montant de ses honoraires entre bien dans votre budget,
- passer du temps à lui expliquer votre problématique métier,
- passer du temps à comprendre les résultats qu'il vous fournit.

4.6.4 Méthode InfiniteInsight

La simplicité et l'automatisation des functionnalités SAP InfiniteInsight® vont vous permettre de mettre en place vous-même l'analyse statistique de votre base de données. De plus, leur rapidité vous permette d'obtenir des résultats en seulement quelques minutes !

SAP InfiniteInsight® utilise les dernières innovations des sciences statistiques et affranchit en même temps l'utilisateur final de la complexité de la démarche associée à l'analyse statistique.

Grâce à SAP InfiniteInsight®, vous êtes en mesure de créer un modèle qui vous permet de :

- déterminer qui sont les individus qui ont la probabilité (*score*) la plus élevée d'être intéressés par votre campagne marketing (modélisation prédictive). Vous pouvez ensuite appliquer le modèle sur la totalité de votre base de données.
- mettre en évidence les facteurs déterminants qui décrivent le phénomène que vous souhaitez modéliser, c'est-à-dire le fait d'être intéressé ou pas par le nouveau produit financier de la banque (modélisation descriptive).

La courbe de profit, véritable outil de validation et de contrôle, permet de comparer la performance des modèles générés avec les fonctionnalités SAP InfiniteInsight® par rapport à celle d'un hypothétique modèle aléatoire ou à celle d'un hypothétique modèle parfait. En même temps, elle vous permet de déterminer le nombre optimal de personnes que vous devez contacter afin de maximiser le profit généré par votre campagne. SAP InfiniteInsight® vous fournit également des indicateurs sur la qualité du modèle (KI) que vous avez créé et sur sa capacité à se généraliser (KR), c'est-à-dire à rester pertinent sur de nouveaux jeux de données.

SAP InfiniteInsight® vous donne les moyens de personnaliser votre campagne de marketing direct par rapport à vos différents profils de clients, et d'augmenter ainsi son pouvoir persuasif.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4.7 Présentation des fichiers exemples

Ce guide est accompagné des fichiers de données exemples suivants :

- un fichier de données *Census01.csv*
- le fichier de description correspondant *desc_census.csv*.

Ces fichiers vous permettent d'évaluer et de faire vos premiers pas avec les fonctionnalités de SAP InfiniteInsight®.

Census.csv est le fichier de données exemple que vous allez utiliser pour suivre les scénarios des composants **InfiniteInsight® Modeler / Régression ou Classement** et **InfiniteInsight® Modeler / Segmentation**. Ce fichier est un extrait de la base de données du Bureau américain du recensement, réalisé en 1994 par Barry Becker.

Remarque

Pour plus d'informations sur le Bureau de recensement américain (*Census bureau*), Census Bureau http://www.census.gov.

Ce fichier présente des données sur 48842 individus américains, âgés au minimum de 17 ans. Chaque individu est caractérisé par 15 données. Ces données, ou variables, sont décrites dans le tableau suivant.

Variable	Description	Exemples de valeurs
age	Age des individus	toute valeur numérique supérieure à 17
workclass	Catégorie socio-professionelle des individus	▪ *Private* (salarié) ▪ *Self-employed-not-inc* (profession libérale)
fnlwgt	Variable de poids, permettant à chaque individu de représenter un pourcentage de la population	toute valeur numérique, telle que "0", "2341" ou 205019".
education	Niveau d'étude, représenté par un niveau scolaire ou par intitulé de diplôme	▪ *11th* (classe de 3ème) ▪ *Bachelors* (équivalent à un diplôme Bac+3, Licence)
education-num	Nombre d'années d'étude, représenté par une valeur numérique	une valeur numérique comprise entre 1et 16
marital-status	Situation maritale	▪ *Divorced* (divorcé) ▪ *Never-married* (jamais marié)
occupation	Profession	▪ *Sales* (profession commerciale) ▪ *Handlers-cleaners* (personnel d'entretien)
relationship	Situation familiale	▪ *Husband* (mari) ▪ *Wife* (épouse)
race	Origine ethnique	▪ *White* (blanc) ▪ *Black* (noir)
sex	Sexe	▪ *Male* (homme) ▪ *Female* (femme)
capital-gain	Gain boursier annuel	toute valeur numérique
capital-loss	Perte boursière annuelle	toute valeur numérique
native country	Pays d'origine	▪ *United States* ▪ *France*
class	Variable indiquant si le salaire d'un individu est supérieur ou inférieur à $50000	▪ "1" si l'individu a un revenu supérieur à $50000 ▪ "0" si l'individu a un revenu inférieur à $50000

Remarque

Afin de ne pas compliquer les scénarios d'utilisation de **InfiniteInsight® Modeler / Régression ou Classement** et **InfiniteInsight® Modeler / Segmentation**, la variable *fnlwgt* est utilisée comme une variable explicative quelconque dans ces scénarios, et non en tant que variable de poids.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

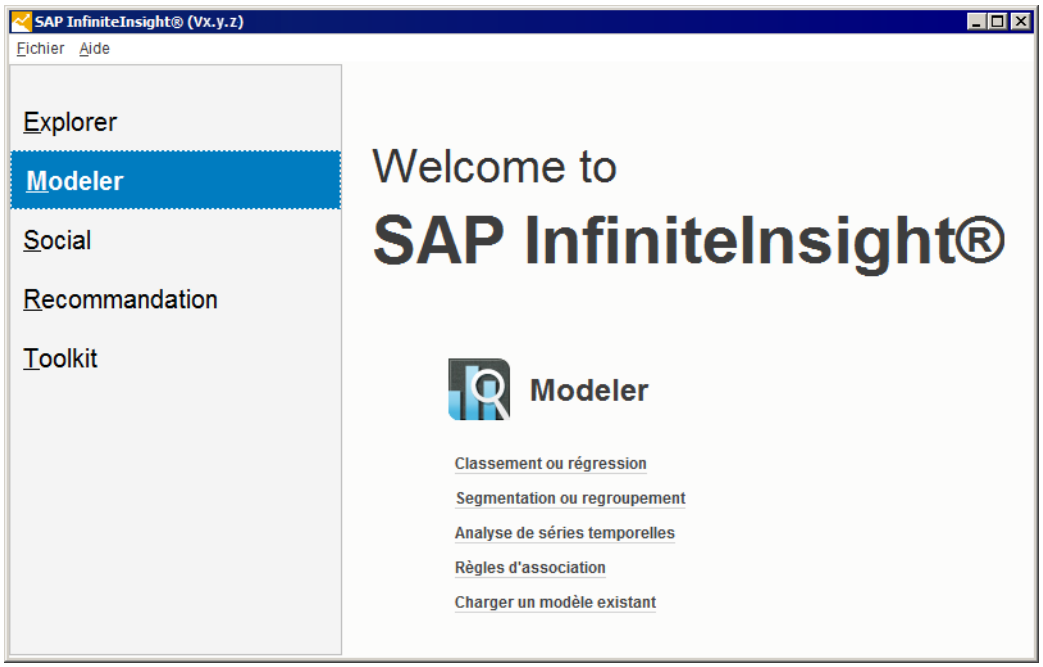
4.8 L'assistant de modélisation

Pour réaliser les deux scénarios, vous utilisez l'assistant de modélisation SAP InfiniteInsight®. Cet assistant vous permet de sélectionner la fonctionnalité avec laquelle vous souhaitez travailler, et vous assiste dans toutes les étapes de la modélisation.

Pour voir plus d'informations sur les fonctionnalités de InfiniteInsight® Modeler, voir la section Architecture et fonctionnement à la page 11.

☑ Pour démarrer l'assistant de modélisation

1 Sélectionnez *Démarrer* > *Programmes* > *SAP Business Intelligence* > *SAP SAP InfiniteInsight®* > *SAP InfiniteInsight®*
L'assistant de modélisation apparaît.



2 Cliquez sur l'action que vous souhaitez réaliser (création de modèle, exploration de données, préparation de données...).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5 Créer un modèle de classement ou de régression avec InfiniteInsight® Modeler

La modélisation de données avec InfiniteInsight® Modeler / Régression ou Classement se subdivise en quatre grandes étapes :

1. **Définition** des paramètres de modélisation
2. **Génération et validation** du modèle
3. **Analyse et compréhension** des résultats d'analyse
4. **Utilisation** du modèle généré

5.1 Etape 1 - Définir les paramètres de modélisation

Pour répondre à votre problématique, vous cherchez à :

- **identifier et comprendre** les facteurs qui déterminent qu'un prospect répond de manière positive ou négative à votre campagne de marketing.
- pouvoir ainsi **prédire** le comportement de nouveaux prospects par rapport à votre campagne.

La fonctionnalité InfiniteInsight® Modeler / Régression ou Classement vous permet de créer des modèles explicatifs et prédictifs.

La première étape du processus de modélisation consiste à définir les paramètres de modélisation, c'est-à-dire à :

1. Sélectionner une **source de données** à utiliser comme jeu de données d'apprentissage. (voir à la page 195)
2. **Décrire le jeu de données** sélectionné.
3. **Sélectionner les variables** (à la page 82) : variable(s) **cible**(s), variable de **poids**, variables **explicatives**.
4. Vérifier les **paramètres de modélisation**.
5. Définir le **degré du modèle** (voir à la page 93). Cette étape est optionnelle.
6. Définir la **valeur des catégories cibles** (voir à la page 96). Cette étape est optionnelle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.1.1 Sélectionner une source de données

Pour ce scénario

Utilisez le fichier *Census01.csv* comme jeu de données d'apprentissage.

Ce fichier représente l'échantillon que vous avez extrait de votre base de données et utilisé pour la phase de test de votre campagne de marketing direct. En accord avec votre plan de test, ce fichier contient donc des données sur 50 000 prospects, dont vous connaissez maintenant le comportement par rapport au nouveau produit financier :

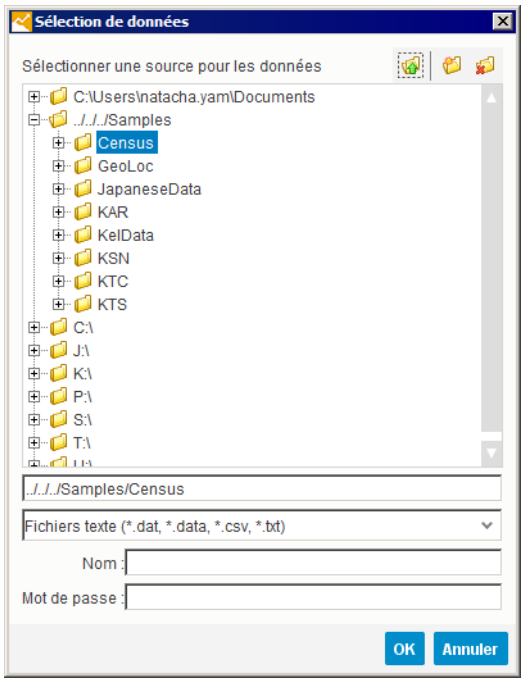
- 25% des prospects se sont montrés clairement intéressés. Ils ont décidé d'accepter un rendez-vous avec un des opérateurs de vos canaux de vente,
- 75% des prospects ont décliné votre invitation.

Dans ce fichier, vous avez créé une nouvelle variable *Class*, qui correspond à la réaction des prospects contactés pour le test. Vous avez assigné :

- la valeur "1" aux prospects ayant répondu de manière positive à votre invitation,
- la valeur "0" aux prospects ayant répondu de manière négative à votre invitation.

Pour sélectionner une source de données

1. Dans l'écran *Données à modéliser*, sélectionnez le format de la source de données à utiliser dans la liste *Type de données*.
2. Cliquez sur le bouton *Parcourir* correspondant au champ *Répertoire*.
   La fenêtre de sélection suivante apparaît.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3. Double-cliquez sur le répertoire `Samples`, puis sur le répertoire `Census.`
4. Cliquez sur le bouton *OK*.
5. Utilisez le bouton *Parcourir* correspondant au champ *Jeu de données* pour sélectionner le fichier `Census01.csv`
6. Cliquez sur *OK*.
   Le nom du fichier apparaît dans le champ *Jeu de données*.
7. Cliquez sur le bouton *Suivant*.
   L'écran *Description des données* apparaît.

8. Passez à la section **Décrire les données** (voir à la page 70).

Cas des données stockées en base de données : le mode "Explain"

Avant de demander des données stockées en base de données Oracle, Teradata ou SQL Server 2005, SAP InfiniteInsight® utilise une fonctionnalité, le **mode "Explain"**, qui classe les performances des requêtes SQL en plusieurs catégories définies par l'utilisateur. Pour plus de rapidité et de légèreté, ce classement est fait **sans que la requête SQL complète soit effectivement exécutée**.

Le but est de permettre d'estimer la charge nécessaire à l'exécution de la requête SQL et de décider --éventuellement grâce à une politique informatique interne-- si la requête SQL en question peut être utilisée ou non.

Ainsi, une politique informatique peut vouloir favoriser l'interactivité et pour cela avoir défini trois catégories de requêtes SQL, chacune ayant une durée maximale d'exécution :

- *Immédiate* : *durée < 1s*. La requête est acceptée et exécutée immédiatement.
- *Différée* : *1s <= durée < 2s*. La requête est acceptée mais ne sera exécutée que lorsque le serveur sera disponible
- *Rejetée* : *2s <= durée*. La requête ne sera jamais exécutée.

Le nombre, les appellations et les limites des catégories sont définies par l'utilisateur afin que ces valeurs correspondent à la configuration du SGBD et à sa politique d'utilisation.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

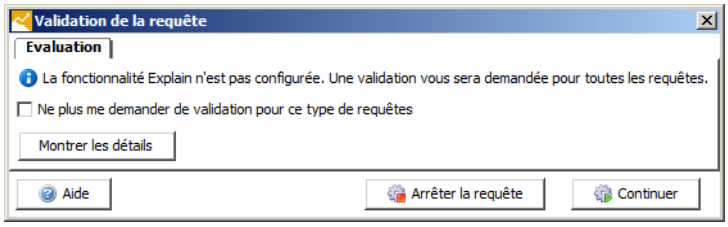
Le mode "Explain" a été configuré

Si le mode "Explain" a été configuré par votre administrateur de bases de données, une requête peut avoir deux résultats :

- **la requête a été acceptée et exécutée** : dans ce cas, le traitement de la requête est transparent pour l'utilisateur, SAP InfiniteInsight® accède aux données directement sans intervention supplémentaire de la part de l'utilisateur.
- **la requête doit être validée avant d'être exécutée** : une fenêtre s'ouvre affichant un message configuré par l'administrateur de bases de données. Une requête nécessitant une validation peut être classifiée de deux façons :
  -  **moyenne**
    Vous devrez probablement vérifier auprès de votre administrateur de bases de données quelle option choisir :
- Si l'administrateur autorise l'exécution de la requête, cliquez sur le bouton *Continuer*. La fenêtre de message se ferme et l'action demandée s'exécute.
- Si l'administrateur n'autorise pas l'exécution de la requête, cliquez sur le bouton *Arrêter la requête*. La fenêtre de message se ferme et aucune action n'est effectuée.
  -  **lourde**
    Cela signifie que la requête prendra trop de temps et de ressources sur le serveur. Dans ce cas, le fonctionnement du bouton *Continuer* dépend de la configuration faite par l'administrateur de bases de données (qui peut, par exemple, rejeter automatiquement les requêtes trop lourdes). Dans tous les cas, vous devez vérifier auprès de lui quelle action effectuer.

Le mode "Explain" n'a pas été configuré

Si votre administrateur de bases de données n'a pas configuré le mode "Explain", la fenêtre de message suivante s'affiche lorsque vous essayez d'accéder aux données :



Vous devez contacter votre administrateur qui vous dira quelle est la marche à suivre et qui configurera le mode "Explain".

Si l'administrateur valide l'exécution de votre requête, vous pouvez vouloir que toutes les requêtes nécessitant le même temps (ou un temps inférieur) soient exécutées sans être validées. Dans ce cas, cochez la case *Ne plus me demander de validation pour des requêtes similaires*. La fenêtre de validation n'apparaîtra que pour des requêtes nécessitant plus de ressources. Cette configuration du mode "Explain" n'est valide que pour la session courante. Pour une configuration définitive, contactez votre administrateur de bases de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

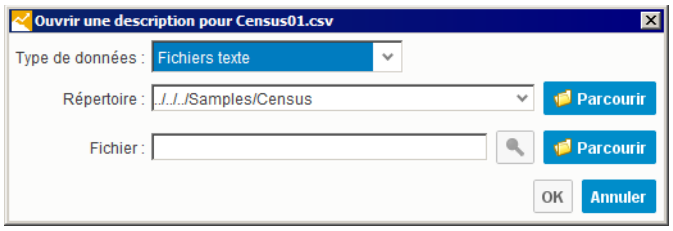
5.1.2 Décrire les données sélectionnées

Pour ce scénario

- Sélectionnez *Fichiers texte* comme type de source de données.
- Utilisez le fichier de description existant `desc_Census01.csv`, correspondant au fichier de données `Census01.csv`.

Pour utiliser un fichier de description existant

1 Dans l'écran *Description des données*, cliquez sur le bouton *Ouvrir*. La fenêtre *Ouvrir une description* s'affiche.



2 Sélectionnez le type de votre source de données dans la liste en haut à droite.

3 Utilisez le bouton *Parcourir* du champ *Répertoire* pour sélectionner le répertoire ou la base de données contenant la source de données.

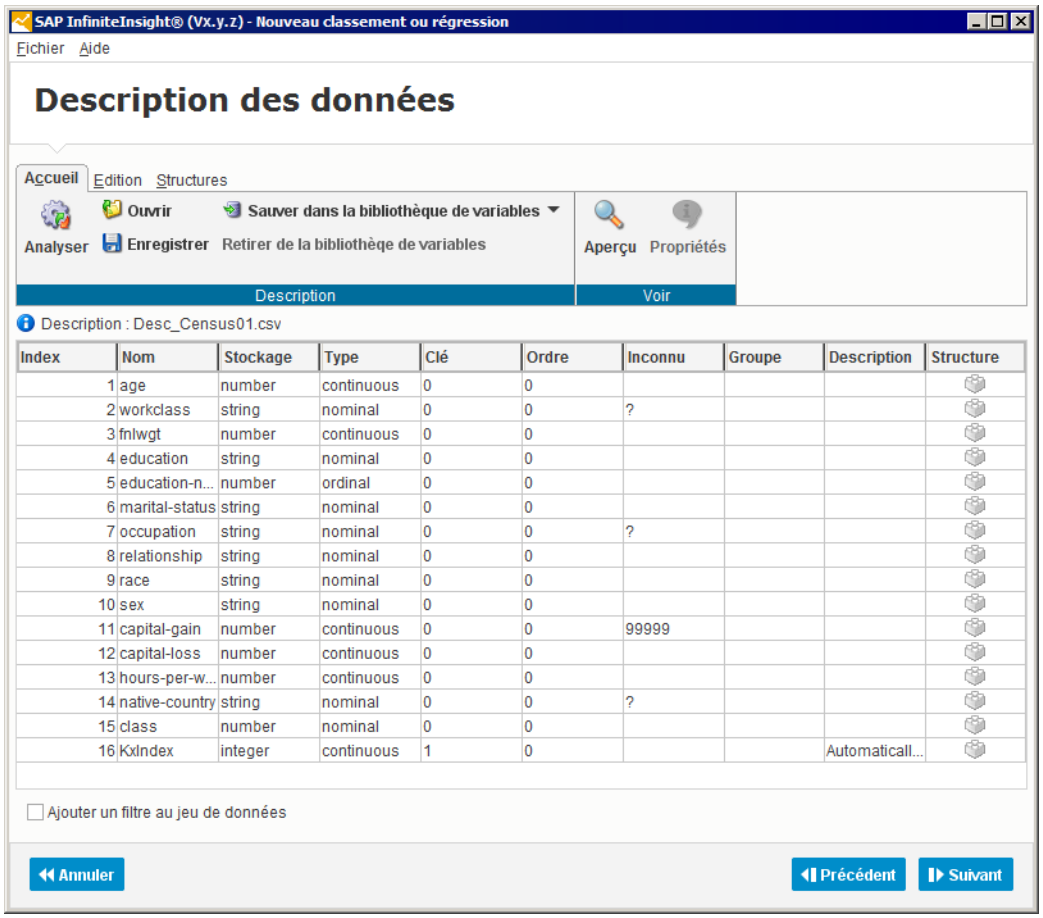
> **Note**
>
> Le répertoire sélectionné par défaut est le même que celui sélectionné à l'étape précédente.

4 Utilisez le bouton *Parcourir* du champ *Fichier* pour sélectionner le fichier ou la table contenant les données.

> **Attention**
>
> Quand l'espace de données utilisé pour la construction du modèle contient une variable physique appelée *KxIndex*, il n'est pas possible d'utiliser un fichier de description ne comportant aucune clé pour l'espace de données courant.
>
> Quand l'espace de données utilisé pour la construction du modèle ne contient pas de variable nommée *KxIndex*, il n'est pas possible d'utiliser un fichier de description incluant une description à propos d'une variable *KxIndex* car cette variable n'existe pas dans l'espace de donnée courant.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5 Cliquez sur le bouton *OK*. La fenêtre *Ouvrir une description* se ferme et la description des données s'affiche dans la fenêtre principale.



Index	Nom	Stockage	Type	Clé	Ordre	Inconnu	Groupe	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

6 Cliquez sur le bouton *Suivant*.
L' écran *Sélection des variables explicatives* apparaît.

7 Passez à la section Sélectionner les variables explicatives.

☑ Pour créer un fichier de description

1 Dans l'écran *Description des données*, cliquez sur le bouton *Analyser*.
La description des données apparaît.

2 Vérifiez l'exactitude de la description obtenue.
Si votre fichier de données initial contient des variables qui ont fonction de clés, elles ne sont pas reconnues automatiquement. Décrivez-les manuellement.

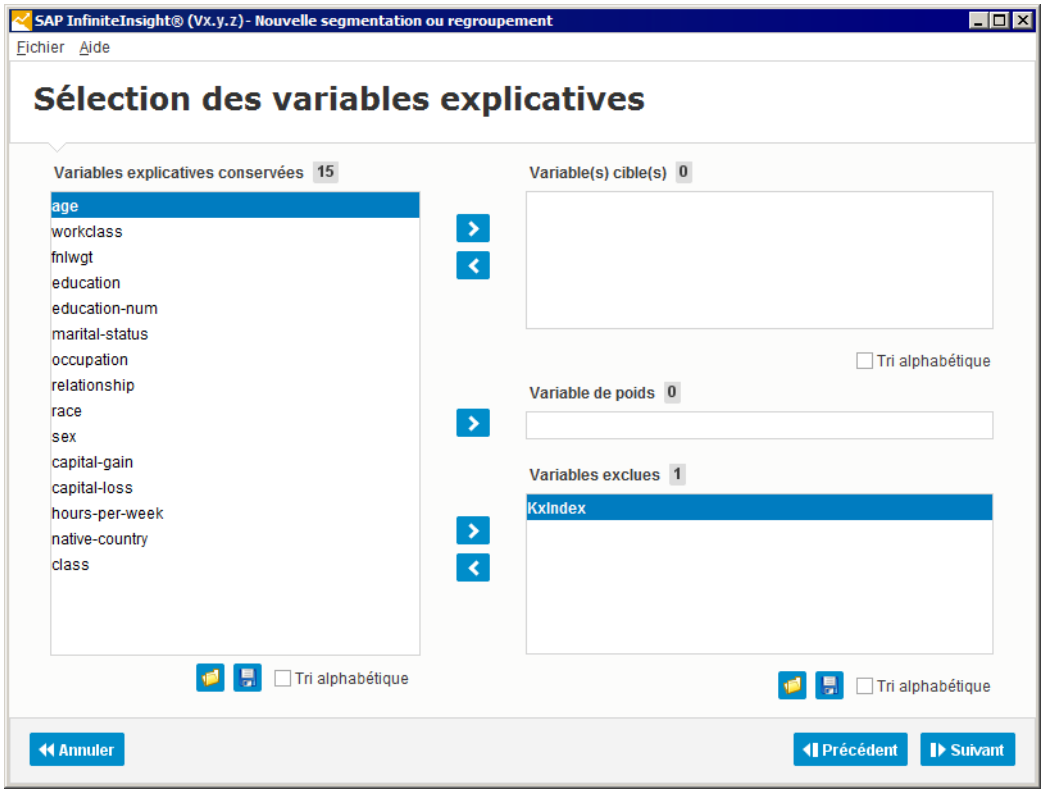
Attention

L'espace de données source utilisé, qu'il s'agisse d'une fichier texte ou d'une base de données ODBC, doit contenir au minimum une variable clé.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Une fois la description des données validée, vous pouvez :
- la sauvegarder en cliquant sur le bouton *Enregistrer*.
- cliquer sur le bouton *Suivant* pour passer à l'étape suivante.

L'écran *Sélection des variables explicatives* apparaît.

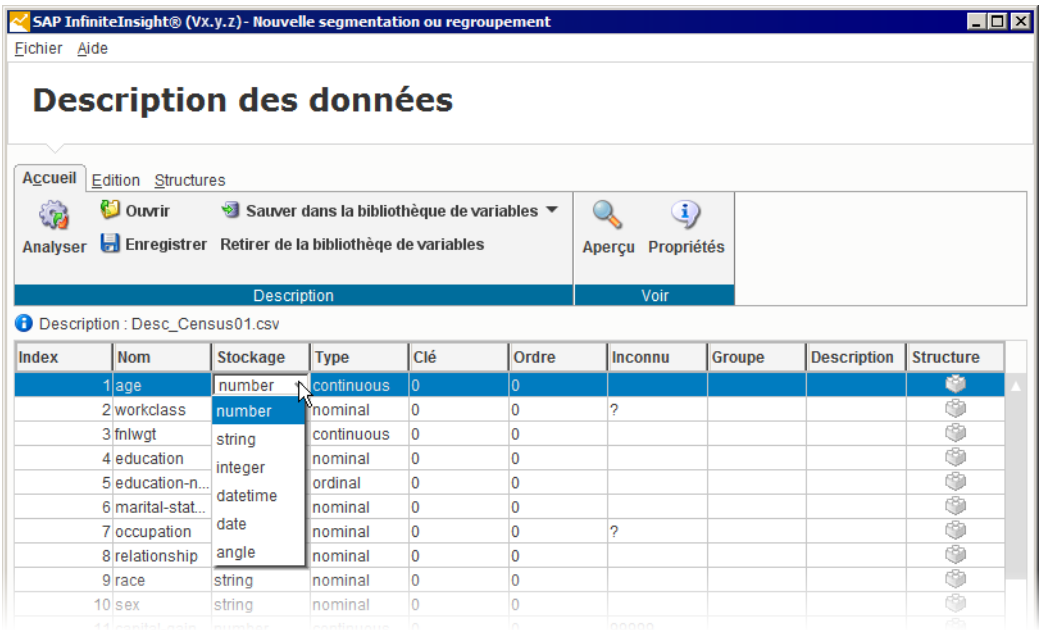


© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4 Passez à la section Sélectionner les variables explicatives.

☑ Pour modifier la description des données

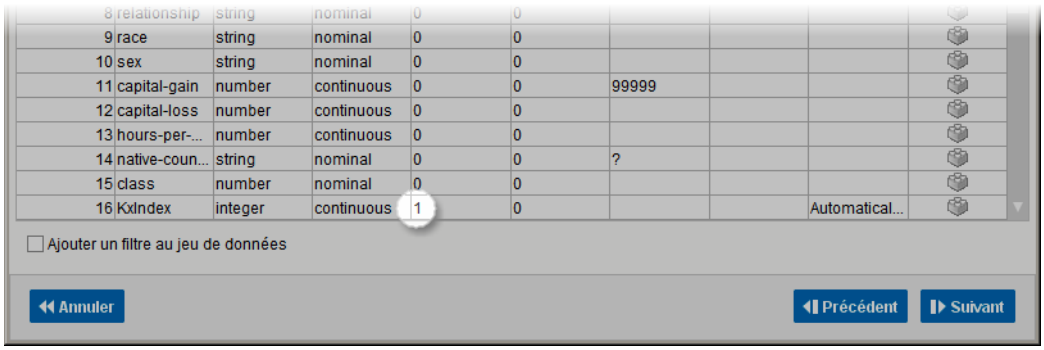
1 Dans la colonne de votre choix, par exemple la colonne *Stockage*, cliquez sur la case que vous souhaitez modifier.
La liste des valeurs possibles apparaît.



2 Sélectionnez la valeur souhaitée dans la liste.

☑ Pour spécifier qu'une variable est une clé

1 Dans la colonne *Clé*, cliquez sur la case correspondant à la ligne de la variable clé.
2 Entrez la valeur "*1*" pour définir cette variable comme clé.



© 2014 SAP SE or an SAP affiliate company. All rights reserved.

> **Note**
>
> Chaque modèle doit contenir une clé, c'est-à-dire qu'une ou plusieurs variables avec un champ clé ayant une valeur de clé différente de zéro. Si aucune clé n'a été détectée pendant le processus d'analyse et qu'aucune variable physique nommée *KxIndex* n'existe dans l'espace de données source, il est impossible d'ajouter une variable appelée *KxIndex* avec sa description. Une variable virtuelle ne peut pas être décrite.
>
> Dans ce cas particulier, en effet, les composants applicatifs de SAP InfiniteInsight® génèrent une variable-clé virtuelle nommée *KxIndex* et une description est ajoutée par les composants applicatifs InfiniteInsight® dans la colonne *Description* : 'Automatically added'.

Pourquoi décrire les données sélectionnées

Pour que vos données soient interprétables et analysables par les fonctionnalités SAP InfiniteInsight®, elles doivent être décrites. En d'autres mots, le fichier de description spécifie la nature de chaque variable en déterminant leur :

- format de **stockage** : nombre (*number*), chaînes de caractère (*string*), date et heure (*datetime*) ou date (*date*).

> **Note**
>
> Lorsqu'une variable est déclarée comme date (*date* ou *datetime*), la fonctionnalité <FR_KDC> (*KDC*) en extrait automatiquement des informations spécifiques telles que le jour du mois, l'année, le trimestre, etc. Des variables contenant ces informations sont créées lors de la génération du modèle et sont utilisées comme variables d'entrée. KDC est activé pour toutes les fonctionnalités SAP InfiniteInsight® à l'exception de InfiniteInsight® Modeler / Séries temporelles (*KTS*).

- **type** : variables continues (*continuous*), nominales (*nominal*) ordinales (*ordinal*) ou textuelle (*textual*).

> **Note**
>
> Toutes les variables décrites doivent se trouver dans la source de données utilisée pour l'apprentissage. Dans le cas où une variable physique décrite n'existe pas dans la source de données, il n'est pas possible de générer un modèle.

Pour plus d'informations sur la description des données, Types de variables à la page 27 et Formats de stockage à la page 30.

> **Note**
>
> La traduction des catégories d'une variable n'a pas d'influence sur sa structure qui doit être définie en fonction des valeurs initiales de la variable.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comment décrire les données sélectionnées

Pour décrire vos données, vous pouvez :

- soit **utiliser un fichier de description existant**, c'est-à-dire issu de votre système d'information ou d'une précédente utilisation des fonctionnalités SAP InfiniteInsight®,
- soit **créer un fichier de description grâce à l'option** *Analyser*, mise à votre disposition dans l'assistant de modélisation SAP InfiniteInsight®. Dans ce cas, vous devez valider le fichier de description obtenu. Vous pouvez sauvegarder ce fichier pour une utilisation ultérieure.

Attention

Le fichier de description obtenu avec l'option *Analyser* résulte de l'analyse des 100 premières lignes du fichier de données initial. Afin d'éviter tout biais, n'hésitez pas à brasser votre jeu données avant de l'analyser.

Le scénario d'utilisation standard [ouverture d'un espace de donnée ODBC - description en utilisant la fonction d'*Analyse* - génération du modèle] ne peut pas être mis en oeuvre lorsque l'espace de données source contient une variable nommée *KxIndex* mais aucune variable ODBC ayant le statut de clé.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

La description d'une variable est composée des champs décrits dans le tableau ci-dessous :

Le champ...	contient...
Nom	le nom de la variable (celui-ci ne peut être modifié)
Stockage	le type de valeurs stockées dans cette variable : ▪ *Number* : la variable contient uniquement des nombres "caculables" (attention : les numéros de téléphone, codes postaux, numéros de compte ne doivent pas être considérés comme des nombres) ▪ *String* : la variable contient des chaînes de caractères. ▪ *Datetime* : la variable contient des dates et des heures ▪ *Date* : la variable contient des dates
Type	le type de la variable : ▪ *Continuous* : une variable numérique pour laquelle la moyenne, la variance, etc. peuvent être calculées. ▪ *Nominal* : variable catégorique, seul type possible pour une chaîne de caractère (les codes postaux, numéros de téléphone, etc. sont généralement de ce type). ▪ *Ordinal* : variable numérique discrète pour laquelle l'ordre est important ▪ *Textual* : variable textuelle contenant des mots, des phrases ou des textes complets. Attention - lors de la création d'un modèle d'analyse textuelle, si aucune variable textuelle n'est définie le bouton *Suivant* est désactivé et il est impossible de passer à l'étape suivante.
Clé	indique si cette variable est une clé ou un identifiant pour l'observation : ▪ *0* la variable l'est pas un identifiant; ▪ *1* clé primaire; ▪ *2* clé secondaire...
Ordre	indique si la variable représente un ordre naturel. Dans un jeu de données d'évènements il doit y avoir au moins une variable marquée comme ordonnée. Attention - si la source de données est un fichier et que la variable marquée comme représentant un ordre naturel n'est pas effectivement ordonnée, un message d'erreur s'affichera au moment de la vérification ou de la génération du modèle.
Inconnu	la chaîne de caractères utilisée dans le fichier de description pour représenter les valeurs manquantes (par exemple "999" ou "#Vide" - sans les guillemets)
Groupe	le nom du groupe auquel appartient la variable. les variables appartenant à un même groupe sont considérées comme apportant la même information et ne seront donc pas croisées dans les modèles d'ordre supérieur à 1. Ce paramètre sera activé dans une future version.
Description	une éventuelle description supplémentaire de la variable
Structure	structure de la variable, c'est-à-dire les groupements des catégories des variables.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Des données redondantes

Certaines informations de la base de données sont redondantes, telles que le "diplôme" et "niveau de formation", ou le "diplôme" et "métier".

Dans le domaine des statistiques, le terme "variables corrélées" est utilisé pour désigner de telles données. Dans toutes analyses statistiques classiques, les variables corrélées doivent faire l'objet d'un traitement particulier. Une autre solution consiste à ne conserver pour l'analyse que l'une des variables sur deux variables corrélées.

N'ayant ni les compétences statistiques ni les moyens pour traiter ce problème de corrélations entre variables, vous décidez de conserver la base de données en l'état.

Un mot sur les clés de base de données

Pour des raisons de gestion des données et de performance, le jeu de données à analyser doit comporter une variable ayant fonction de clé. Deux cas se présentent :

- Si le jeu de données initial ne contient **pas de variable clé**, une variable index *KxIndex* est automatiquement créée par les fonctionnalités SAP InfiniteInsight®. Elle correspondra au numéro de la ligne de données traitée.

  > **Note**
  >
  > Il n'est pas possible de forcer l'indice de clé (Key Level) à 0 pour une clé virtuelle si aucune autre clé n'a été définie.

- Si le fichier contient **une ou plusieurs variables clés**, ces dernières ne sont pas automatiquement reconnues. Vous devez alors le spécifier manuellement dans la description des données en renseignant l'indice de clé à la valeur appropriée. Se reporter à la procédure **Pour spécifier qu'une variable est une clé**. Par ailleurs, si vos données sont stockées dans une base de données, elles seront automatiquement reconnues.

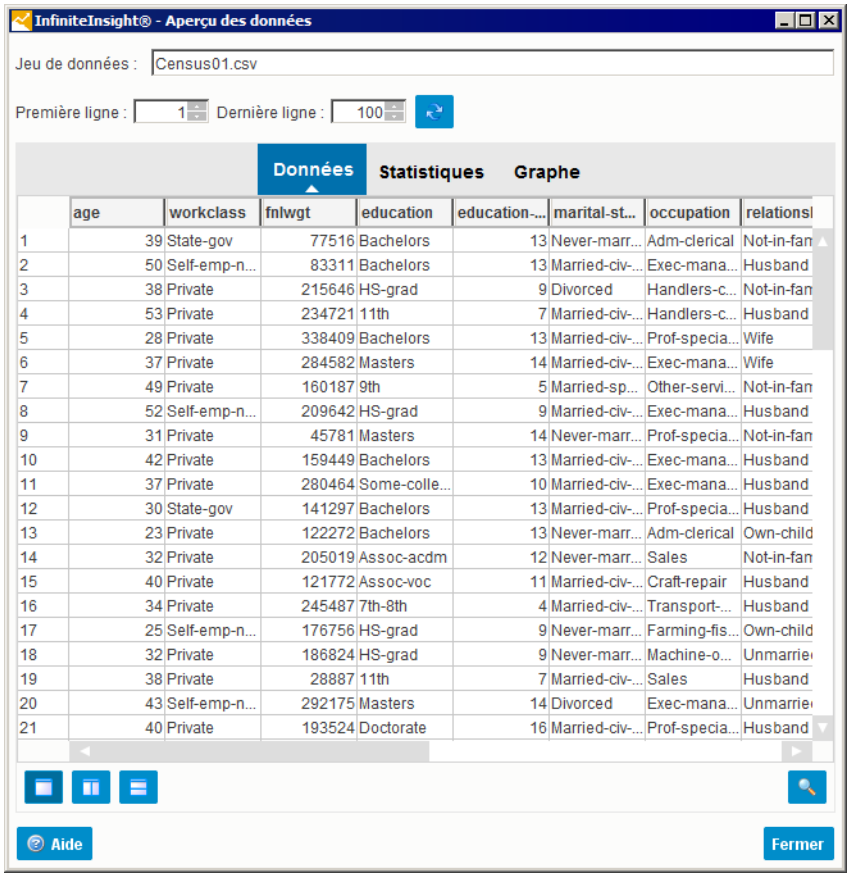
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Voir les données

Pour vous aider à valider la description obtenue par analyse, vous pouvez afficher le contenu de votre jeu de données.

☑ Pour voir les données

1 Cliquez sur le bouton *Aperçu*. Une nouvelle fenêtre s'ouvre affichant les cent premières lignes du jeu de données.



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	23	Private	122272	Bachelors	13	Never-marr...	Adm-clerical	Own-child
14	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
15	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
16	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
17	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
18	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
19	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
20	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
21	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband

2 Dans le champ *Première ligne*, saisissez le numéro de la première ligne à afficher.

3 Dans le champ *Dernière ligne*, saisissez le numéro de la dernière ligne à afficher.

4 Cliquez sur le bouton  (*Rafraîchir)* pour afficher les lignes sélectionnées.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.1.3 Ajouter un filtre au jeu de données

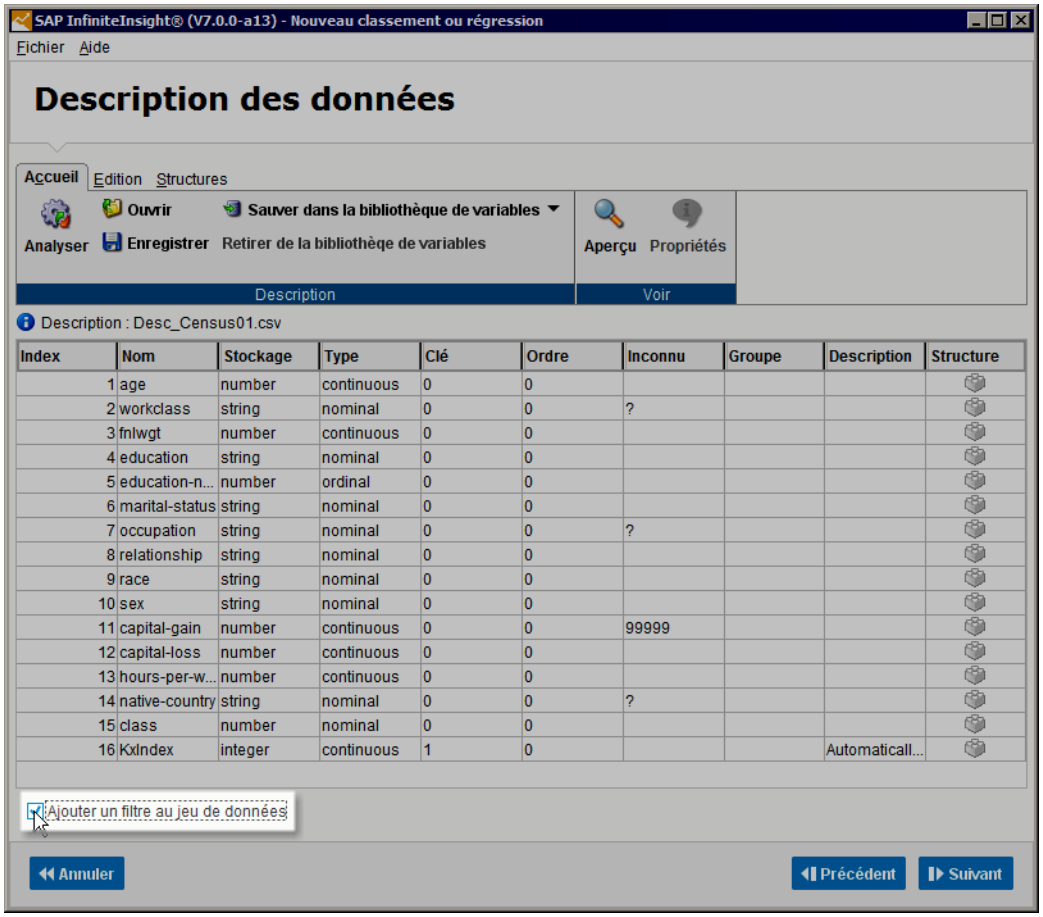
Vous avez la possibilité d'appliquer un filtre à votre jeu de données afin d'accélérer le processus d'apprentissage et d'optimiser le modèle qui en résulte.

Pour ce scénario

N'utilisez pas de filtre pour votre jeu de données.

Ajouter un filtre

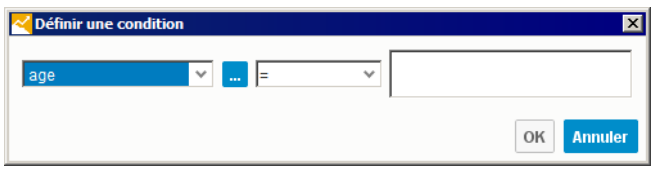
1 Cochez la case *Ajouter un filtre au jeu de données*.



2 Cliquez sur *Suivant*.

Ajouter une condition

1 Cliquez sur le bouton *Ajouter une condition*.
La fenêtre *Définir une condition* s'ouvre.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 Choisissez une variable dans la première liste déroulante.
3 Choisissez un opérateur dans la deuxième liste.
4 Indiquez une valeur dans la troisième liste :
   - Pour une variable du type *Number* entrez une valeur.
   - Pour une variable du type *String* choisissez une variable dans la liste. Si cette liste est vide, cliquez sur le bouton  pour extraire les catégories.
5 Cliquez sur *OK*.

> **Note**
>
> Vous pouvez modifier une condition en double-cliquant dessus.

☑ Ajouter une conjonction logique

Cliquez sur le bouton *Ajouter un "ET" logique* ou sur le bouton *Ajouter un "OU" logique*.

> **Note**
>
> Vous pouvez modifiez le type de conjonction en double-cliquant dessus.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Changer l'ordre

Vous pouvez changer l'ordre des noeuds pour accélérer l'application du filtre en mettant les conditions, qui ont une grande probabilité de s'avérer fausse, en haut de la liste.

1 Sélectionnez le noeud que vous voulez déplacer vers le haut ou vers le bas.

2 Utilisez les boutons et pour changer sa position dans le filtre.

☑ Supprimer un noeud

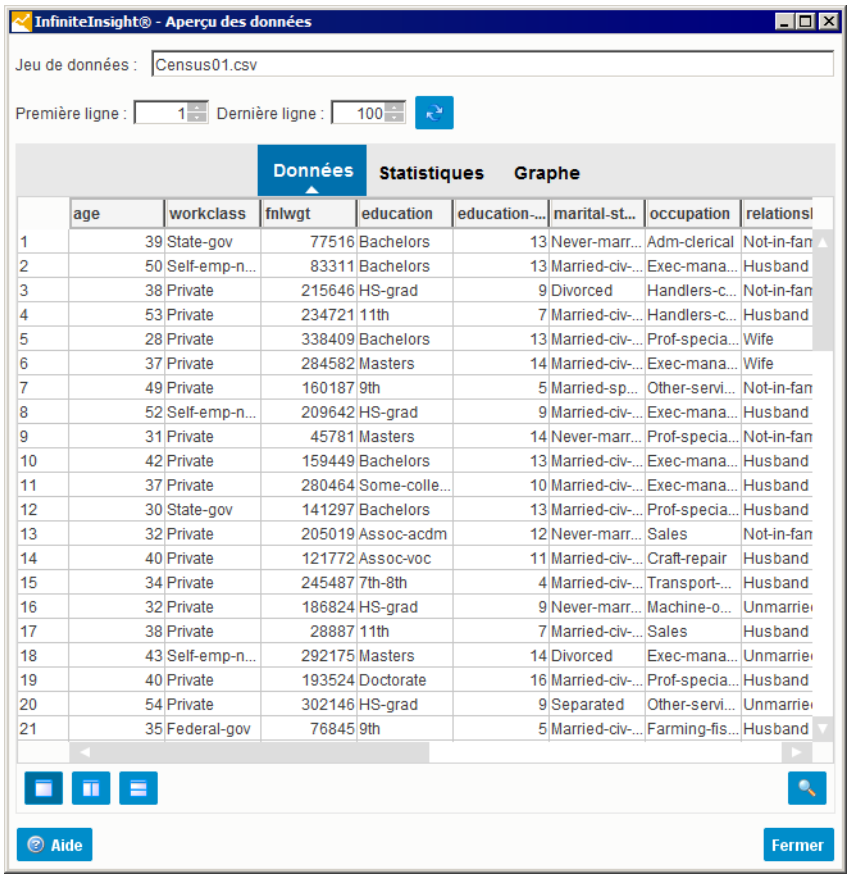
1 Sélectionnez le noeud que vous voulez supprimer.

2 Cliquer sur le bouton *Supprimer le noeud sélectionné*.

☑ Afficher le jeu de données filtré

Vous pouvez visualiser le jeu de données qui vous obtiendrez en appliquant le filtre.

Cliquez sur le bouton *Aperçu*.
Une nouvelle fenêtre s'ouvre.



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
14	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
15	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
16	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
17	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
18	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
19	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband
20	54	Private	302146	HS-grad	9	Separated	Other-servi...	Unmarrie
21	35	Federal-gov	76845	9th	5	Married-civ-...	Farming-fis...	Husband

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Enregistrer un filtre

Vous pouvez enregistrer le filtre créer pour le réutiliser ultérieurement sans être obligé de recréer un filtre avec les mêmes conditions.

1. Cliquez sur le bouton *Enregistrer ce filtre*.
   La fenêtre *Enregistrer ce filtre* s'ouvre.
2. Dans la liste *Type de données*, sélectionnez le format de l'enregistrement.
3. Utilisez le bouton *Parcourir* à droite du champ *Répertoire* pour choisir un répertoire ou une base de données pour l'enregistrement.
4. Dans le champ *Description*, entrez le nom du fichier ou de la table.
5. Cliquez sur *OK*.

☑ Charger un filtre existant

Pour filtrer un jeu de donnée, vous pouvez utiliser un filtre préalablement créé avec SAP InfiniteInsight® pour ce jeu de données.

1. Cliquez sur le bouton *Charger un filtre existant*.
   La fenêtre *Charger un filtre existant* s'ouvre.
2. Utilisez la liste déroulant Type de données pour sélectionner le format du filtre.
3. Utilisez le bouton *Parcourir* à droite du champ *Répertoire* pour choisir le répertoire ou la base de données où se trouve le filtre.
4. Utilisez le bouton *Parcourir* à droite du champ *Description* pour choisir le fichier ou la table contenant le filtre.
5. Cliquez sur *OK*.

5.1.4 Sélectionner les variables

Une fois le jeu de données d'apprentissage et sa description chargés, vous devez sélectionner :

- la ou les variables à utiliser comme variables cibles (voir "Sélectionnez les variables cibles" à la page 83),
- éventuellement une variable de poids (voir "Sélectionner la variable de poids" à la page 84),
- les variables explicatives (voir "Sélectionner les variables explicatives" à la page 86).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

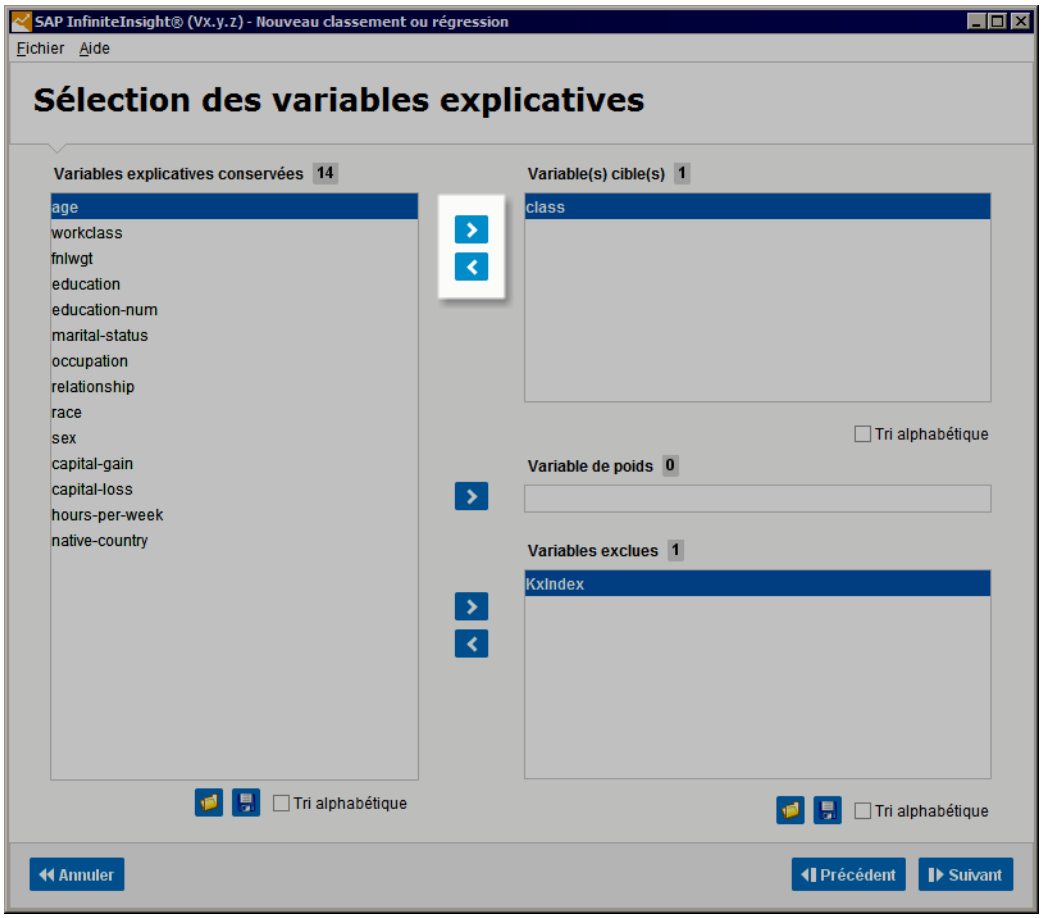
Sélectionnez les variables cibles

Pour ce scénario

Sélectionnez pour **variable cible** la variable *Class*, c'est-à-dire la variable indiquant la probabilité d'un individu à répondre de manière positive ou négative à votre campagne.

Pour sélectionner la variable cible

1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez la ou les variables choisies comme cibles.



> **Remarque**
>
> Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

2 Cliquez sur le bouton > situé gauche du champ *Variable(s) cible(s)*.
Les variables sélectionnées passent dans la partie *Variable(s) cible(s)*.

3 Pour retirer une ou plusieurs variables de la liste des variables cibles, sélectionnez celles-ci dans la liste puis cliquez sur le bouton <.

4 Passez à la section Sélectionner la variable de poids (à la page 84).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Sélectionner la variable de poids

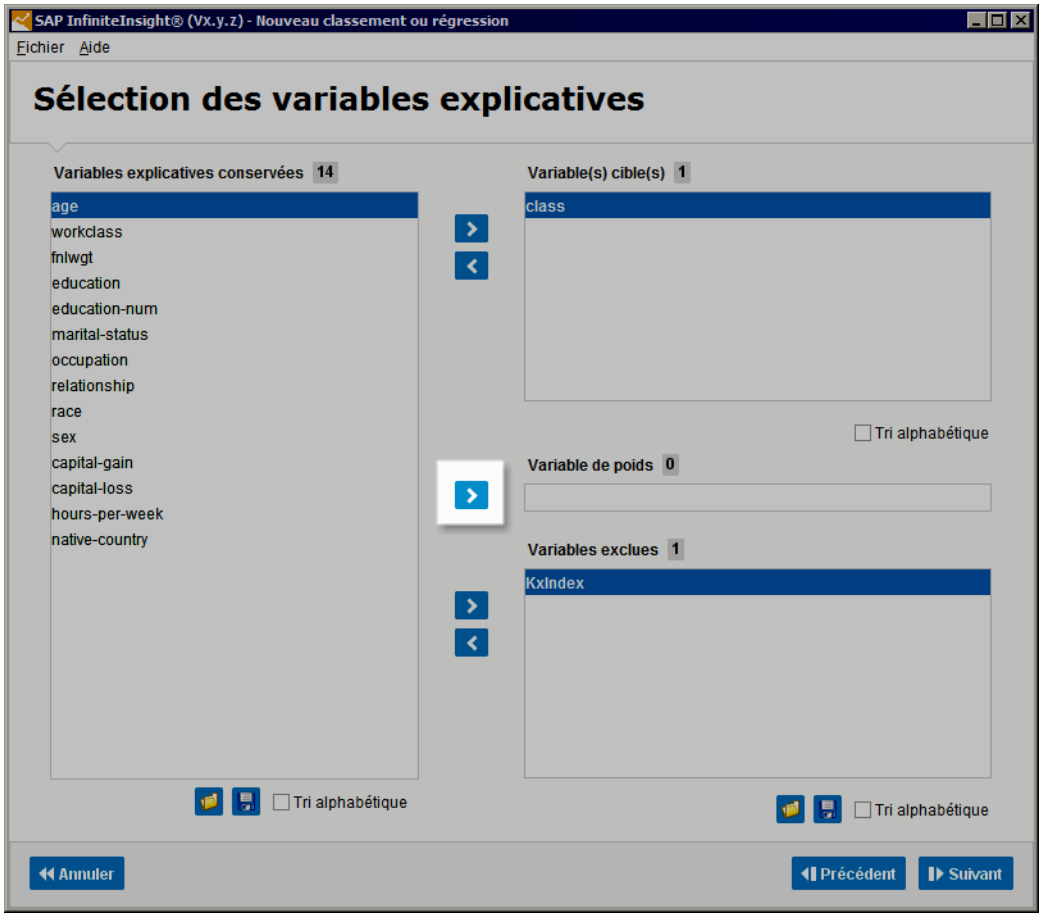
La sélection d'une variable de poids vous permet d'utiliser l'option **Poids de référence** dans les **Paramètres avancés du modèle**.

Pour ce scénario

Ne sélectionnez **aucune** variable de poids.

Pour sélectionner une variable de poids

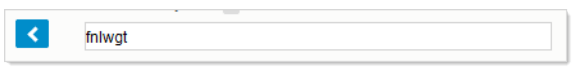
1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez la variable à utiliser comme variable de poids.



> **Remarque**
>
> Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

2 Cliquez sur le bouton > situé gauche du champ *Variable de poids*.
La variable passe dans le champ *Variable de poids*.

3 Pour supprimer la variable de poids, cliquez sur le bouton <.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

4 Passez à la section Sélectionner les variables explicatives (à la page 86).

Sélectionner les variables explicatives

Par défaut, et à l'exception des variables clés, toutes les variables contenues dans votre jeu de données sont prises en compte pour la génération du modèle. Vous pouvez exclure certaines de ces variables.

Pour la première analyse de votre jeu données, il est conseillé de **conserver toutes les variables**. Il est notamment important de conserver les variables qui n'ont à priori aucun impact sur la variable cible. Si ces variables n'ont aucun impact sur la variable cible, le modèle le confirmera. A l'opposé, le modèle vous permettra de découvrir des corrélations entre ces variables et la variable cible. Exclure des variables de l'analyse sur simple intuition présente le risque de se priver d'une forte valeur ajoutée des modèles SAP InfiniteInsight®: la découverte d'**information non intuitive**.

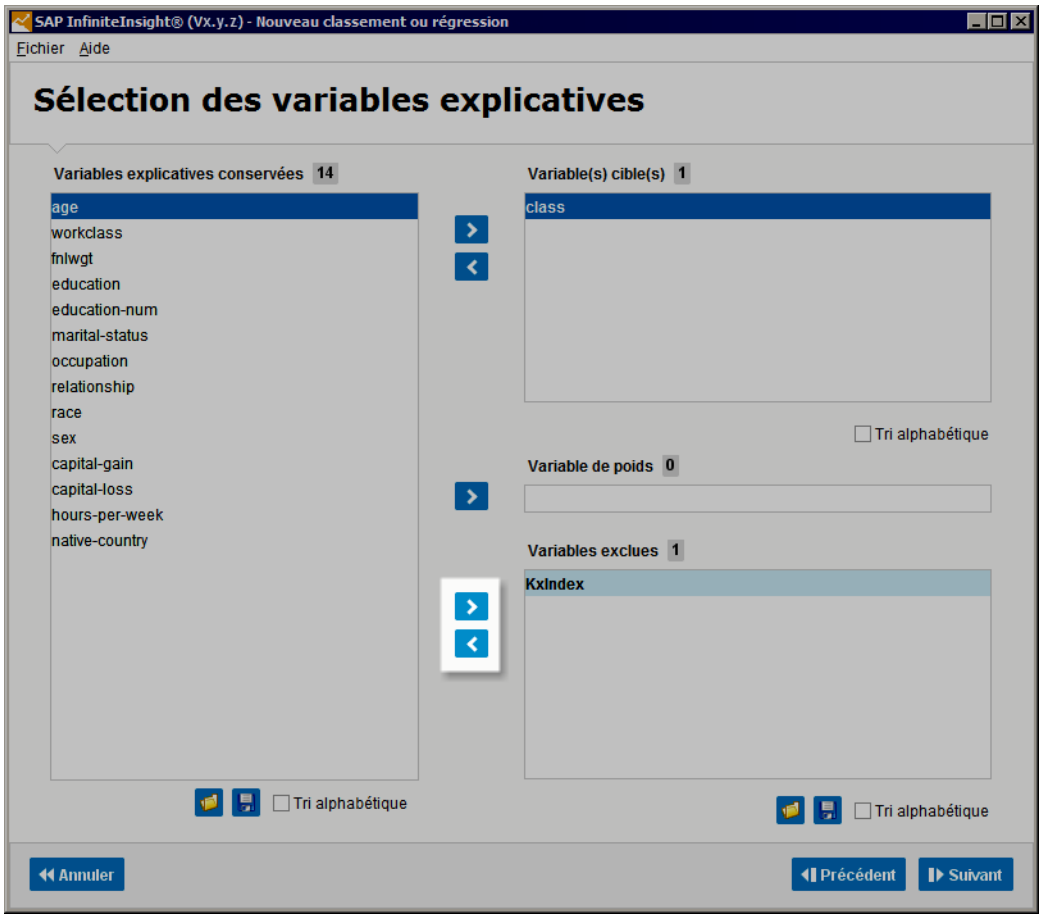
En fonction des résultats obtenus avec une première analyse incluant la totalité des variables du jeu de données, vous pouvez générer un second modèle en excluant les variables trop corrélées à la variable cible. Une fonctionnalité prévue à cet effet est proposée dans le menu d'utilisation du modèle.

Pour ce scénario

- Laissez la variable *KxIndex* exclue. Cette variable est une variable clé. Le jeu de données initial ne contenant pas de variable clé, les composants SAP InfiniteInsight® ont généré automatiquement la variable *KxIndex*.
- Conservez toutes les autres variables.

Pour exclure des variables de l'analyse des données

1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez les variables à exclure.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 Cliquez sur le bouton > situé gauche du champ *Variables exclues*.
Les variables sélectionnées passent dans la partie *Variables exclues*.

3 Pour retirer une ou plusieurs variables de la liste des variables exclues, sélectionnez celles-ci dans la liste des variables exclues, puis cliquez sur le bouton <.

> **Note**
>
> Par défaut, toute variable définie comme clé est exclue automatiquement : elle figure dans la section *Variables Exclues*. Cependant, l'utilisateur a la possibilité de déplacer une variable clé dans la section *Variables Explicatives Conservées* s'il veut que cette variable joue un tel rôle.

4 Cliquez sur le bouton *Suivant*.
L'écran *Récapitulatif des paramètres de modèlisation* apparaît.

5 Passez à la section Vérifier les paramètres de modélisation.

> **Remarque**
>
> Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

5.1.5 Traduire les catégories de variables

Vous pouvez traduire les catégories des variables nominales, enregistrer la traduction ou charger une traduction existante. Cette traduction n'influence pas la structure de la variable, qui doit être définie en fonction des valeurs originales de la variable.

> **Note**
>
> La variable "Catégorie cible", utilisée par exemple dans les paramètres avancés, ne prend pas en compte une éventuelle traduction quand les valeurs possibles de cette variable sont affichées. Pour cette raison des valeurs entrées manuellement ne peuvent pas être traitées correctement, si elles ne correspondent pas aux valeurs d'origine.

☑ Traduire les catégories de variables

1 Faites un clique droit sur la variable nominale dont vous souhaitez traduire les catégories. Un menu contextuel est affiché.

2 Sélectionnez l'option *Traduire les catégories de <nom_de_la_variable>*.

3 Choisissez dans quelles langues vous voulez traduire. Par défaut, la langue de l'interface utilisateur est affichée comme colonne.

4 Cliquez sur le bouton pour extraire les catégories de variables du jeu de données.

5 Traduisez les catégories.

> **Note**
>
> Vous n'êtes pas obligé de renseigner tous les champs.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

6 Cliquez sur *OK*.

☑ Enregistrer la traduction des catégories

1 Traduisez les catégories de variables comme expliqué ci-dessus.
2 Cliquez sur le bouton *Enregistrer*.
3 Choisissez un *Type de données*.
4 Sélectionnez un *Répertoire*.
5 Entrez un *Nom* pour le fichier ou la table.
6 Cliquez sur *OK*.

☑ Charger une traduction existante

1 Faites un clique droit sur une variable nominale. Un menu contextuel est affiché.
2 Sélectionnez l'option *Traduire les catégories de <nom_de_la_variable>*.
3 Cliquez sur le bouton *Charger*.
4 Sélectionnez le format de la traduction dans la liste *Type de données*.
5 Utilisez le bouton *Parcourir* situé à droite du champ *Répertoire* pour choisir le répertoire ou la base de données contenant la traduction.
6 Utilisez le bouton *Parcourir* situé à droite du champ *Table ou fichier* pour choisir la traduction des catégories de variables.
7 Cliquez sur le bouton *OK*.
8 Cliquez sur le bouton  *Rafraîchir* pour actualiser l'affichage des catégories.
9 Si les colonnes ne sont pas nommées correctement, utilisez les Paramètres avancés  (voir paragraphe suivant) pour choisir la ligne d'en-tête et actualisez à nouveau.
10 Mettez les noms des langues en correspondance avec les langues de la traduction chargée en cliquant sur les catégories et en choisissant la langue qui correspond dans le menu contextuel.
11 Cliquez sur le bouton *OK*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.1.6 Vérifier les paramètres de modélisation

L'écran *Récapitulatif des paramètres de modélisation* vous permet d'effectuer une dernière vérification des paramètres de modélisation avant de générer le modèle.

Note

L'écran *Récapitulatif des paramètres de modélisation* présente également un bouton *Avancé*. Ce bouton vous permet d'accéder à l'écran *Paramètres spécifiques du modèle* dans lequel vous pouvez définir des paramètres avancés tels que le degré du modèle à générer. Pour plus d'informations, voir la section suivante.

- Le **nom du modèle** est renseigné automatiquement. Il correspond au nom de la variable cible (`class` pour notre scénario), suivi du signe underscore ("_") et du nom de la source de données, sans son extension de fichier (`Census01` pour notre scénario).
- Vous pouvez afficher les résultats générés par InfiniteInsight® Modeler / Régression ou Classement sous la forme d'un arbre de décision basé sur les cinq variables les plus contributives. Pour activer cette option, cochez la case *Calculer l'arbre de décision*.
- Le bouton *Sauvegarde automatique* vous permet de spécifier que le modèle doit être automatiquement enregistré dès la fin de la génération du modèle. Les informations d'enregistrement sont paramétrables dans le panneau *Sauvegarde automatique*. Lorsque la sauvegarde automatique est activée, une coche verte s'affiche sur le bouton.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Activation de la sauvegarde automatique

Le panneau *Sauvegarde automatique* vous permet d'activer l'enregistrement automatique du modèle à la fin de ma génération et de saisir les informations nécessaires.

☑ Pour activer la sauvegarde automatique

1 Dans le panneau *Récapitulatif des paramètres de modélisation*, cliquez sur le bouton *Sauvegarde automatique*.

2 Cochez l'option *Activer la sauvegarde automatique du modèle*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Renseignez les champs décrits dans le tableau ci-dessous.

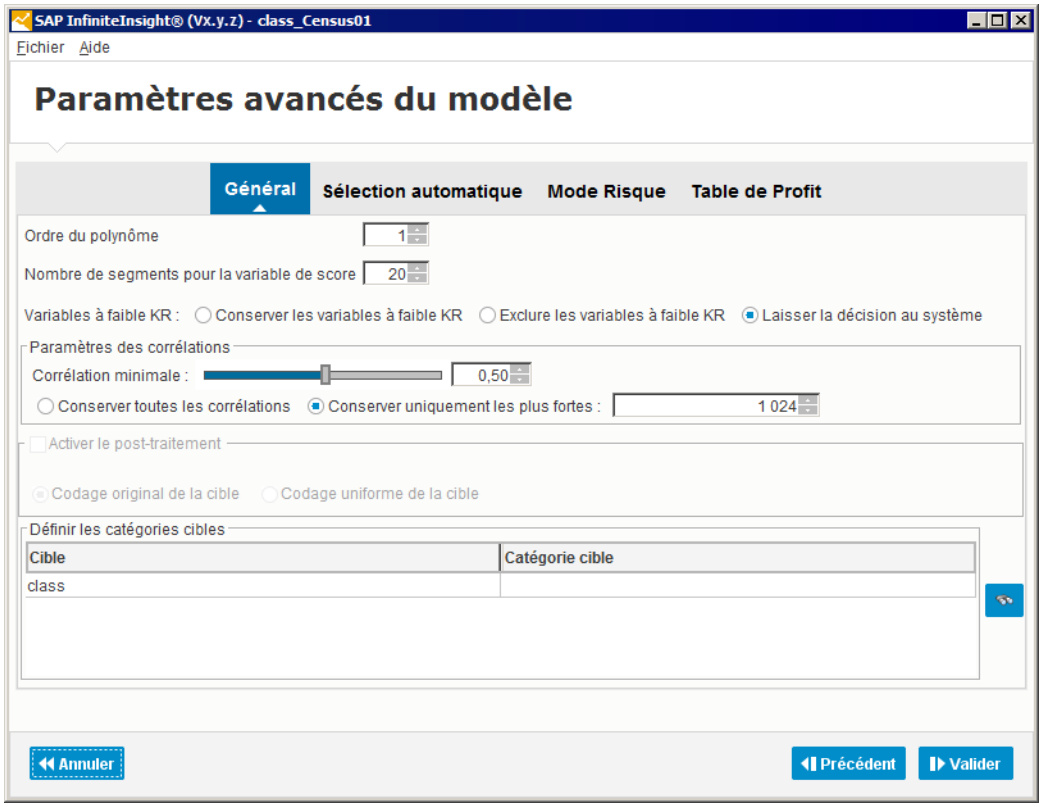
Champs	Description
Nom du modèle	Ce champ vous permet d'associer un nom au modèle. Ce nom est utilisé dans la liste des modèles qui vous est proposée quand vous chargez un modèle existant.
Description	Ce champ vous permet d'entrer des informations de votre choix, telles que le nom du jeu de données d'apprentissage utilisé, l'ordre du polynôme ou la capacité prédictive et la reproductibilité obtenus pour ce modèle. Ces informations peuvent vous être utiles ultérieurement pour identifier le modèle. Notez que cette description sera utilisée à la place de celle saisie dans le panneau *Récapitulatif des paramètres de modélisation*.
Type de données	Cette liste vous permet de sélectionner dans quel format votre modèle sera enregistré. Les formats suivants sont proposés : ▪ *Fichiers texte*, pour enregistrer le modèle dans un fichier texte, ▪ *Bases de données*, pour enregistrer le modèle dans une table ODBC, ▪ *Espace de stockage mémoire*, pour enregistrer le modèle en mémoire. Le modèle sera conservé jusqu'à la fermeture de l'interface graphique de SAP InfiniteInsight®. Notez que selon votre licence d'autres formats peuvent être disponible (comme SAS, par exemple).
Répertoire	En fonction de l'option que vous avez sélectionnée, ce champ vous permet de spécifier la source ODBC ou le répertoire dans lequel vous souhaitez enregistrer le modèle .
Fichier/Table	Ce champ vous permet d'entrer le nom du fichier ou de la table qui contiendra le modèle. Le nom de fichier doit contenir l'une des deux extensions de format `.txt` (fichier texte dans lequel les données sont séparées par des tabulations) ou `.csv` (fichier texte dans lequel les données sont séparées par des virgules).

4 Cliquez sur le bouton *Valider*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.1.7 Définir les paramètres spécifiques du modèle

Dans l'écran *Récapitulatif des paramètres de modélisation* cliquez sur le bouton *Avancé*. L'écran *Paramètres avancés du modèle* s'affiche.



Onglet "Général"

L'onglet *Général* vous permet de définir les paramètres généraux du modèle, tels que le degré du modèle, le nombre de segment de la variable de score, le nombre de corrélations à afficher, la catégorie cible de la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Définir le degré du modèle (optionnel)

Le modèle généré par InfiniteInsight® Modeler / Régression ou Classement est représenté par un polynôme. Ce polynôme peut être de degré 1, 2, 3 ou plus. En définissant l'ordre du polynôme, vous définissez le degré de complexité du modèle.

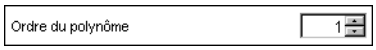
Il est fortement conseillé de toujours utiliser un ordre 1 pour la première analyse d'un jeu de données. Utiliser un ordre de polynôme élevé ne garantit pas l'obtention du modèle le plus performant dans tous les cas. Pour plus d'informations sur le paramétrage de l'ordre du polynôme, voir Méthodologie à la page 38.

Pour ce scénario

Utilisez un polynôme d'ordre 1 (valeur par défaut).

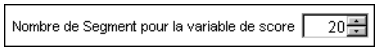
☑ Pour définir le degré de complexité du modèle

Dans l'écran *Paramètre avancés du modèle*, dans le champ Valeur de la section *Ordre du polynôme*, entrez la valeur correspondant au degré de complexité du modèle que vous souhaitez obtenir.



Définir le nombre de segments pour la variable de score

Cette option vous permet de définir le nombre de segments de score à créer. La valeur saisie doit être entre 20 et 100, en effet un nombre inférieur ou supérieur de segments nuirait à la qualité du modèle.



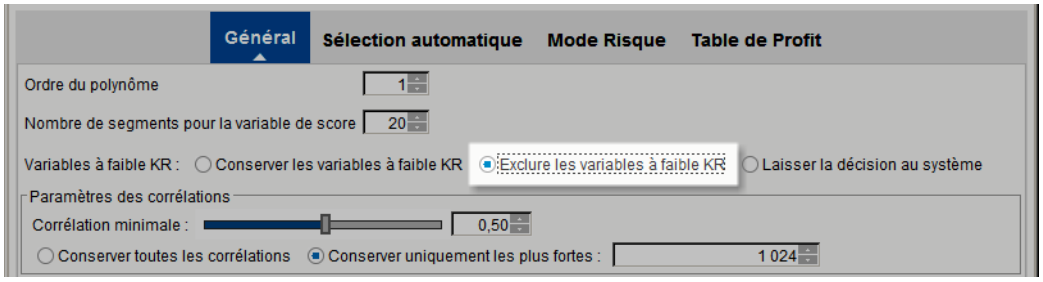
Exclusion des variables à faible KR

Cette option vous permet d'activer l'exclusion des variables d'après la valeur de leur KR (c'est-à-dire de leur reproductibilité). Pour déterminer si la reproductibilité d'une variable est trop faible, InfiniteInsight® calcule un seuil qui dépend principalement de la taille du jeu de données et de la distribution de la cible.

Dans les versions antérieures à la version 6.1.0, InfiniteInsight® excluait automatiquement les variables dont la reproductibilité était trop faible. Depuis la version 6.1.0, ce comportement a été désactivé par défaut. Si vous n'activez pas cette option, aucune variable ne sera exclue à cause de la valeur de sa reproductibilité.

☑ Pour exclure automatiquement les variables à faible KR

Cochez l'option *Exclure les variables à faible KR*.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

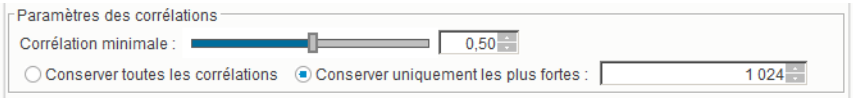
Nombre maximum de corrélations conservées

Cette option vous permet de choisir combien de corrélations devront être affichées dans le panneau de débriefing *Corrélations*.

Deux variables hautement corrélées contiennent les même informations par rapport à la variable cible. A chaque corrélation correspondent donc deux variables et un taux de corrélation. Lorsque vous modifiez le nombre de corrélations à afficher, le moteur supprime celles dont le taux de corrélation est le moins élevé, conservant ainsi uniquement les plus significatives.

☑ Pour modifier les corrélations à conserver

1 Dans la section *Paramètres des corrélations*, déplacez le curseur pour indiquer à partir de quel coefficient de corrélation celles-ci doivent être conservées.



2 Cochez l'option *Conserver uniquement les plus fortes*.

☑ Pour conserver toutes les corrélations

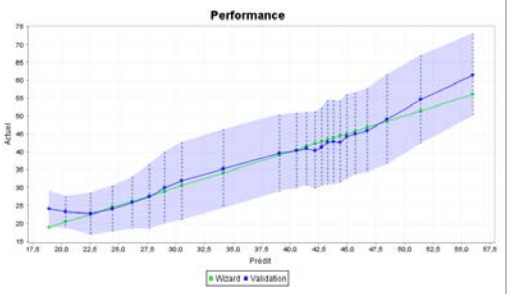
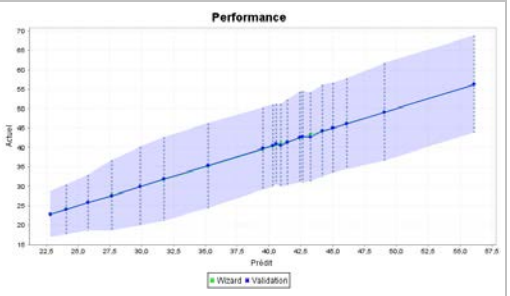
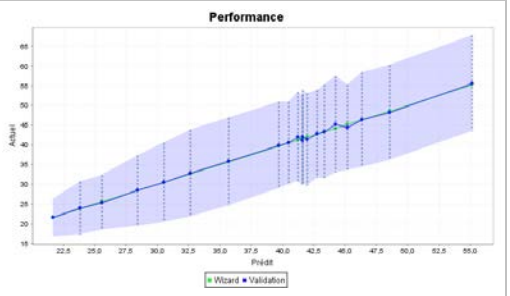
Cochez l'option *Conserver toutes les corrélations*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Activer le Post-traitement

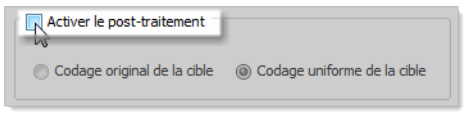
Cette section vous permet de paramétrer la régression selon trois stratégies. Cette option ne peut être activée que si le modèle contient au moins une variable cible continue.

La description de ces stratégies et un exemple de courbe de performances pour chaque stratégies sont proposés dans le tableau ci-dessous.

Stratégie de régression	Description	Exemple de courbe de performances
Pas de post-traitement	Cette stratégie consiste à désactiver la phase de redressement des prédictions lors de l'apprentissage du modèle afin de créer une régression similaire aux versions de SAP InfiniteInsight® antérieures à la 3.3.2. Dans ce cas, une régression standard est effectuée. Aucune optimisation spécifique n'est appliquée aux scores finaux. Les valeurs cible d'origine sont utilisées et les valeurs de score brutes sont générées en sortie.	
Codage original de la cible	La seconde consiste à activer la phase de redressement des prédictions et à utiliser la valeur de la cible directement lors de l'apprentissage du modèle pour calculer les coefficients de régression. Pendant la phase de post-traitement, le résultat de la régression est ensuite transformé afin d'aligner les moyennes des segments du score à celles de la variable cible. Note - C'est la stratégie utilisée par défaut dans SAP InfiniteInsight®.	
Codage uniforme de la cible	La dernière consiste à activer la phase de redressement des prédictions et à utiliser un codage normalisé de la cible lors de l'apprentissage du modèle afin d'obtenir une distribution uniforme : c'est la phase de prétraitement. Ensuite, les coefficients de régression sont calculés et les scores sont convertis dans l'espace d'origine de la cible. Note - Cette stratégie peut être choisie lorsque la stratégie par défaut ne produit pas des modèles de qualité satisfaisante, ce qui est souvent le cas avec des distributions dissymétriques des valeurs de cible.	

☑ Régression sans redressement

Décochez la case *Activer le post-traitement*.



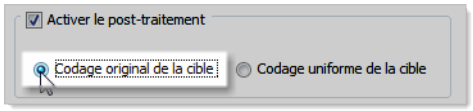
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Note

Il n'est pas possible de choisir le type de codage utilisé pour la cible quand la phase de redressement est désactivée.

☑ Régression utilisant la valeur cible

1. Cochez la case *Activer le post-traitement*.
2. Sélectionnez le bouton radio *Codage original de la cible*.



Note

Ce type de régression correspond aux régressions des versions 3.3.1 à 3.3.6 incluses. Cette stratégie de régression est la stratégie par défaut.

☑ Régression utilisant la valeur codée de la cible

1. Cochez la case *Activer le post-traitement*.
2. Sélectionnez le bouton radio *Codage uniforme de la cible*.

Définir la valeur des catégories cibles

SAP InfiniteInsight® vous donne la possibilité de définir les valeurs des catégories cibles des variables cibles lorsque celles-ci sont binaires. Par défaut, SAP InfiniteInsight® utilise comme catégorie cible la catégorie la moins représentée dans l'ensemble de données.

L'écran Paramètres Spécifiques du Modèle liste l'ensemble des variables cibles binaires, vous permettant ainsi de déterminer pour chacune la valeur de sa catégorie cible, c'est-à-dire la valeur attendue de la variable cible.

Pour ce scénario

Ne définissez aucune valeur pour la variable cible. SAP InfiniteInsight® sélectionnera automatiquement la valeur "1" comme catégorie cible pour la variable Class.

☑ Définir la catégorie cible d'une variable cible

Dans l'écran Paramètre spécifique du modèle, dans le champ Catégorie Cible correspondant à la variable cible choisie, entrez la valeur de la catégorie cible de cette variable.

Onglet "Sélection automatique"

L'onglet *Sélection automatique* vous permet de définir les paramètres de la sélection automatique des variables.

Sélection automatique des variables

Ces paramètres vous permettent de réduire automatiquement le nombre de variables du modèle par rapport à des critères de qualité. Cette sélection se fait par itérations successives. Il existe deux modes de sélection, un basé sur le nombre de variables à conserver, et l'autre sur la quantité d'information à conserver. La quantité d'information correspond à la somme des contributions des variables.

- *Nombre de variables conservées*
  L'interface vous permet de fixer le nombre de variables supprimées par itération et le nombre final de variables.
- *Quantité d'information conservée*
  L'interface vous permet de fixer la quantité d'information conservée par itération, ainsi que plusieurs critères d'arrêts tels que :
  - *Qualité* et *Perte autorisée*
    Pour une itération, la qualité de la sélection automatique de variables se base sur un indicateur définis soit par la somme du de la capacité prédictive (KI) et de la reproductibilité (KR), soit par la capacité prédictive uniquement ou la reproductibilité uniquement. On peut définir la perte de qualité autorisée pour cet indicateur.
  - *Variables min.*
    Ce critère d'arrêt permet de fixer le nombre minimal de variables du modèle final.

Il est aussi possible de copier dans l'arbre des paramètres les itérations successives du processus de sélection en sélectionnant l'option *Sauvegarder les étapes intermédiaires*. Ces informations sont accessibles après la génération du modèle dans *Protocols/Default/Transforms/Kxen.RobustRegression[...]/SelectionProcess/Iterations*.

☑ Pour utiliser la sélection automatique des variables

- Cochez la case *Activer la sélection automatique des variables*. Les options correspondantes sont activées.

  Les paramètres par défaut sont : "*Sélectionner le meilleur modèle en conservant entre 1 et toutes variables*."
- Chaque paramètre modifiable est signalé sous forme de lien hypertexte (bleu, souligné).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Mode de sélection

☑ Pour choisir le mode de sélection

1 Cliquez sur le lien caractérisant le type d'information à conserver à chaque itération du processus de sélection. Par exemple, *le meilleur modèle* dans la phrase "*Sélectionner le meilleur modèle en conservant entre 1 et toutes variables.*"

Une liste déroulante s'affiche, proposant les choix suivants:

- *le meilleur modèle*
- *le dernier modèle généré.*

2 Sélectionnez l'option de votre choix.

3 Cliquez sur *Validez*.

☑ Pour choisir le nombre de variables

Ce critère d'arrêt est obligatoire et permet de fixer le nombre minimal de variables du modèle final.

1 Dans la phrase "*Sélectionner le meilleur modèle en conservant entre 1 et toutes variables*", cliquez sur le nombre de variables minimum (par exemple, *1 variable*) et le nombre de variables maximum (par exemple, *toutes les variables*).
Pour sélectionner le nombre minimum de variables, un curseur allant de 1 au nombre total de variables du modèle s'affiche.

Pour sélectionner le nombre maximum de variables, vous pouvez soit confirmer ce minimum en cochant *Garder toutes les variables* ou choisir un nombre maximum de variables.

2 Cliquez sur *Valider.*

Critères d'arrêt

Vous avez le choix entre deux paramètres de sélection des variables :

- *Chaque étape retire 1 variable.*

Cette option vous permet de paramétrer le nombre de variables qui devraient être exclues à chaque itération.

- *Chaque étape conserve 95,0% de l'information.*

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Cette option vous permet de paramétrer la quantité d'information qui devrait être conservé à chaque itération, limitant ainsi la perte d'information.

Sélectionnez l'option de votre choix.

☑ Pour paramétrer le nombre de variables restantes

1. Cliquez sur le lien indiquant la nombre de variables dans la phrase "*Chaque étape retire 1 variable.*" Un curseur allant de 1 au nombre total de variables du modèle s'affiche.
2. Déplacez le curseur pour sélectionnez le nombre de votre choix.
3. Cliquez sur *Valider*.

☑ Pour paramétrer la quantité d'information

1. Cliquez sur le lien indiquant la quantité d'information à conserver dans la phrase "*Chaque étape conserve 95,0% de l'information*". Un curseur s'affiche.
2. Déplacez le curseur pour sélectionnez la quantité de votre choix.
3. Cliquez sur *Valider*.

☑ Pour paramétrer la perte de qualité autorisée

La perte de qualité est paramétrée dans la phrase "*La recherche s'arrête en cas de diminution de 5,0% du KI et du KR*".

1. Cliquez sur le lien indiquant le pourcentage de perte (par exemple, *5,0%*). Un curseur s'affiche.
2. Sélectionnez le pourcentage maximal autorisé de perte de qualité.
3. Cliquez sur *Valider*.
4. Cliquez sur le critère de qualité. Une liste déroulante s'affiche proposant les options suivantes :
   - *Basé sur KI + 2KR*, la perte de qualité est basée sur la capacité prédictive (KI) et deux fois la reproductibilité (KR)
   - *KI et KR*, la perte de qualité est limitée à la fois pour la capacité prédictive (KI) et pour la reproductibilité (KR). C'est la valeur par défaut.
   - *KI*, la perte de qualité est seulement limitée pour la capacité prédictive (KI).
   - *KR*, la perte de qualité est seulement limitée pour la reproductibilité (KR).
5. Sélectionnez l'option de votre choix.
6. Cliquez sur *Validez*.

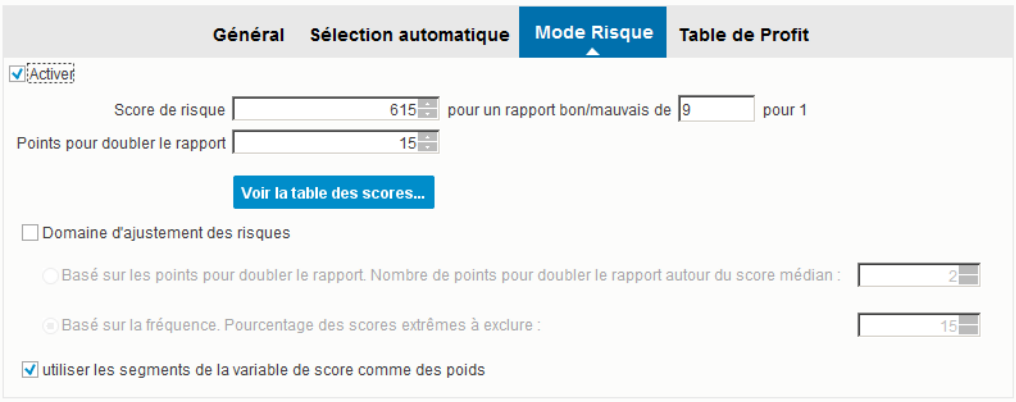
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Onglet "Mode Risque"

Cet onglet vous permet de sélectionnez un mode d'apprentissage spécifique pour votre modèle.

☑ Pour activer un mode d'apprentissage spécifique

1 Sélectionnez l'onglet *Mode Risque*.



2 Cochez la case *Activer*. L'onglet s'active et les paramètres du mode "Risque" s'affichent.

Activer le Mode "Risque"

Le mode "Risque" permet aux utilisateurs avancés de demander à un modèle de classement de traduire les équations internes qu'il a obtenues sans contrainte vers une échelle de scores spécifiée associées au rapport bons/mauvais.

Quand ce mode est activé, les différents codages internes des variables continues et ordinales sont rassemblés en une seule représentation qui permet une vision simplifiée des équations internes du modèle. Ceci est particulièrement intéressant lorsque l'utilisation de modèles prédictifs est soumise à des restrictions légales : les équations du modèle sont désormais assez simples pour être comprises par les services juridiques et peuvent être présentées, non seulement dans un langage de programmation comme avant, mais également en termes simples.

La technologie sous-jacente est également utilisée pour afficher les 'cartes de score'.

L'utilisation de ce mode nécessite que vous choisissiez :

- un *score de risque* associé à *un rapport bons/mauvais*

> **Note**
>
> Le rapport bons/mauvais est égal à (1-p)/p, où p est la probabilité du risque.

- le *nombre de points pour doubler le rapport*

> **Note**
>
> Les points pour doubler le rapport sont le nombre de points de risque nécessaires pour doubler le rapport bons/mauvais.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Exemple

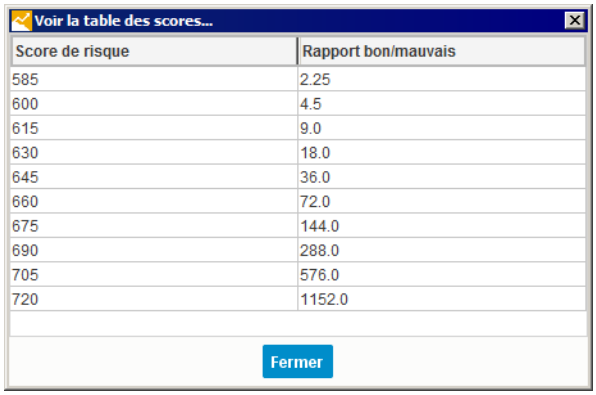
Si on considère un score de risque de 615, un rapport bons/mauvais de 9 pour 1 et 15 points pour doubler le score, InfiniteInsight® ré-échelonnera automatiquement les scores internes vers des scores dans l'espace du mode "Risque" et associera un rapport bons/mauvais à chacun de ces scores.

Dans ce scénario

- N'activez pas le mode "Risque".

Pour définir les paramètres du mode "Risque"

1. Dans le champ *Score de risque*, saisissez le score que vous voulez associer à rapport bon/mauvais.
2. Dans le champ *pour un rapport bon/mauvais de*, saisissez le rapport.
3. Dans le champ *Points pour doubler le rapport*, indiquez le nombre de points dont le score doit augmenter pour doubler le rapport.
4. Cliquez sur le bouton *Voir la table de score* pour afficher un tableau des scores associés aux rapports bon/mauvais correspondants.



Domaine d'ajustement des risques

Cette option permet à l'utilisateur de paramétrer la manière dont l'ajustement des scores de risque est effectué, c'est-à-dire comment InfiniteInsight® ajuste ses propres scores aux scores de risque.

L'option d'ajustement des scores a deux modes :

- *Basé sur les points pour doubler le rapport* : l'aire d'ajustement des scores est égale à `[Score médian - N*PDR ; Score médian + N*PDR]`. `N` (nombre de points pour doubler le rapport autour du score médian) doit être spécifié par l'utilisateur. Par défaut, il est égal à `2`.

Note

`PDR` signifie Points pour doubler le rapport.

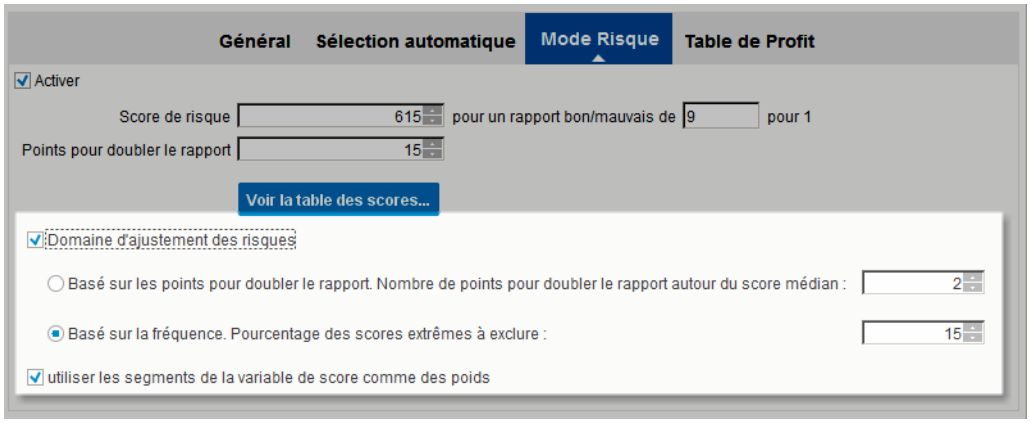
- *Basé sur la fréquence* : l'aire d'ajustement des scores est égale à `[Quantile(Freq) ; Quantile(1.0 - Freq)]`. La fréquence des scores extrêmes à exclure doit être spécifié par l'utilisateur. Par défaut, elle est égale à `15%`.

Si vous ne cochez pas la case *Domaine d'ajustement des risques*, le mode **Basé sur la fréquence** sera utilisé par défaut.

L'ajustement des scores peut être pondéré.

☑ Pour paramétrer l'ajustement des risques

1 Cochez la case *Domaine d'ajustement des risques*.



2 Sélectionnez le mode que vous souhaitez utiliser.

3 Selon le mode choisi, saisissez la valeur appropriée dans le champ correspondant.

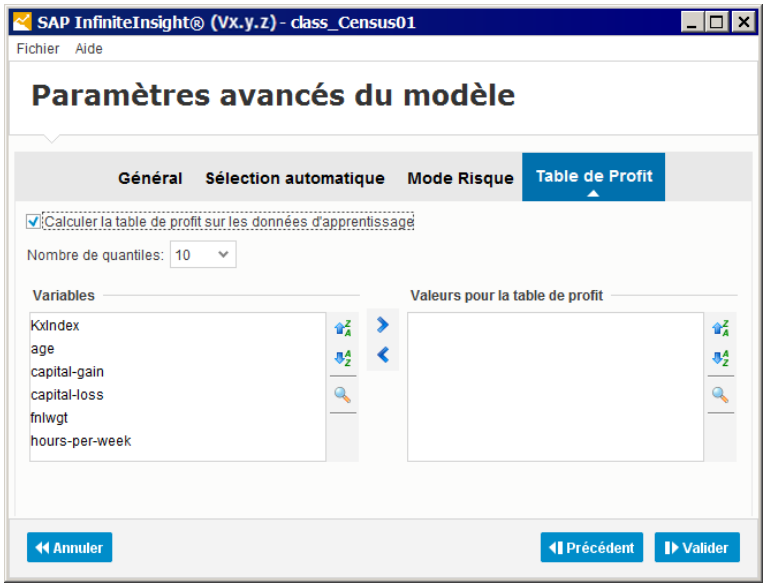
4 Si vous voulez pondérer l'ajustement des risques, cochez la case *utiliser les segments de la variable de score comme des poids*.

Onglet Table de profit

Ce panneau vous permet de calculer la table de profit pour le jeu de données d'application, c'est-à-dire de trier vos données par ordre de score décroissant et de les répartir de façon égale en quantiles (déciles, vingtiles ou centiles). Cette option peut être utile pour vérifier la performance du modèle sur le jeu de validation.

☑ Pour calculer la table de profit

1 Sélectionnez l'onglet *Table de Profit*.



2 Cochez la case *Calculer la table de profit sur les données d'apprentissage*.

3 Dans la liste, sélectionnez le *Nombre de quantiles* que vous souhaitez obtenir.

4 Vous pouvez ajouter des variables supplémentaires pour estimer le profit pour chaque segment de la population :

   1. Dans la liste *Variables*, sélectionnez les variables que vous souhaitez ajouter à la table de profit. Utilisez la touche *CTRL* de votre clavier pour sélectionner plusieurs variables à la fois.
   2. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Valeurs pour la table de profit*.

5 La somme de chaque variable sélectionnée sera calculée pour chaque segment de la population.

6 Cliquez sur le bouton *Valider* pour enregistrer les paramètres avancés et revenir au panneau *Appliquer un modèle*.

☑ Résultats

Vous pouvez également retrouver résultat du calcul de la table de profit dans la section *Performance du modèle* dans le panneau *Rapports de modélisation*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.2 Etape 2 - Générer et valider le modèle

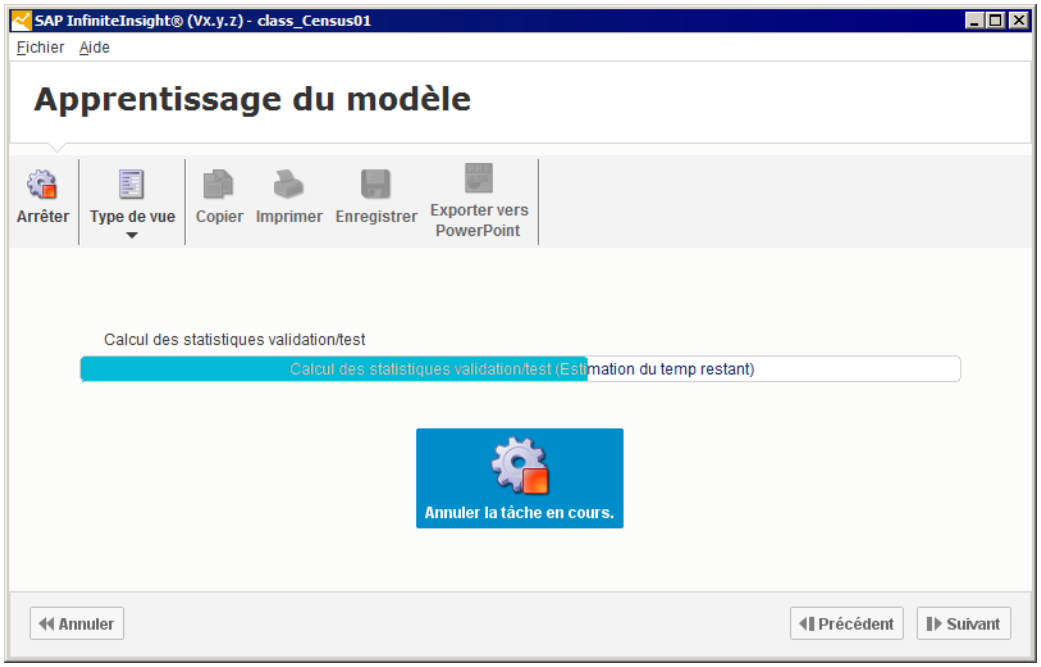
Une fois les paramètres de modélisation définis, vous pouvez générer le modèle. Vous devez ensuite valider ses performances grâce à la capacité prédictive (KI) et à la reproductibilité (KR) :

- Si le modèle est suffisamment performant, vous pouvez analyser les réponses qu'il apporte par rapport à votre problématique (étape 3 à la page 108, à la page 225), puis l'appliquer sur de nouveaux jeux de données (étape 4).
- Sinon, vous pouvez modifier les paramètres de modélisation de manière à ce qu'ils soient plus adaptés à votre jeu de données et à votre problématique, et générer ainsi de nouveaux modèles plus performants.

5.2.1 Générer le modèle

☑ Pour générer le modèle

1 Dans l'écran *Récapitulatif des paramètres du modèle*, cliquez sur le bouton *Générer*.
L'écran *Apprentissage du modèle* apparaît. La génération du modèle est en cours. Une barre de progression vous permet de suivre le déroulement des différentes étapes.



2 Une fois le modèle généré, passez à la section Valider le modèle généré (voir à la page 70).

5.2.2 Suivi du processus de génération

Il existe deux manières de suivre la progression du processus de génération du modèle :

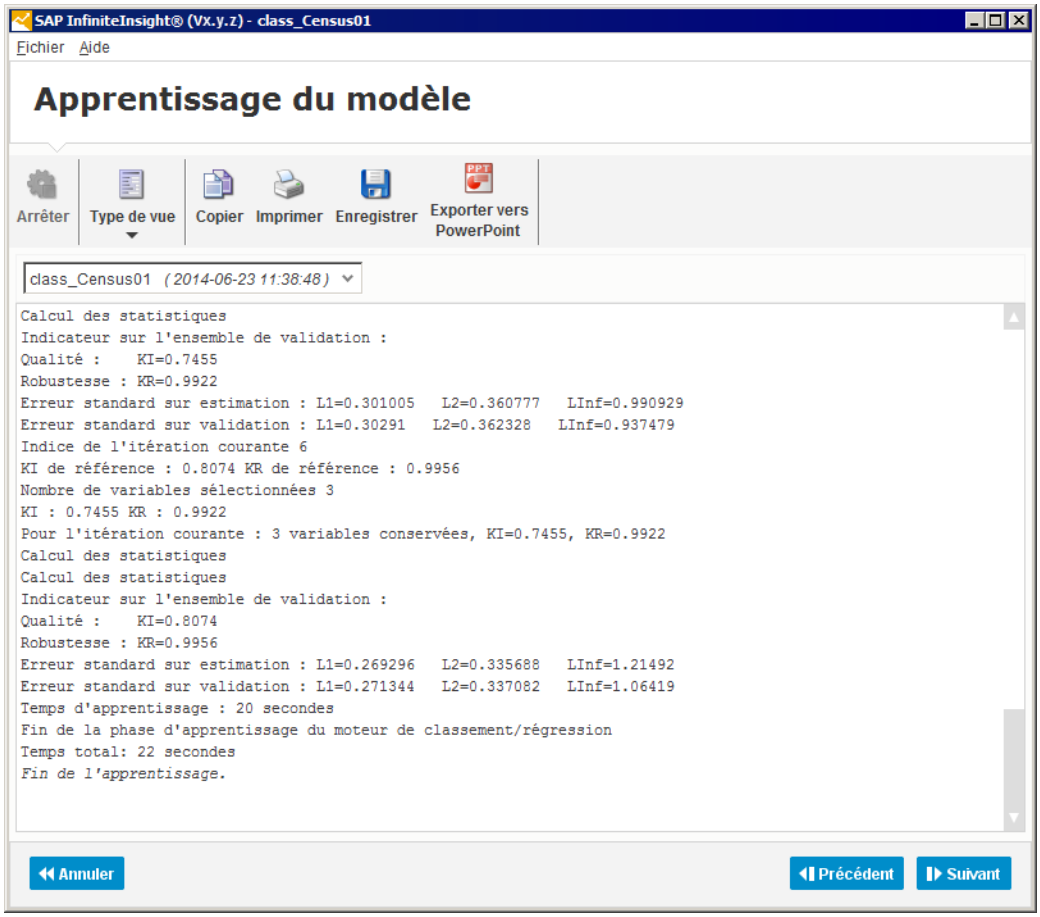
- La **Barre de progression** affiche la progression de chaque étape. C'est l'écran par défaut..
- Le **Détail du processus** affiche des messages détaillés pour chaque étape.

☑ Pour afficher la barre de progression

Cliquez sur le bouton  (*Affiche la progression*).
La barre de progression s'affiche.

☑ Pour afficher le détail du processus

Cliquez sur *Type de Vue* >  *Détails des messages*.
L'écran ci-dessous s'affiche.



☑ Pour arrêter le processus d'apprentissage

1 Cliquez sur le bouton  (*Arrêter*).
Une boîte de dialogue de confirmation s'affiche.

2 Cliquez sur le bouton *Précédent*.
L'écran *Récapitulatif des paramètres de modélisation* s'affiche.

3 Reportez-vous à la section Vérifier les paramètres de modélisation.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.2.3 Valider le modèle généré

Une fois le modèle généré, vous devez vérifier sa validité en observant les indicateurs de performance :

- la **capacité prédictive** vous permet de connaître le **pouvoir explicatif** du modèle, c'est-à-dire sa capacité à expliquer les valeurs de la variable cible sur le jeu de données d'apprentissage. Un modèle parfait possède une capacité prédictive égale à 1 et un modèle purement aléatoire possède une capacité prédictive égale à 0.
- la **reproductibilité** vous permet de connaître le **degré de robustesse** du modèle, c'est-à-dire sa capacité à conserver le même pouvoir explicatif sur un nouveau jeu de données. En d'autres mots, le degré de robustesse correspond à la capacité prédictive du modèle sur un jeu de données d'application.

Pour savoir comment sont calculées la capacité prédictive et la reproductibilité, voir Capacité prédictive, reproductibilité et courbes de profit à la page 232.

Remarque

La validation du modèle est une phase primordiale dans le processus global de Data Mining. Accordez toujours une importance majeure aux valeurs obtenues pour la capacité prédictive et la reproductibilité d'un modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Pour ce scénario

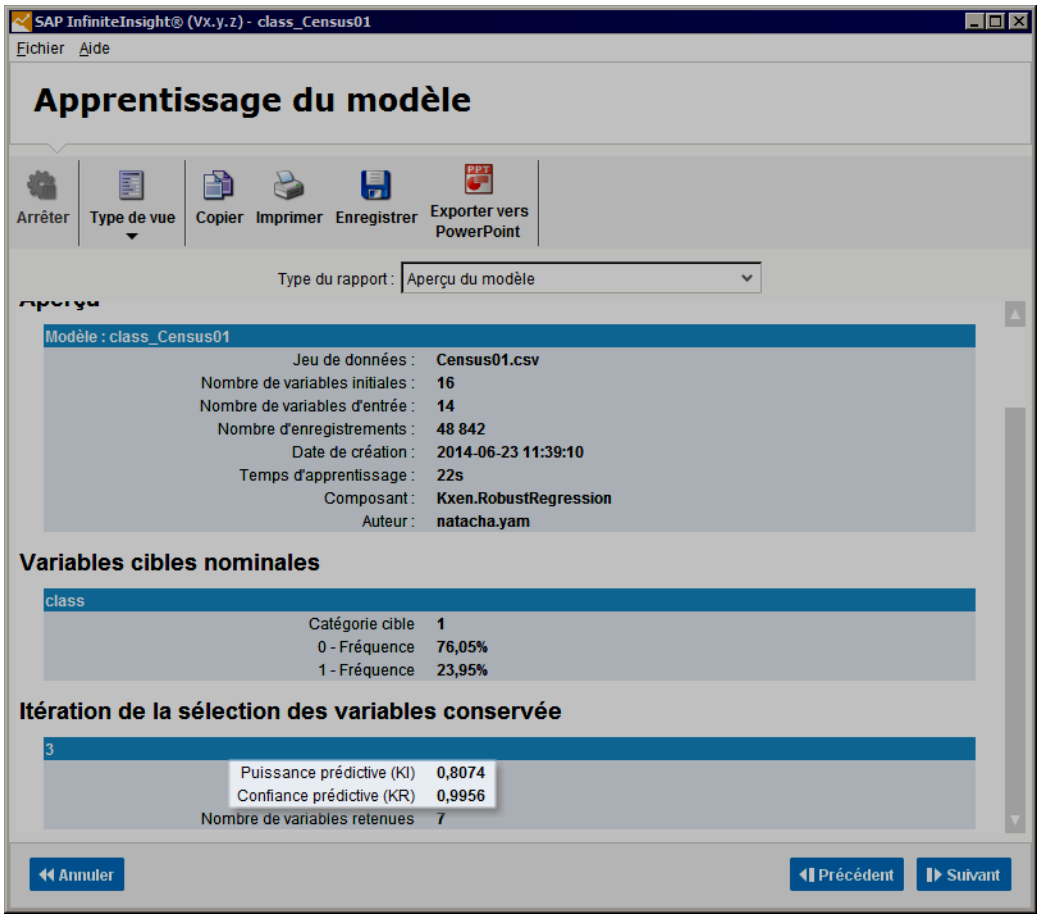
Le modèle généré possède :

- un indicateur de qualité KI égal à **0,8074**,
- un indicateur de robustesse KR égal à **0,9956**.

Le modèle est suffisamment performant. Vous n'avez pas besoin d'en générer un autre.

☑ Pour valider le modèle généré

Vérifiez les indicateurs de qualité *KI* et de robustesse *KR* du modèle. Ces indicateurs sont encerclés sur la figure suivante.
a) Si les performances du modèle vous conviennent, passez à l'étape 3 "Analyser et comprendre le modèle généré à la page 108, à la page 225"
b) Sinon, passez à la procédure Pour générer un nouveau modèle (voir à la page 70).



☑ Pour générer un nouveau modèle

Vous avez deux options. Dans l'écran *Apprentissage du modèle*, vous pouvez :

- soit cliquer sur le bouton *Précédent* pour revenir sur les paramètres de modélisation initialement définis. Vous pouvez alors modifier les paramètres un à un.
- soit cliquer sur le bouton *Annuler* pour revenir à la page d'accueil de l'assistant de modélisation. Vous devez alors redéfinir tous les paramètres de modélisation.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

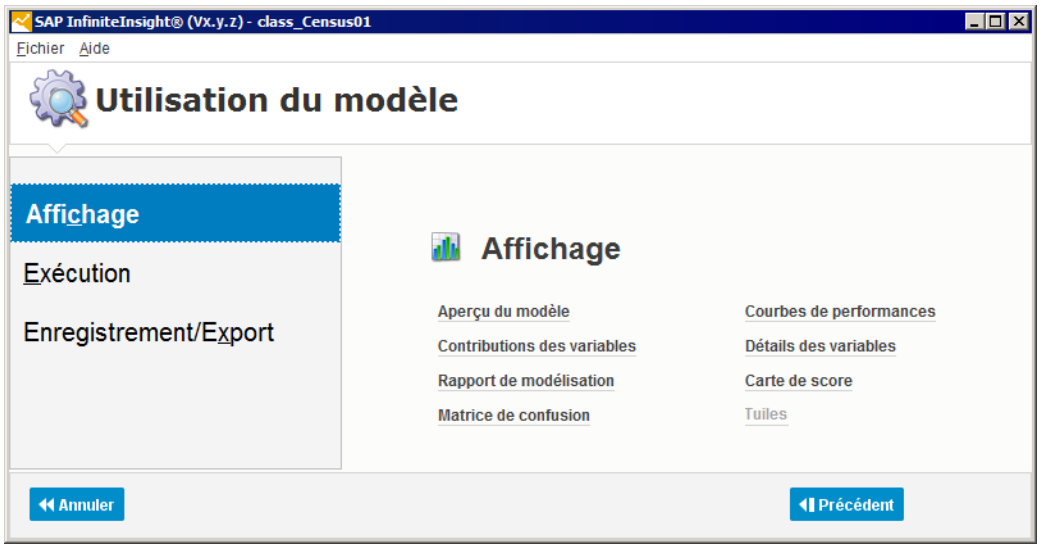
5.3 Etape 3 - Analyser et comprendre le modèle généré

Un ensemble d'outils graphiques vous permet d'analyser le modèle généré et de connaître :

- la **performance du modèle** par rapport à un hypothétique modèle parfait et un modèle de type aléatoire,
- la **contribution de chacune des variables explicatives** par rapport à la variable cible,
- l'**importance des différentes catégories de chaque variable** par rapport à la variable cible.

5.3.1 Menu d'utilisation

Une fois le modèle généré, cliquez sur le bouton *Suivant*. L'écran *Utilisation du modèle* apparaît.



L'écran *Utilisation du modèle* présente les différentes options d'utilisation du modèle, qui vous permettent :

- d'afficher les informations relatives au modèle généré (groupe *Affichage*), c'est-à-dire l'aperçu du modèle, les graphiques des courbes d'évaluation, des contributions des variables et des différentes variables, des rapports statistiques détaillés au format HTML, des tables d'analyse. Certaines informations ne sont affichées qu'à la demande de l'utilisateur : ainsi l'affichage des résultats de InfiniteInsight® Modeler / Régression ou Classement sous forme d'arbre de décision doit être spécifié lors du paramétrage du modèle et l'accès aux paramètres du modèle doit être spécifié dans les options générale de l'assistant.
- d'appliquer et de simuler le modèle généré sur de nouvelles données, et d'affiner le modèle en effectuant une sélection automatique des variables explicatives à prendre en compte (groupe *Exécution*).
- d'enregistrer le modèle, ou de générer les codes source correspondants (groupe *Enregistrement/Export*).

5.3.2 Aperçu du modèle

L'*aperçu du modèle* reprend les informations récapitulée à la fin du processus de génération.

Ces informations sont détaillées dans les sections ci-dessous.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Aperçu

Modèle	Nom du modèle créé à partir du nom de la variable cible et du nom du jeu de données
Jeu de données	Nom du fichier de données
Nombre de variables initiales	Nombre de variables dans le jeu de données
Nombre de variables d'entrée	Nombre de variables explicatives conservées
Nombre d'enregistrements	Nombre d'enregistrements de la source de données
Date de création	Date et heure de la création du modèle
Temps d'apprentissage	temps d'apprentissage total (par défaut le temps est indiquée en seconde)
Composant	Selon le composant utilisé pour créer le modèle : ▪ Kxen.RobustRegression ▪ Kxen.SmartSegmenter ▪ Kxen.TimeSeries ▪ Kxen.AssociationRules ▪ Kxen.EventLog ▪ Kxen.SequenceCoder ▪ Kxen.SocialNetwork

Notifications

Variables Monotones Détectées	Indique si des variables monotones ont été trouvées dans le jeu de données, c'est-à-dire des variables dont le sens de variation est constant, dans l'ordre de lecture des données dans le jeu d'estimation.
Variables Suspectes Détectées	Ce rapport présente une liste de variables qui sont considérées comme suspectes. Ces variables suspectes ont un KI > 0.9, elles sont très fortement corrélées à la variable cible. Cela signifie que ces variables apportent probablement une information biaisée et qu'elles ne devraient pas être utilisées pour la modélisation. Une attention particulière doit être accordée à ces variables. Un rapport plus détaillé liste quelles variables particulères sont suspectes et dans quelle mesure (voir Rapports Statistiques > Compte Rendu Expert > Variables Suspectes).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variables cibles nominales

Pour chaque variable cible nominale

<Nom de la variable cible>	Le nom de la variable cible nominale concernée
Catégorie cible	Valeur de la catégorie cible
<Catégorie non-cible> - Fréquence	Proportion d'enregistrements pour lesquels la valeur de la variable cilbe **n'est pas égale** à la catégorie cible
<Catégorie cible> - Fréquence	Proportion d'enregistrements pour lesquels la valeur de la variable cible **est égale** à la catégorie cible

Variables cibles continues

Pour chaque variable cible continue

<Nom de la variable cible>	Le nom de la variable cible continue concernée
Min	La valeur minimum trouvée pour cette variable cible
Max	La valeur maximum trouvée pour cette variable cible
Moyenne	La moyenne des valeurs de cette variable cible
Ecart Type	L'écart type des valeurs de cette variable cible

Indicateurs de performance

Pour chaque variable cible:

rr_<variable cible>	nom du modèle, identifié par le préfixe `rr_` suivi du nom de la variable cible. Par exemple, `rr_class`.
KI	Indicateur de qualité. Pour plus d'information sur le KI, reportez-vous à la section **Indicateurs de performances** (voir à la page 40).
KR	Indicateur de robustesse. Pour plus d'information sur le KR, reportez-vous à la section **Indicateurs de performances** (voir à la page 40)

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Options

☑ Pour copier l'aperçu du modèle

1. Cliquez sur le bouton  (*Copier*).
   L'application copie le code HTML correspondant à l'aperçu du modèle.
2. Collez les paramètres dans l'application de votre choix.

☑ Imprimer l'aperçu du modèle

1. Cliquez sur le bouton  (*Imprimer*).
   Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
2. Sélectionnez l'imprimante et les options d'impression.
3. Cliquez sur *OK*.
   L'impression est lancée.

☑ Pour enregistrer l'aperçu du modèle

1. Cliquez sur le bouton  (*Enregistrer*).
   Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
2. Entrez un nom de fichier.
3. Choisissez le dossier de destination.
4. Cliquez sur *OK*.
   Les informations du modèle sont sauvegardées dans un fichier texte.

Exporter vers PowerPoint

☑ Pour exporter vers PowerPoint

Cliquez sur  (*Exporter vers PowerPoint*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.3.3 Les courbes de performances

Définition

Selon le type de cible, le graphique des **courbes de performances** vous permet de :

- visualiser le **profit réalisable** par rapport à votre problématique en utilisant le modèle généré lorsque la cible est nominale.
- **comparer les performances du modèle généré** à celles d'un modèle de type aléatoire et celles d'un modèle hypothétique parfait lorsque la cible est nominale.
- comparer la **valeur prévue** à la **valeur réelle** lorsque la cible est continue.

Sur le graphique, les courbes représentent :

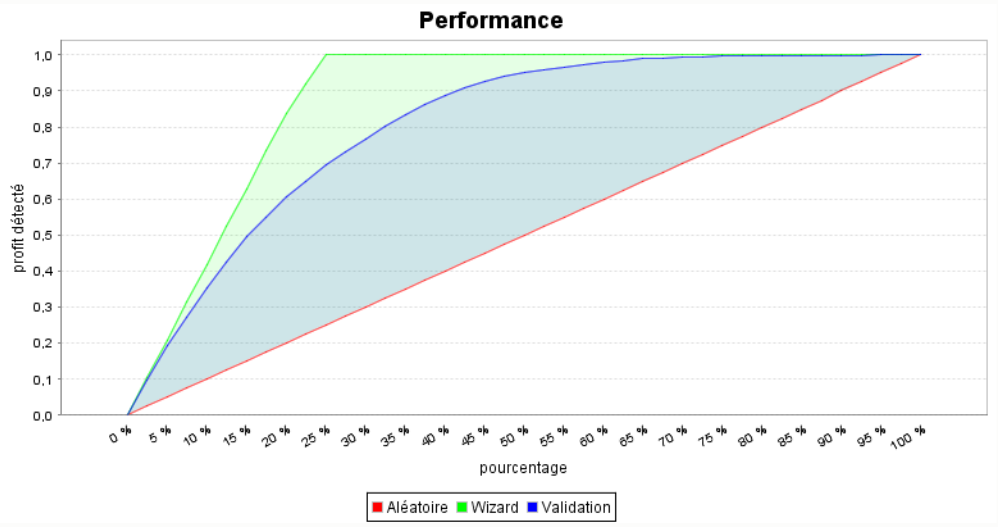
- le profit réalisable (axe des ordonnées) en fonction du taux d'observations sélectionnées sur la totalité du jeu de données initial (axe des abscisses) pour une cible nominale,
- la valeur prédite par rapport à la valeur réelle pour une cible continue.

Afficher le graphique des courbes de profit

☑ Pour afficher le graphique des courbes de performances

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Courbes de performances*.
Les courbes de performances s'affichent.

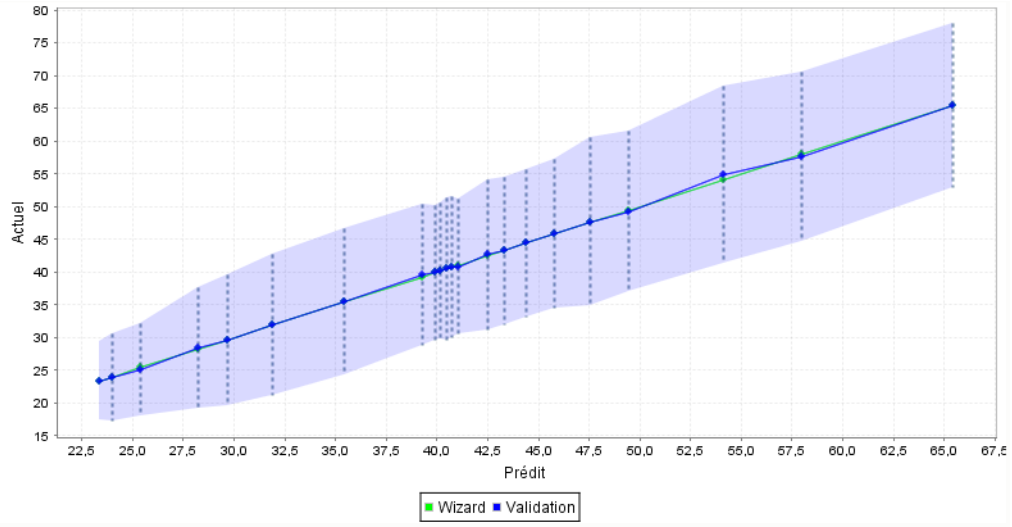
Lorsque la variable cible est de type **nominal**, une courbe du type suivant s'affiche.



Les paramètres par défaut affichent les courbes de profit correspondant au sous-jeu de *Validation*, à un hypothétique modèle parfait (*Wizard*) et à un modèle aléatoire (*Random*). Le type de profit utilisé est profit *Détecté*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Lorsque que la variable cible est de type **continu**, une courbe du type suivant s'affiche :



Les paramètres par défaut affichent les courbes correspondant au sous-jeu de *Validation* et à un hypothétique modèle parfait (*Wizard*). Le type de courbe utilisé est *Prédit/Réel*. La zone bleue correspond à la déviation standard du modèle en cours.

2 Dans le cas où il existe plusieurs variables cibles, sélectionnez dans la liste déroulante *Modèles* celui pour lequel vous souhaitez voir les courbes de performances.

> **Note**
>
> A chaque variable cible correspond un modèle. Le nom de chaque modèle est constitué du préfixe *rr_* (*Robust Regression*) et du nom de la variable cible concernée.

3 Sélectionnez les options de visualisation qui vous intéressent.
Pour plus d'informations sur les options de visualisation, Options de visualisation (à la page 113).

Options de visualisation

☑ Pour copier une courbe de performances

1 Cliquez sur le bouton  (*Copier*).
2 Sélectionnez l'option désirée.
L'application copie les paramètres de la courbe.
3 Collez les paramètres dans l'application de votre choix. Vous pouvez par exemple les utiliser pour générer un graphique dans un tableur (Excel, ...).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Pour imprimer une courbe de performances

1 Cliquez sur le bouton  (*Imprimer*).
Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
2 Sélectionnez l'imprimante et les options d'impression.
3 Cliquez sur *OK*.
L'impression est lancée.

☑ Pour enregistrer une courbe de performances

1 Cliquez sur le bouton  (*Enregistrer*).
Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
2 Entrez un nom de fichier.
3 Choisissez le dossier de destination.
4 Cliquez sur *OK*.
La courbe est enregistrée au format PNG dans le dossier sélectionné.

☑ Pour afficher les courbes des sous-jeux d'estimation, de validation et de test

1 Dans l'écran *Courbes de performances*, cliquez sur *Jeux de données* et sélectionnez l'une des options suivantes :

-  *Tous les jeux de données*.
-  *Validation uniquement*.

Pour exporter au format Excel

☑ Pour exporter au format Excel

Cliquez sur  (*Exporter au format Excel*).

Pour ouvrir la vue courante dans une nouvelle fenêtre

☑ Pour ouvrir la vue courante dans une nouvelle fenêtre

Cliquez sur  (*Punaiser la vue*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Pour un modèle à cible nominale

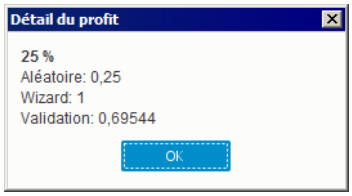
Sur le graphique des courbes de performances, différentes options vous permettent de visualiser :

- les **valeurs exactes** d'un point pour toutes les courbes représentées.
- les courbes de profit associées aux **sous-jeux d'estimation et de test**.
- les différentes courbes profit en fonction des **types de profit:**
  - *Détecté*,
  - *Lift*,
  - *Normalisé*,
  - *ROC*
  - *Lorenz 'Bon'* et *'Mauvais'*
  - *Densité 'Bon'*, *'Mauvais'* et *'Tous'*
  - *Personnalisé*.

Pour plus d'informations sur les courbes de profit (voir "Types de profit" à la page 46).

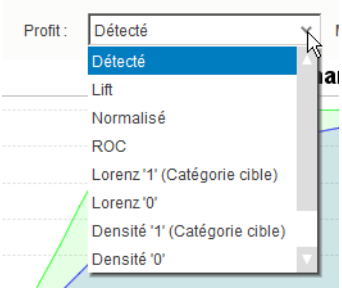
☑ Pour afficher les valeurs de profit exactes pour un point donné

Dans l' écran *Courbes de performances*, sur le graphique, cliquez sur un point de l'une des courbes représentées.
Par exemple, en cliquant sur un point de l'une des courbes ayant pour valeur en abscisse 25%, les valeurs de profit exactes apparaissent.



☑ Pour sélectionner un type de profit

1 Dans l'écran *Courbes de performances*, au-dessus du graphique, cliquez sur la liste déroulante associée au champ *Profit*.
La liste des types de profit apparaît.



2 Sélectionnez un type de profit.
Les courbes correspondantes s'affichent.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

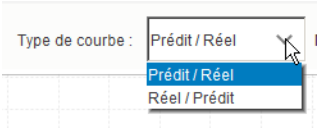
Pour un modèle à cible continue

Sur le graphique des courbes de performances, différentes options vous permettent de visualiser :

- les **valeurs exactes** d'un point pour toutes les courbes représentées.
- les courbes associées aux **sous-jeux d'estimation et de test**.
- la courbe en fonction des **types** *Prédit/Réel* **ou** *Réel/Prédit*.

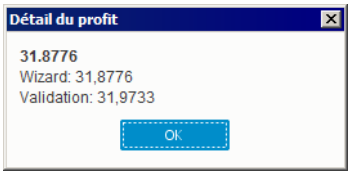
☑ Pour afficher les valeurs de profit exactes pour un point donné

Dans l'écran *Courbes de performances*, sur le graphique, cliquez sur un point de l'une des courbes représentées.
Par exemple, en cliquant sur un point de l'une des courbes ayant pour valeur en abscisse 29 ans, les valeurs exactes prédites et réelles s'affichent.



☑ Pour sélectionner un type de courbe

1 Dans l'écran *Courbes de performances*, sous le titre, cliquez sur la liste déroulante associée au champ *Type de courbe*.
La liste des types de courbe apparaît.



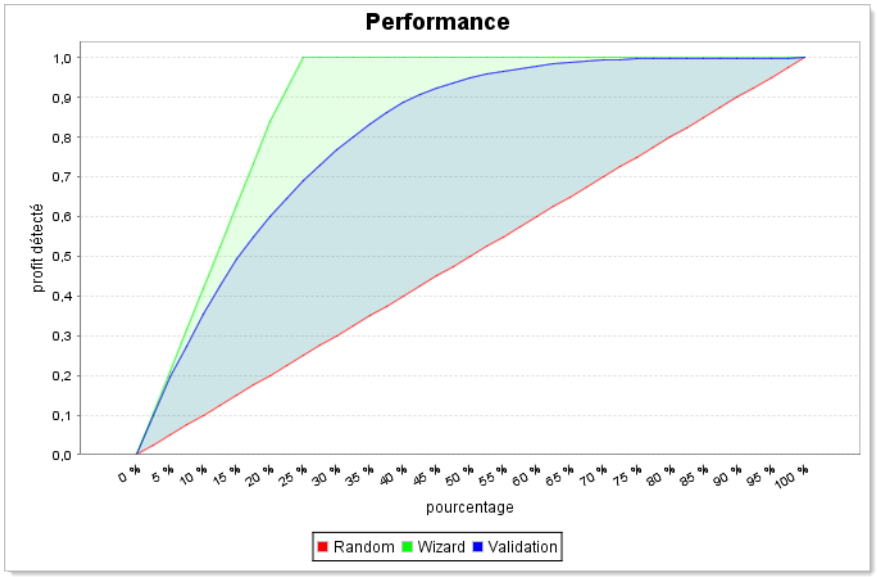
2 Sélectionnez un type de courbe.
Les courbes correspondantes s'affichent.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre les courbes de performances

Pour un modèle à cible nominale

La figure ci-dessous représente le graphique des courbes de profit utilisant les paramètres par défaut.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-avec InfiniteInsight® Modeler

Sur le graphique, les courbes représentent pour chaque type de modèle le profit réalisable (axe des ordonnées), c'est-à-dire le pourcentage d'observations appartenant à la variable cible, en fonction du taux d'observations sélectionnées sur la totalité du jeu de données initial (axe des abscisses). Sur l'axe des abscisses, les observations sont ordonnées de manière décroissante en fonction de leur "score", c'est-à-dire par probabilité décroissante d'appartenir à la catégorie cible de la variable cible.

Dans ce scénario d'utilisation, les courbes de profit représentent le taux de prospects susceptibles de répondre de manière positive à votre campagne marketing sur la totalité des prospects référencés dans votre base de données.

Le profit *Détecté* est le type de profit proposé par défaut. Avec ce type de profit :

- la valeur "0" est affectée aux observations n'appartenant pas à la catégorie cible de la variable cible,
- la valeur "1/(fréquence de la variable cible dans le jeu de données)" est affectée aux observations appartenant à la catégorie cible de la variable cible.

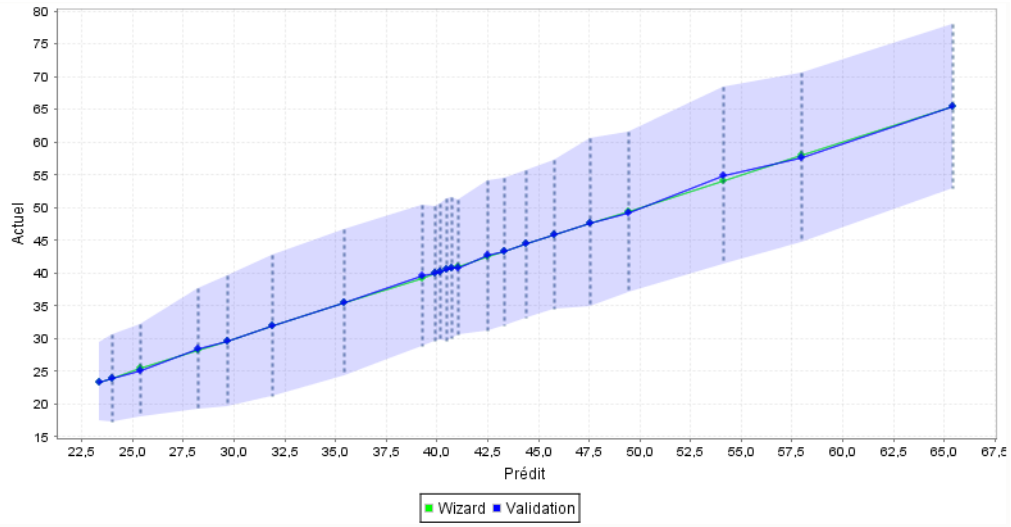
Le tableau suivant décrit les trois courbes représentées sur le graphique utilisant les paramètres par défaut.

La courbe...	Représente...	Par exemple, en sélectionnant...
Wizard (courbe verte, la plus haute)	le profit réalisable en utilisant un hypothétique *modèle parfait*, permettant de *connaître de manière absolue* la valeur de la variable cible pour chaque observation du jeu de données	25% des observations sur la totalité de votre jeu de données à l'aide d'un modèle parfait, 100% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées. Le profit maximum est alors atteint. **Remarque** Ces 25% correspondent au pourcentage de prospects ayant répondu de manière positive à votre campagne marketing, lors de votre phase de test. Pour ces prospects, la valeur de la variable cible, ou profit, est égale à 1.
Validation (courbe bleue, du milieu)	le profit réalisable en utilisant le *modèle généré par InfiniteInsight® Modeler*, permettant de prédire au mieux la valeur de la variable cible pour chaque observation du jeu de données	25% des observations de votre jeu de données initial à l'aide du modèle généré, 69% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées
Aléatoire (courbe rouge, la plus basse)	le profit réalisable en utilisant un *modèle aléatoire*, ne permettant de connaître en aucun cas la valeur de la variable cible pour chaque observation du jeu de données.	25% du jeu de données initial à l'aide d'un modèle aléatoire, 25% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Pour un modèle à cible continue

La figure ci-dessous représente le graphique des courbes de performances lorsque la cible est continue.



Les graphiques par défaut affichent les valeurs de la cible réelle (axes des ordonnées) en fonction des valeurs de la cible prédite (axes des abscisses). Deux courbes sont tracées : une pour le jeu de données *Validation* (représentée par une courbe bleue) et une autre pour le modèle parfait (représentée par une courbe verte). Par exemple, lorsque le modèle prédit 35, la moyenne de la valeur réelle est 37. La courbe du *Wizard* correspond simplement à X=Y, ce qui signifie que chaque valeur prédite est égale à la valeur réelle. Ce graphique permet de voir facilement et rapidement les erreurs du modèle. Lorsque la courbe s'éloigne trop du modèle parfait, cela signifie que la valeur prédite est suspecte.

Le graphique est calculé comme suit :

- les valeurs prédites sont réparties dans environ 20 segments ou groupes. Chacun de ses segments représente environ 5 % de la population.
- pour chacun de ces segments des statistiques basiques sont calculées sur la valeur réelle, telles que la moyenne du segment (*SegmentMean*), la moyenne associée à la cible (*TargetMean*) et la variance de la cible sur ce segment (*TargetVariance*). Par exemple pour une valeur prédite dans [17; 19], si la moyenne est égale à 18,5, la moyenne réelle est égale à 20,5 et la variance de la valeur réelle est égale à 9. Dans ce cas on peut dire que, si la valeur prédite se situe entre 17 et 19, le modèle sous-estime légèrement la valeur réelle.

Pour chaque courbe, un point est défini comme la moyenne d'un segment (*SegmentMean*) en abscisse et la moyenne associée à la cible en ordonnée (*TargetMean*).

La zone bleue représente la déviation standard attendue du modèle courant. Cette zone représente environ 70% des valeurs de la cible attendues.

Il est à noter que cet intervalle de prédiction (c'est dire la moitié de la zone bleue) est égal à la déviation standard de la cible observée pour un segment de valeurs prédites. En d'autres mots, cela signifie que, dans la cas d'une distribution Gaussienne, 70 % des valeurs réelles se situent dans cette zone.

> **Note**
>
> Il s'agit évidemment d'un pourcentage théorique qui peut varier.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Les valeurs extrêmes de l'intervalle de prédiction se calculent de la façon suivante :
`{TargetMean - (sqrt(TargetVariance)); TargetMean + (sqrt(TargetVariance))}`

> **Note**
>
> La déviation standard est égale à `sqrt(TargetVariance)`.

KI, KR et courbes de performances

Sur le graphique des courbes de performances pour un modèle dont la cible est continue :

- pour le jeu de données d'estimation (graphique par défaut), l'indicateur **KI** correspond au rapport entre "la surface se trouvant entre la courbe du **modèle généré** et celle du **modèle aléatoire**" et "la surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire**". Ainsi plus la courbe du modèle généré se rapproche de la courbe du modèle parfait, plus le KI se rapproche de 1.
- pour les jeux de données d'estimation, de validation et de test (sélectionnez l'option correspondante dans la liste *Jeu de données*, située sous le titre), l'indicateur **KR** correspond au rapport entre la "surface se trouvant entre la courbe du **jeu d'estimation** et celle du **jeu de validation**" et la "surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire**".

5.3.4 Contribution des variables

Définition

Le graphique des **contributions des variables** vous permet de visualiser l'**importance relative de chacune des variables dans le modèle**. Sur ce graphique, chaque barre représente la contribution d'une variable explicative par rapport à la variable cible.

Les quatre types de graphiques suivants permettent de visualiser les contributions des variables :

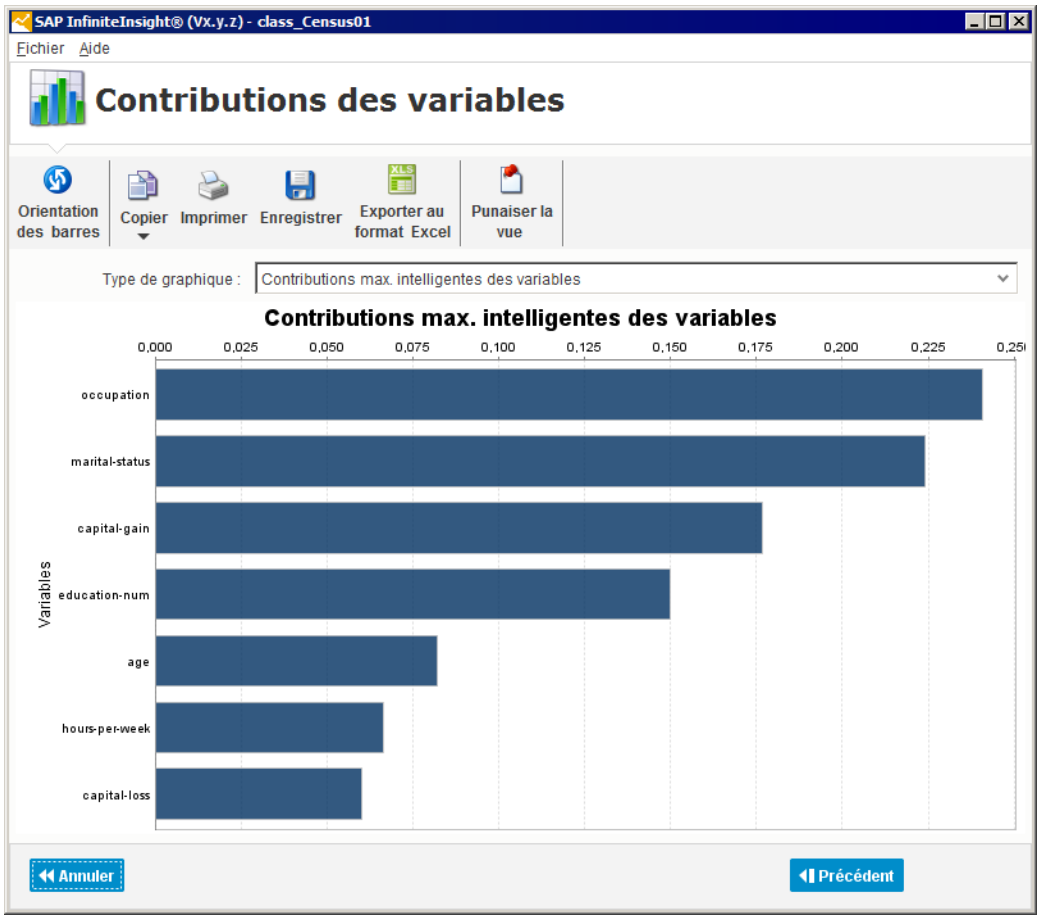
- *Contribution des variables*
- *Poids des variables*
- *Contributions intelligentes des variables*
- *Contributions maximales intelligentes des variables*

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Afficher les contributions des variables

☑ Pour afficher le graphique des contributions des variables

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Contributions des variables*.
Le graphique de *Contributions des variables* apparaît.
Le type de graphique défini par défaut est *Contributions maximales intelligentes des variables*.

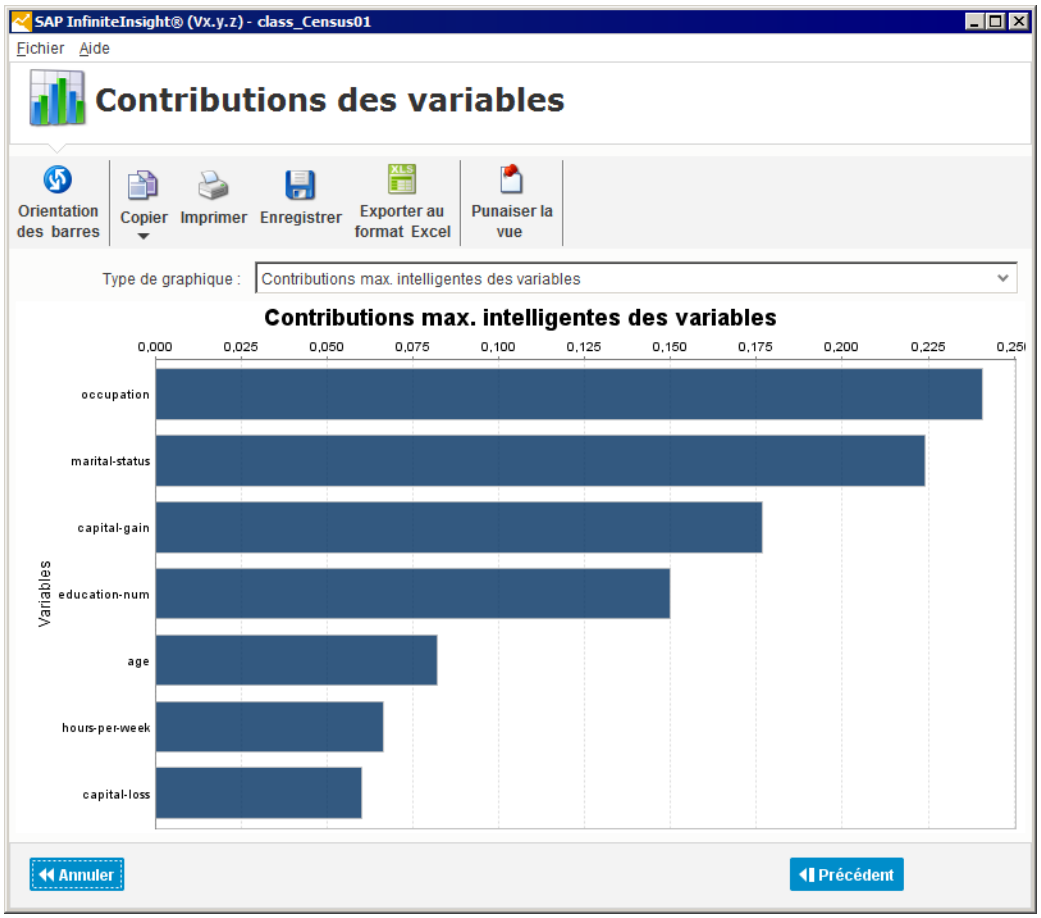


Si votre jeu de données contient des variables de type Date ou Datetime, des variables générées automatiquement peuvent apparaître dans ce panneau. Pour plus d'information, reportez-vous à la section Variables de date : les variables générées automatiquement (voir "Variables de Date : les variables générées automatiquement" à la page 31).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre les contributions des variables

L'option *Afficher les contributions des variables* permet de visualiser l'importance de la contribution de chacune des variables explicatives par rapport à la variable cible. Cette importance est relative : l'importance d'une variable donnée est calculée en fonction de l'importance des autres variables explicatives.



Sur le graphique ci-dessus, correspondant au modèle généré, les deux variables qui contribuent le plus à l'explication de la variable cible sont :

- *marital-status*,
- *capital-gain*.

En d'autres mots, les variables *marital-status* (statut marital) et *capital-gain* (gains en bourse) sont celles qui déterminent le plus si un prospect répond de manière positive ou négative à votre campagne marketing. Parmi toutes les variables contenues dans le jeu de données, ce sont les **variables les plus discriminantes** par rapport à la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Variables corrélées

Dire que des variables sont **corrélées** signifie qu'elles sont en partie redondantes, qu'elles apportent en partie la même information par rapport à la variable cible. Deux variables fortement corrélées décrivent donc en grande partie une même information, un même concept.

Le graphique *Contributions intelligentes des variables* rend compte des corrélations qui peuvent exister entre les différentes variables explicatives. Quand deux variables A et B sont fortement corrélées :

- la variable A, qui a une contribution plus forte que B par rapport à la variable cible, devient la "variable primaire" : le graphique représente **tout son apport**, y compris l'information qu'elle a en commun avec la variable B.
- la variable B, qui a une contribution plus faible que A par rapport à la variable cible, devient la "variable secondaire" : seul son **apport marginal** est représenté sur le graphique, c'est-à-dire les informations qu'elle ne partage pas avec la variable A. Cette différence d'information est notée `[variable_B]-[variable_A]`.

Variables codées

Pour créer un modèle, SAP InfiniteInsight® utilise non seulement les variables originales, mais également, dans le cas de variables continues ou ordinales, leur valeur codées par InfiniteInsight® Modeler / Codeur analytique. C'est ce qu'on appelle le codage double. Cela permet à SAP InfiniteInsight® d'extraire toute l'information contenue dans chaque variable.

Les variables codées sont indiquées par le préfixe `c_` dans les graphiques de contributions. Ainsi, la version codée de la variable `age` est notée `c_age`.

> **Note**
>
> Dans InfiniteInsight® Modeler, dans le panneau Description des données, si vous activez le codage naturel pour une variable donnée, la valeur codée de cette variable (*c_NomVariable*) ne sera pas générée.

5.3.5 Détails des variables

Définition

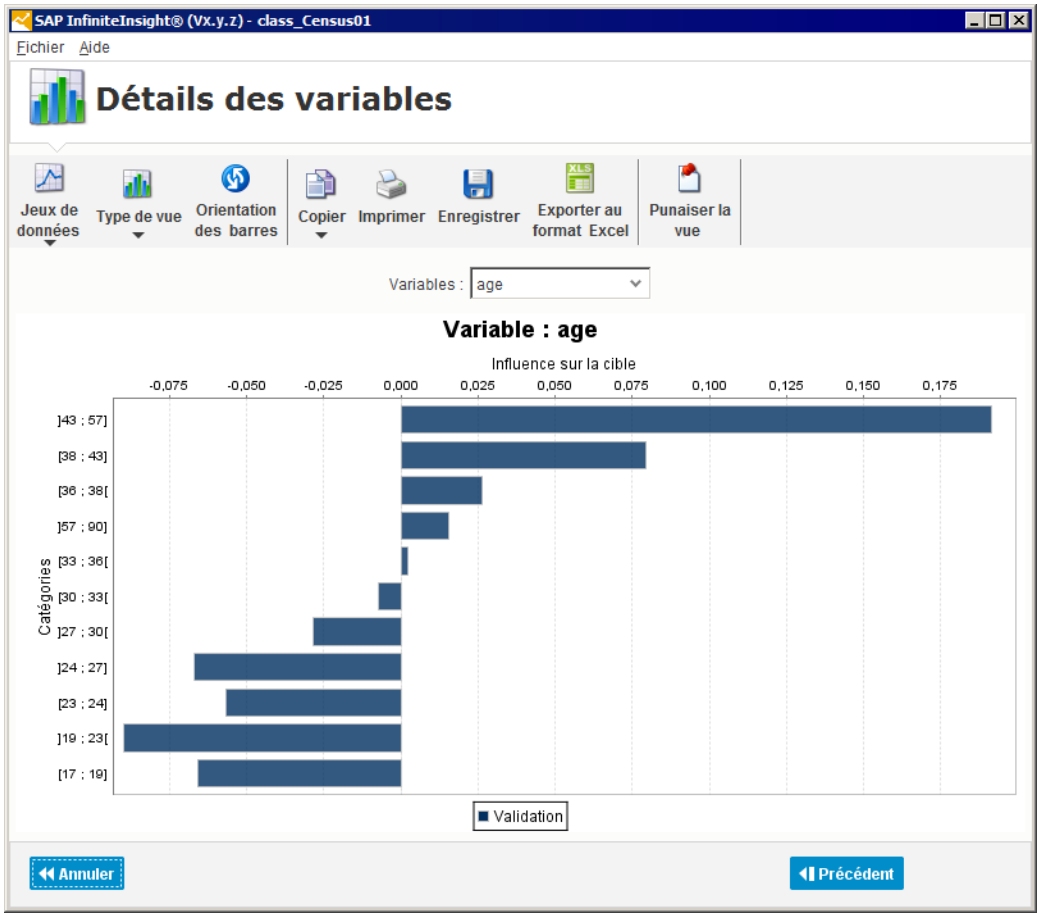
Le **graphique de détails de variable** présente l'importance des catégories d'une variable donnée par rapport à la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Afficher le graphique de détails d'une variable

☑ Pour afficher le graphique de détails d'une variable

1 Dans l'écran *Utilisation du modèle*, cliquez sur *Détails des variables*.
Le graphique de détails des variables apparaît.



2 Au-dessus du graphique, dans la liste *Variables*, sélectionnez la variable dont vous souhaitez afficher les catégories.

Si votre jeu de données contient des variables de type Date ou Datetime, des variables générées automatiquement apparaîtront dans cette liste. Pour plus d'information, reportez-vous à la section Variables de date : les variables générées automatiquement (voir "Variables de Date : les variables générées automatiquement" à la page 31).

Note

Vous pouvez afficher les détails d'une variable directement à partir du graphique *Contributions des variables*, en double-cliquant la barre représentant la variable qui vous intéresse.

Dans le cas où aucune structure utilisateur n'a été définie pour une variable continue, le graphe de détail des variables affiche les catégories créées automatiquement en utilisant le paramètre de *nombre de segments*. Le nombre de catégories affichées correspond à la valeur du paramètre de nombre de segments. Pour plus d'information au sujet de la configuration du paramètre de *nombre de segments*, reportez-vous à la section Nombre de segments pour les variables continues.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Options

En haut du panneau, une barre d'outils vous est proposée vous permettant de modifier l'affichage du graphique, de l'imprimer, copier ses données ou l'enregistrer.

Options d'affichage

☑ Pour afficher et masquer les sous-jeux d'Estimation et de Test

Cliquez sur *Jeux de données* et sélectionnez l'une des options suivantes :

-  *Tous les jeux de données*.
-  *Validation uniquement*.

☑ Pour afficher un histogramme

Cliquez sur *Type de vue* et sélectionnez  (*Histogramme*).
L'histogramme des catégories de la variable sélectionnée s'affiche.

☑ Pour afficher une courbe

Cliquez sur *Type de vue* et sélectionnez  (*Courbe de profit*).
La courbe de performances de la variable sélectionnée s'affiche.

☑ Pour ouvrir la vue courante dans une nouvelle fenêtre

Cliquez sur  (*Punaiser la vue*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Options d'utilisation

☑ Pour imprimer

1. Cliquez sur le bouton  (*Imprimer*).
   Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
2. Sélectionnez l'imprimante et les options d'impression.
3. Cliquez sur *OK*.
   L'impression est lancée.

☑ Pour enregistrer

1. Cliquez sur le bouton  (*Enregistrer*).
   Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
2. Entrez un nom de fichier.
3. Choisissez le dossier de destination.
4. Cliquez sur *OK*.
   Le graphique est enregistré au format PNG dans le dossier sélectionné.

☑ Pour copier

1. Cliquez sur le bouton  (*Copier*) et sélectionnez l'option désirée.
   L'application copie les paramètres du graphique.
2. Collez les paramètres dans l'application de votre choix. Vous pouvez par exemple les utiliser pour générer un graphique dans un tableur (Excel, ...).

☑ Pour exporter au format Excel

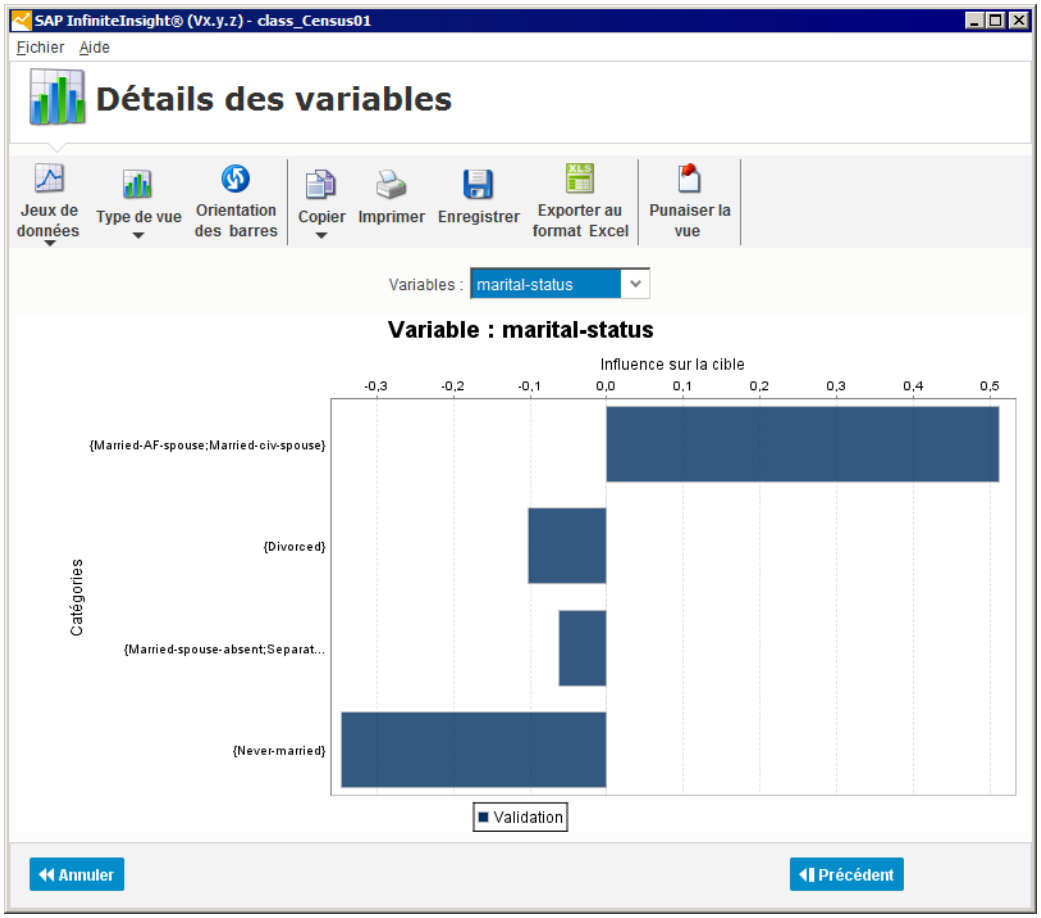
Cliquez sur  (*Exporter au format Excel*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre les graphiques de variables

Pour ce scénario

Sélectionnez la variable *marital-status*, qui est la variable explicative qui contribue le plus à la variable cible *Class*.



Ce graphique présente l'impact des catégories de la variable *marital-status* sur la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Catégories des variables et profit

Pour le graphique de détails d'une variable, le type de profit utilisé est **profit normalisé**, c'est-à-dire le profit qui permet de mesurer ce que le modèle apporte par rapport à un modèle de type aléatoire.

Sur ce type de graphique :

- **Plus une catégorie est située haut sur le graphique**, plus elle a un impact positif sur la catégorie cible (ou valeur souhaitée) de la variable cible. En d'autres mots, plus une catégorie est en haut sur le graphique, plus le taux de la catégorie cible est important dans cette catégorie..
- **La longueur d'une barre** correspond au profit apporté par la catégorie. Pour une catégorie donnée, une barre positive indique que cette catégorie contient plus d'observations appartenant à la catégorie cible de la variable cible que la moyenne. Une barre négative indique que la catégorie est moins concentrée en catégorie cible de la variable cible que la moyenne.

> **Note**
>
> Vous pouvez afficher les courbes de profit de la variable sélectionnée en cliquant sur le bouton  (*Courbe de profit*).

L'importance d'une variable dépend à la fois de sa différence par rapport à la moyenne de la catégorie cible et du nombre de cas représentés. Une importance élevée peut être le résultat :

- d'une forte divergence entre la catégorie et la moyenne de la catégorie cible de la variable cible,
- ou d'une faible divergence conjuguée à un grand nombre d'enregistrements dans cette catégorie,
- ou encore d'un mélange des deux.

La longueur de la barre montre le profit de cette catégorie. Les barres positives correspondent aux catégories ayant un nombre d'enregistrements supérieur à la moyenne de la catégorie cible, et les barres négatives correspondent aux catégories ayant un nombre d'enregistrements inférieur à la moyenne de la catégorie cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Axes du graphique

Les catégories des variables sont affichées sur l'**axe des ordonnées**. Les catégories ayant le même impact sur la variable cible sont regroupées. Elles apparaissent comme suit : `[Category_a;Category_b;Category_c]`. Les catégories ne contenant pas suffisamment de données pour fournir une information robuste sont regroupées dans la catégorie `KxOther`. Quand une variable a trop de valeurs manquantes, celles-ci sont regroupées dans la catégorie `KxMissing`. Ces deux catégories sont créées automatiquement par SAP InfiniteInsight®.

L'**axe des abscisses** montrent l'influence des catégories d'une variable sur la cible. La signification des différents nombres présents sur l'**axe des abscisses** est détaillée dans le tableau ci-dessous.

Le nombre est...	Indique que la catégorie a...
positif	une influence positive sur la cible
égal à 0	aucune influence sur la cible (le comportement est le même que le comportement moyen de l'ensemble de la population)
négatif	une influence négative sur la cible

Définition de l'importance des catégories

La définition ci-dessous s'applique aux cibles continues ; la formulation peut être en partie simplifiée pour les cibles binaires. Les formules suivantes peuvent également être appliquées au cas d'une cible binaire (dans ce cas, utilisez les catégories et non les segments).

Nous considérons le cas où un modèle de régression InfiniteInsight® Modeler / Régression ou Classement est utilisé en apprentissage sur une cible ou un signal continu S, à l'aide d'une variable d'entrée X.

InfiniteInsight® Modeler / Régression ou Classement segmente tout d'abord la cible continue S en B segments: $S_{1,\ldots,}S_{B}$ puis calcule les statistiques de base et les statistiques croisées des entrées par rapport à la cible.

Nous supposerons que l'entrée X est une variable nominale (catégorique), même si tout le processus peut être étendu facilement aux cas de variables ordinales ou continues.

Nous supposerons que X comporte N catégories : $X_1, \ldots, X_N$.

Nous souhaitons évaluer l'importance d'une catégorie Xi par rapport à la cible S.

L'importance d'une catégorie dépend de deux facteurs :

- le fait que la répartition de la cible pour cette catégorie est fortement biaisée en faveur de valeurs faibles ou élevées par rapport à la répartition de la cible sur l'ensemble de la population ;
- la fréquence de cette catégorie.

L'une des causes suivantes peut engendrer une importance de niveau élevé :

- une forte disparité entre la répartition de la cible pour les cas associés à cette catégorie et la répartition de la variable cible pour l'ensemble de la population ;
- une légère disparité combinée à un grand nombre d'enregistrements dans cette catégorie ;
- une combinaison des deux.

SAP InfiniteInsight® utilise un réglage non paramétrique où l'importance de la catégorie est définie ainsi :

$$CategoryImportance(Xi) = normalProfit(Xi) * \frac{Freq(Xi)}{Z}$$

où :

- *normalProfit(Xi)* correspond au profit standard de la catégorie Xi (voir la définition ci-dessous),
- *Freq (Xi)* correspond à la fréquence globale de la catégorie Xi,
- *Z* correspond à une constante de normalisation.

Nous indiquons ci-dessous le calcul détaillé de ces valeurs.

Profit standard

Chaque catégorie de la cible S_j est associée à un profit *profit(Sj)* défini ainsi :

$$\sum_{j=1}^{j=B} profit(S_j)\ Freq(Sj) = 0$$

Le profit d'une catégorie cible correspond à une valeur située dans la plage *[-1; +1]*. Il est défini de la manière suivante à partir des fréquences (cumulées) des catégories cibles:

$$profit(S_j) = 2\sum_{k=1}^{j-1} Freq(S_k) + Freq(S_j) - 1$$

Le profit standard d'une catégorie *Xi* est alors défini ainsi :

$$normalProfit(X_i) = \sum_{j=1}^{j=B} profit(S_j)\ Prob[S_j \mid X_i]$$

où *Proba[Sj |Xi]* correspond à la probabilité conditionnelle de voir apparaître la catégorie cible *Sj* dans la catégorie de la variable *Xi* (statistiques croisées) :

$$Prob[S_j \mid X_i] = \frac{Freq(S_j\ ;\ X_i)}{Freq(X_i)}$$

Ces formules reposant uniquement sur des fréquences, elles sont résistantes à toute transformation monotone de la cible *S*.

Constante de normalisation

La normalisation peut être approximative pour les cibles continues non pathologiques (c'est-à-dire les cibles continues sans pic de répartition (Dirac)), comme :

$$Z = P[S > median(S)] * (1 - P[S > median(S)])$$

Dans la plupart des cas, la valeur 0,25 constitue une bonne approximation.

Propriétés de profit standard

Plusieurs points sont à souligner au sujet du profit standard :

Le profit standard des catégories est indépendant des valeurs cibles en elles-mêmes (l'utilisateur peut modifier la valeur cible par le biais de transformations monotones ; le profit standard ne changera pas pour cette cible). Il s'agit de mesures non paramétriques.

Une conséquence du point 1 est que cette mesure est résistante aux valeurs aberrantes : s'il existe quelques occurrences de la cible dont la valeur est très élevée par rapport au reste de la répartition des valeurs cibles, la notion de profit standard n'est pas altérée.

La somme pondérée du profit standard pour toutes les catégories de variables données est toujours égale à 0.

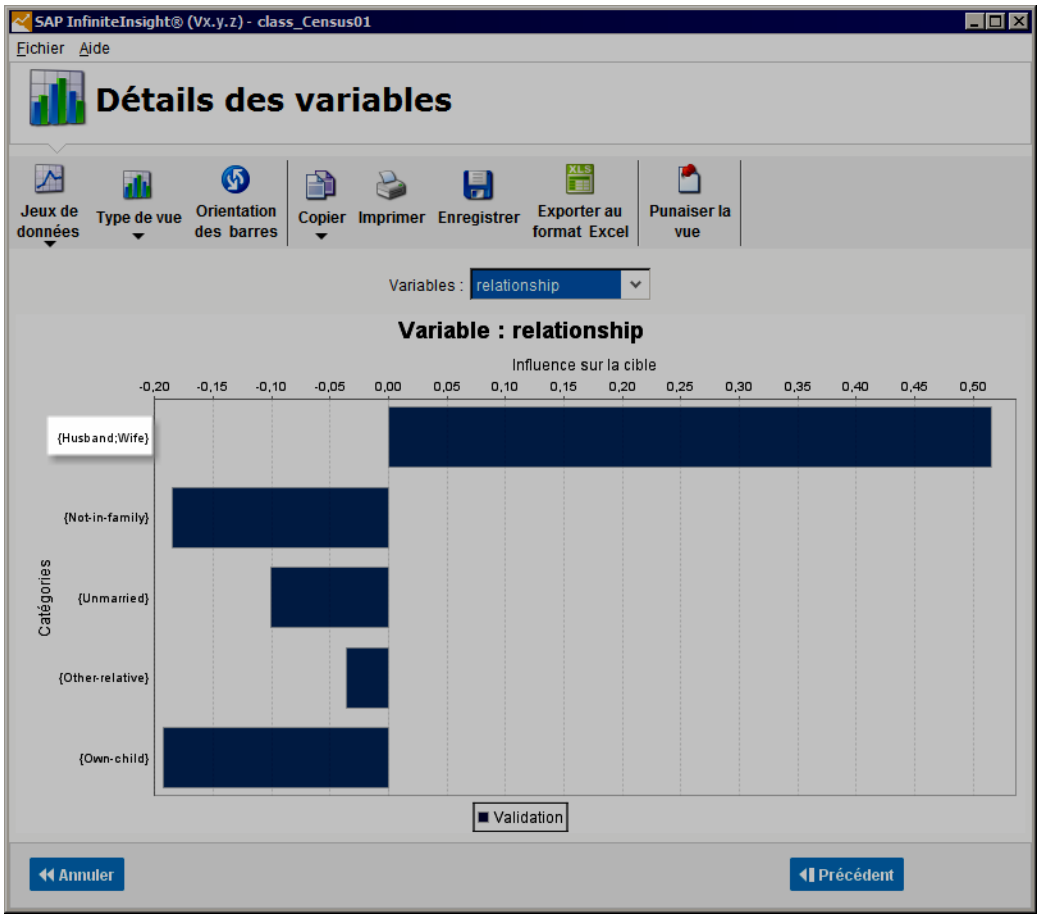
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Regroupement de catégories

Sur le graphique de détails d'une variable, des catégories peuvent apparaître groupées.

Quand l'option *Activer l'optimisation des regroupements basés sur la variable cible effectués par InfiniteInsight® Modeler / Codeur analytique pour toutes les variables* est activée, SAP InfiniteInsight® regroupe les **catégories ayant le même impact sur la variable cible**. Par exemple, pour la variable *relationship* (statut marital), les catégories *husband* (mari) et *wife* (femme) sont regroupées. Si la variable explicative est continue, SAP InfiniteInsight® repère les points de changements de comportement vis à vis de la variable cible et découpe ainsi automatiquement la variable en intervalles ayant un comportement homogène vis à vis de la cible.

Pour plus d'information, reportez-vous à la section Optimisation des regroupements.



Quand des **catégories ne sont pas assez représentées** pour apporter une information robuste, elles sont regroupées dans la catégorie *KxOther*, qui est alors automatiquement créée.

Quand une variable possède **trop de valeurs manquantes**, ces valeurs manquantes sont regroupées dans la catégorie *KxMissing*, alors automatiquement créée.

Pour comprendre l'intérêt des catégories *KxOther* et *KxMissing*, imaginons le cas suivant. La base de données des entreprises clientes d'une entreprise contient la variable "Adresse Web". Cette variable contient l'adresse du site Web des entreprises clientes référencées dans la base. Certaines entreprises possèdent une adresse Web, d'autres n'en possèdent pas. De plus, chaque adresse web est unique. Dans un tel cas, SAP InfiniteInsight® transforme automatiquement la variable "Adresse Web" en une variable binaire avec deux valeurs possibles : *KxOther* (l'entreprise a un site web) et *KxMissing* (l'entreprise n'a pas de site Web).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.3.6 Rapports de modélisation

SAP InfiniteInsight® vous propose un ensemble de *Rapports de modélisation* vous permettant une analyse fine de votre modèle. Ces tables sont regroupées en plusieurs niveaux :

- les *statistiques descriptives*, qui fournissent des statistiques sur les variables, leurs catégories et les jeux de données ainsi que les statistiques croisées des variables par rapport aux variables cibles.

> **Note**
>
> Si votre jeu de données contient des variables de type Date ou Datetime, des variables générées automatiquement apparaîtront dans ces rapports. Pour plus d'information, reportez-vous à la section Variables de date : les variables générées automatiquement (voir "Variables de Date : les variables générées automatiquement" à la page 31).

- les *performances du modèle*, dans lesquelles vous trouverez les indicateurs de performance du modèle, les individus non assignés, ainsi que les statistiques détaillées du score.
- la *vérification des déviations*, qui vous permet de vérifier la présence de déviation pour chaque variable et catégorie de variable entre les jeux de données de validation et de test.
- les *rapports avancés*, dans lesquels vous trouverez d'autres indicateurs de performance, l'encodage des variables, ...

Options des rapports de modélisation

Une barre d'outils vous est proposée vous permettant de modifier l'affichage du rapport courant, de le copier, l'imprimer, le sauvegarder ou l'exporter sous format Excel.

Options d'utilisation

Copier		Cette option permet de copier les données de la vue courante du rapport affiché. Les informations ainsi copiées peuvent être collées dans un éditeur de texte, un tableur, un document de traitement de texte.
		Si le rapport courant contient plusieurs vues (pour différentes variables, différents jeux de données, etc.) Cette option permet de copier l'ensemble des vues pour ce rapport.
		Si le rapport en cours est affiché sous forme de graphique, cette option vous permet de le copier au format image et de le coller dans un éditeur de texte ou dans un logiciel graphique.
Imprimer	Cette option permet d'imprimer la vue courante du rapport sélectionné selon le mode d'affichage choisi (rapport HTML, graphique, ...).	
Exporter		Cette option permet d'enregistrer sous différents formats (texte, html, pdf, rtf) les données de la vue courante du rapport affiché.
		Cette option permet d'enregistrer sous différents formats (texte, html, pdf, rtf) les données de l'ensemble des vues du rapport affiché.
		Cette option, qui est disponible pour toutes les formes d'affichage, permet d'exporter la vue courante vers Excel (compatible avec Excel 2002, 2003, XP et 2007).

		Cette option vous permet de sauvegarder tous les rapports.
		Cette option vous permet de sauvegarder la personnalisation des rapports.

Options d'affichage

 *Vue*		Cette option permet d'afficher la vue courante du rapport dans un tableau graphique qui peut être triés par colonne.
		Cette option permet d'afficher la vue courante du rapport sous forme de tableau HTML.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme d'histogramme. Cet histogramme peut être trié par ordre ascendant ou descendant des valeurs ainsi que par ordre alphabétique ascendant ou descendant. Vous pouvez également choisir quelles données afficher.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme de secteurs.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme de courbe.
 *Trier*		Quand le rapport en cours est affiché sous la forme d'un histogramme cette option vous permet de modifier son orientation (d'horizontal à vertical et inversement).
		Cette option vous permet d'afficher le rapport courant sans triage.
		Cette option vous permet de trier les valeurs du rapport courant par ordre ascendant.
		Cette option vous permet de trier les valeurs du rapport courant par ordre descendant.
		Cette option vous permet de trier les noms du rapport courant par ordre ascendant.
		Cette option vous permet de trier les noms du rapport courant par ordre descendant.
 *Séries*	Cette option permet de sélectionner quelles informations afficher dans le rapport courant.	

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.3.7 Carte des scores

Ce panneau vous fournit les coefficients associés à chaque catégorie pour toutes les variables d'un **modèle de regression**.

☑ Pour obtenir un score

Additionnez les coefficients correspondants à la valeur de chaque variable pour le cas étudié.

Remarque

Dans le cas d'une variable continue, la carte des scores comprend toujours un nombre de catégories supérieur à celui de la structure utilisateur définie ou du paramètre de *nombre de segments* si aucune structure utilisateur n'a été définie. En effet, l'encodage des variables pour la carte de score introduit des points de continuité pour augmenter la précision de codage par rapport au jeu de données d'apprentissage. Ces points de continuité scindent certaines catégories existantes et augmentent donc le nombre de catégories dans la carte de score.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

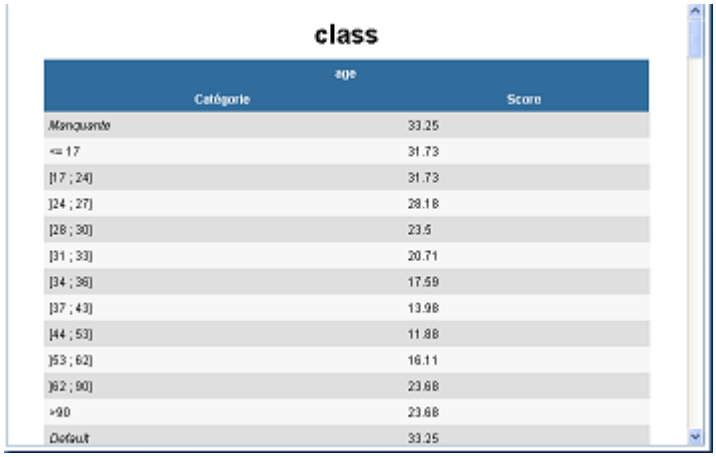
Mode "Risque"

La lecture de l'équation du modèle et l'interprétation de la carte de score sont facilitées dans le mode "Risque" en raison de l'encodage par palier pour les variables ordinales et continues.

En mode "Risque", il est facile d'identifier quelle catégorie a un effet positif ou négatif sur le score du risque, sur le ratio bons/mauvais ou sur la probabilité du risque.

Afin de mieux illustrer les avantages de la carte de scores pour l'interprétation des résultats, nous utilisons la variable "age" dans cet exemple.

Le segment]24;27] a un score de risque d'environ 30 et le segment [37;43] d'environ 15. Selon le paramètre PDO (points pour doubler le score, ici il vaut 15), on peut conclure que les individus appartenant au segment [37;43] sont deux fois plus risqués ou que le ratio bons/mauvais pour le segment [37;43] est deux fois moins élevé que pour le segment]24;27].



class

age	
Catégorie	Score
Manquante	33.25
<= 17	31.73
]17 ; 24]	31.73
]24 ; 27]	28.18
]28 ; 30]	23.5
]31 ; 33]	20.71
]34 ; 36]	17.59
]37 ; 43]	13.98
]44 ; 53]	11.88
]53 ; 62]	16.11
]62 ; 90]	23.68
>90	23.68
Défaut	33.25

Options de la carte des scores

Une barre d'outils en tête de panneau vous permet de copier le code HTML de la carte des scores, de l'enregistrer au format HTML ou de l'imprimer.

☑ Pour copier la carte des scores

1 Cliquez sur le bouton  (*Copier*).
L'application copie le code HTML correspondant à l'aperçu du modèle.
2 Collez les paramètres dans l'application de votre choix.

☑ Pour imprimer la carte des scores

3 Cliquez sur le bouton  (*Imprimer*).
Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
4 Sélectionnez l'imprimante et les options d'impression.
5 Cliquez sur *OK*.
L'impression est lancée.

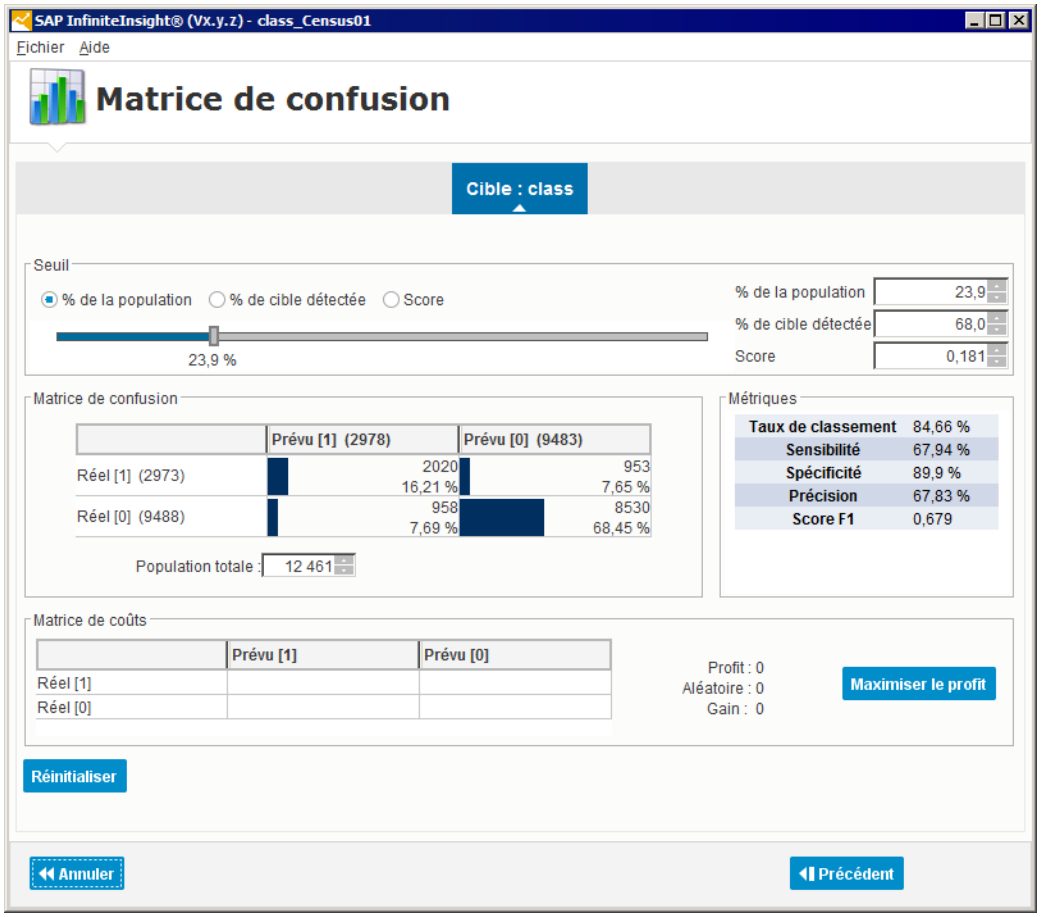
☑ Pour enregistrer la carte des scores

6 Cliquez sur le bouton  (*Enregistrer*).
Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
7 Entrez un nom de fichier.
8 Choisissez le dossier de destination.
9 Cliquez sur *OK*.
Les informations du modèle sont sauvegardées dans un fichier texte.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.3.8 Matrice de confusion

Le panneau *Matrice de confusion* permet de visualiser les valeurs de la cible prédites par le modèle par rapport aux valeurs réelles et de fixer le score à partir duquel les observations seront considérées comme positives, c'est-à-dire pour lesquelles la valeur de la cible est celle recherchée. Ce panneau vous permet également de faire des simulations de profit selon le score choisi comme seuil ou d'adapter automatiquement le seuil pour obtenir un profit maximal.



Définitions

On appelle "*Observation positive*", toute observation appartenant à la population cible.

On appelle "*Observation négative*", toute observation n'appartenant pas à la population cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre la matrice de confusion

Il y a trois façon de paramétrer le score utilisé pour séparer les observations positives des observations négatives en utilisant l'échelle affichée :

- en sélectionnant le **pourcentage de population visé** si la population est triée par ordre descendant de score (*% de la population*)
- en sélectionnant le **pourcentage d'observations positives** que vous souhaitez détecter (*% de cible détectée*)
- en sélectionnant directement le **score** à utiliser comme seuil (*Score*). Toute observation dont le score est supérieur au seuil est considérée comme positives et toute observation dont le score est inférieur au seuil est considérée comme négative.

L'échelle est graduée du plus petit score (à gauche), au plus grand (à droite). Les valeurs correspondant à chaque option sont affichées dans des champs situés sous l'échelle.

Lorsque vous déplacez le curseur sur l'échelle, la matrice de confusion est modifiée en conséquence. Le tableau ci-dessous indique comment lire la matrice de confusion.

	Prévu[Catégorie cible] Observations positives prédites	*Prévu[Catégorie non-cible]* Observations négatives prédites
Réel[Catégorie cible] Observations positives réelles	Nombre d'observations positives correctement prévues	Nombre d'observations réellement positives mais prédites négatives
Réel[Catégorie non-cible] Observations négatives réelles	Nombre d'observations réellement négatives mais prédites positives	Nombre d'observations négatives correctement prévues

Par défaut, la *Population totale* est égale au nombre d'enregistrements dans le jeu de données de validation. Vous pouvez modifier ce nombre pour visualiser la matrice sur la population sur laquelle vous voulez appliquer votre modèle.

Les Métriques

- Le *Taux de classement* correspond à la proportion de données correctement classée par le modèle lors de son application sur le jeu de données d'apprentissage.
- La *Sensibilité* d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée.
- La *Spécificité* d'un test mesure sa capacité à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.
- La *Précision* correspond à la proportion de mesures répétées à donner le même résultat, dans des conditions demeurant inchangées.

Le *Score* indique à quel point la fonction de vraisemblance dépend de son paramètre.
La fonction de vraisemblance est une fonction de probabilités conditionnelles qui décrit les valeurs d'une loi statistique en fonction de paramètres supposés connus.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre la matrice de coût

Cette section vous permet de visualiser votre profit selon le score choisi comme seuil ou de choisir automatiquement le meilleur seuil d'après vos paramètres.

Pour chaque catégorie d'observations, saisissez un profit ou un coût par observation. Le profit total s'affiche automatiquement à droite du tableau.

Pour connaître le seuil vous permettant d'atteindre un profit maximal pour le tableau de profit/coût que vous avez paramétré, cliquez sur le bouton *Maximiser le profit*.

Si on considère le tableau de profit/coût ci-dessous, chaque observation positive correctement identifiée rapportera 15€, par contre chaque observation négative identifiée comme étant positive coûtera 8€.

Catégorie	*Prévu[1]*	*Prévu[0]*
Réel[1]	15	0
Réel[0]	-8	0

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.3.9 Arbre de décision

Le panneau *Arbre de décision* vous permet d'afficher les résultats générés par InfiniteInsight® Modeler / Régression ou Classement sous la forme d'un arbre de décision basé sur les cinq variables les plus contributives.

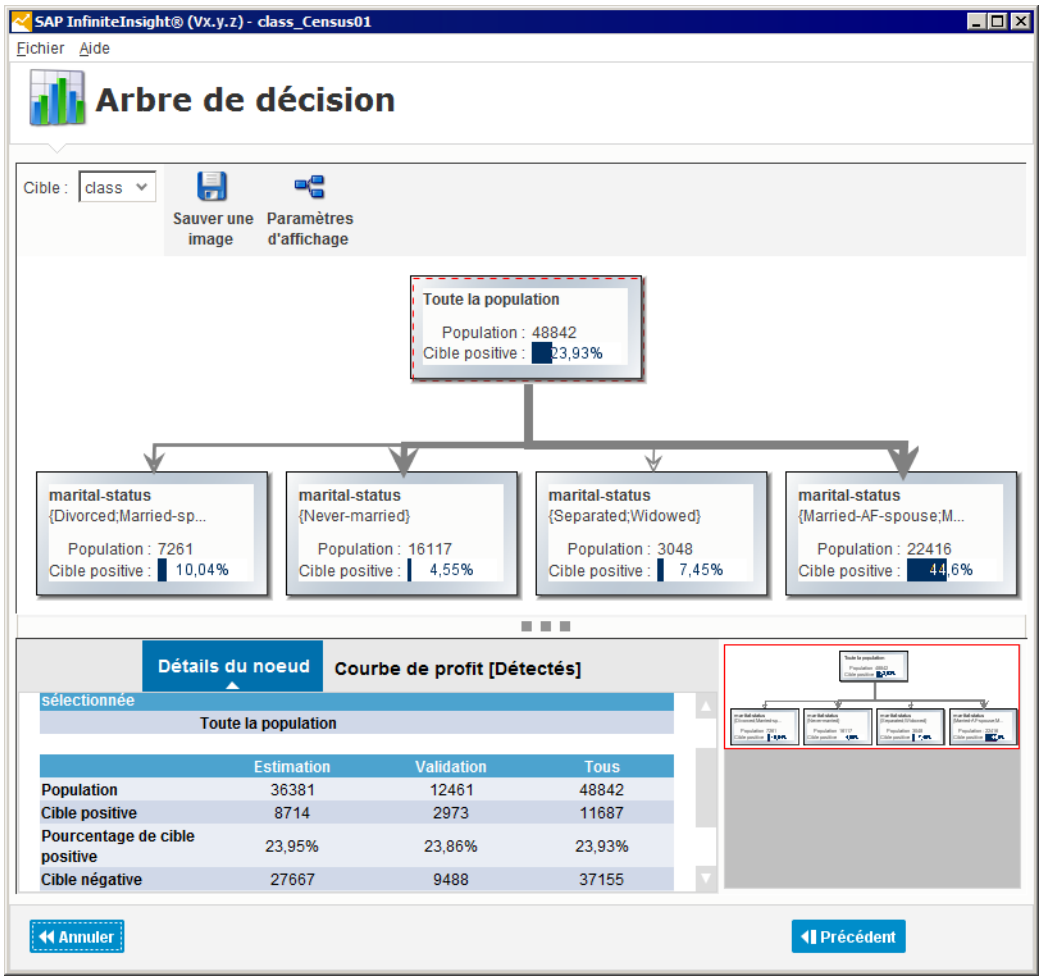
DANS CE CHAPITRE

 © 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Afficher l'arbre de décision

☑ Pour afficher l'arbre de décision pour une variable cible

1 Dans la liste *Cible*, choisissez la variable cible pour laquelle vous souhaitez afficher l'arbre de décision.

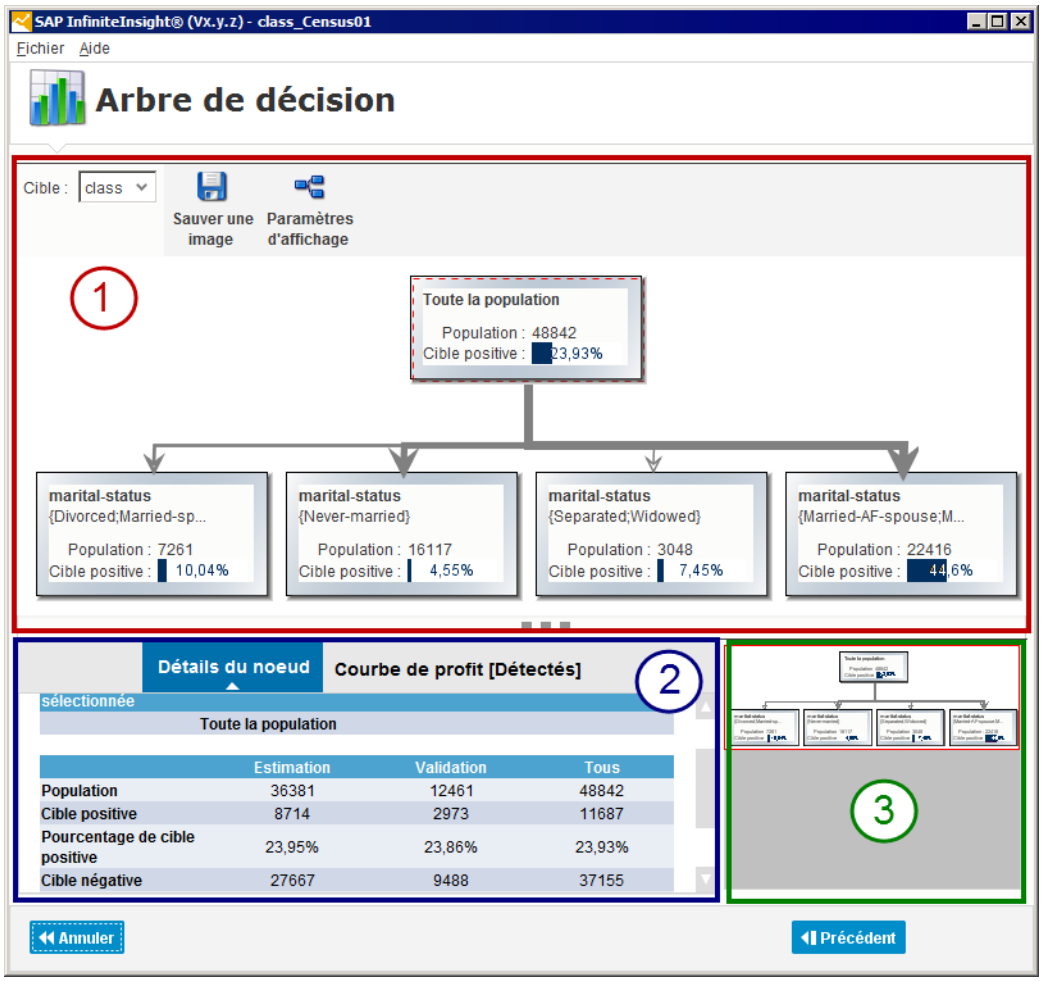


© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre l'arbre de décision

Le panneau *Arbre de décision* est divisé en trois parties :

1. l'**arbre de décision** lui-même, affiché en première partie du panneau,
2. deux onglets situés en bas du panneau vous permettent de visualiser les **informations des noeuds** ainsi que la **courbe de profit** correspondant à l'arbre de décision affiché.
3. une **fenêtre de navigation**, située en bas à droite du panneau, vous permet de visualiser quelle section de l'arbre vous être en train d'étudier.

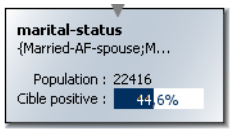
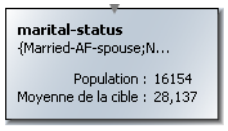


© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

L'arbre de décision

Chaque noeud de l'arbre indique :

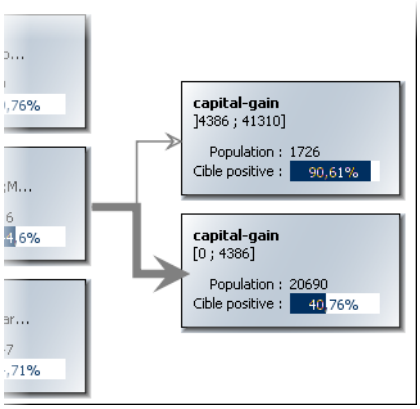
- Le nom de la variable déployée, par exemple `Marital-status`.
- Les catégories ayant servi à filtrer la population du noeud, par exemple `{Married-AF-spouse;Never-married}`.
- La *Population* totale du noeud.
- Le pourcentage de *Cible positive* (pour une cible nominale) ou la *Moyenne de la cible* (pour une cible continue).

Exemple pour une cible nominale	*Exemple pour une cible continue*
	

Lorsque vous survolez un noeud, plusieurs options sont disponibles :

	Sélectionner une variable pour déployer le niveau suivant de l'arbre de décision.
	Déployer automatiquement le niveau suivant, en fonction de la variable la plus contributive non encore utilisée dans l'arbre de décision.
	Replier la section affichée sous le noeud.

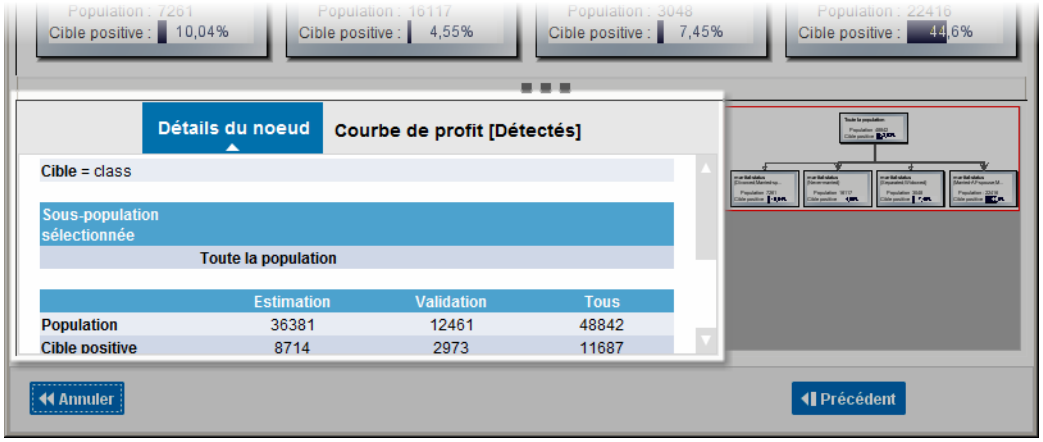
L'épaisseur des flèches est relative à la quantité de population contenue dans le noeud pointé. Dans l'exemple suivant, la flèche pointant le noeud correspondant à la catégorie `[0;4386[` de la variable `capital-gain` est significativement plus épaisse car ce noeud contient une population nettement plus importante que le noeud `capital-gain ]4386;41310]`.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Le détail des noeuds

Lorsque vous sélectionnez un noeud, les informations correspondantes s'affichent dans l'onglet *Détails du noeud* (partie inférieure gauche du panneau).



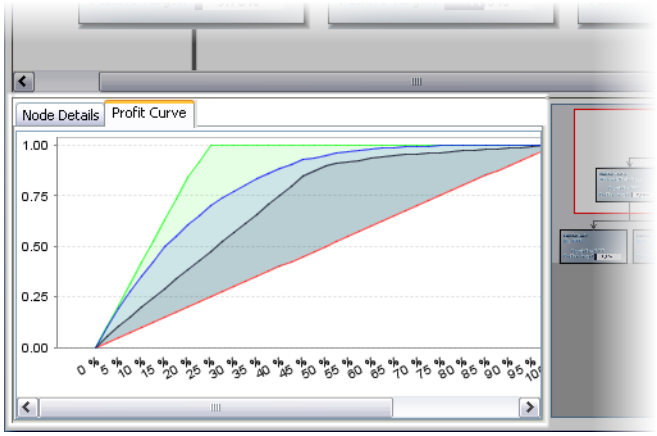
Cet onglet indique pour quelle cible l'arbre de décision est déployé et vous fournit les informations suivantes pour chaque jeu de données du modèle :

- *Population*, c'est-à-dire le nombre d'enregistrements existants pour le noeud,
- Pour une cible continue :
  - *Moyenne de la cible*, c'est-à-dire la moyenne de la cible pour le noeud
- Pour une cible nominale :
  - *Cible positive*, c'est-à-dire le nombre d'enregistrements pour lesquels la cible est positive
  - *Pourcentage de cible positive*, c'est-à-dire le pourcentage de la population du noeud pour laquelle la cible est positive,
  - *Cible négative*, c'est-à-dire le nombre d'enregistrements pour lesquels la cible est négative,
  - *Pourcentage de cible négative*, c'est-à-dire le pourcentage de la population du noeud pour laquelle la cible est négative,
  - la *Variance*,
  - *Population pondérée*, c'est-à-dire le nombre d'enregistrements lorsque une variable de poids est utilisée.

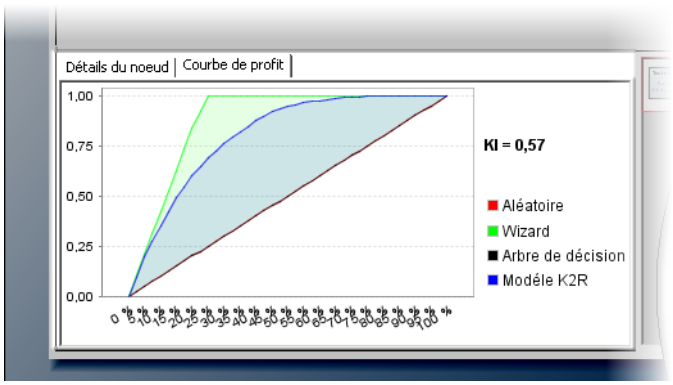
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

La courbe de profit

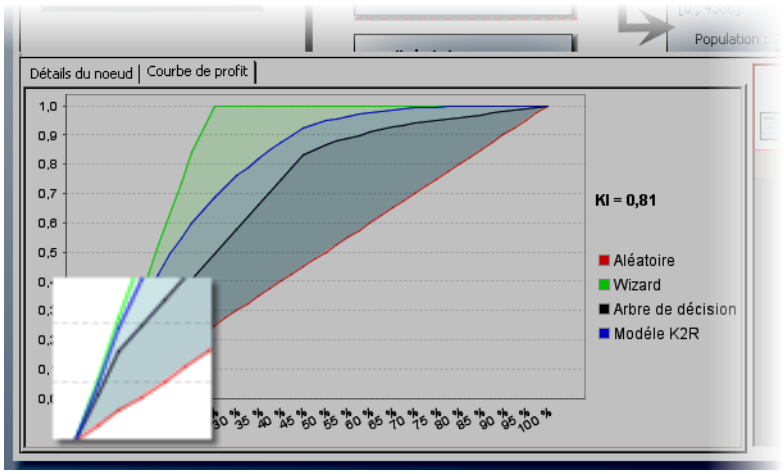
La courbe de profit pour l'arbre de décision est affichée dans l'onglet *Courbe de profit* (partie inférieure gauche du panneau). La courbe évolue en fonction des modifications faites sur l'arbre de décision.



La courbe de profit correspondant au noeud qui contient la population totale est égale à la courbe aléatoire.

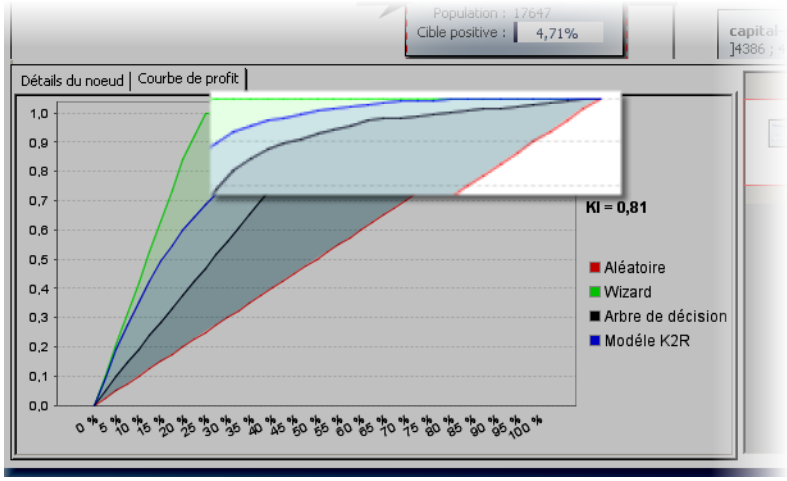


Lorsque vous développé le noeud contenant **le plus haut pourcentage de cible positive, la courbe de profit s'améliorera sur les premiers percentiles**, c'est-à-dire que le modèle détectera d'avantage de cas dans la population ayant les plus hauts scores.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Au contraire, si vous développez le noeud contenant le plus faible pourcentage de cible positive, la courbe de profit s'améliorera sur les derniers percentiles.

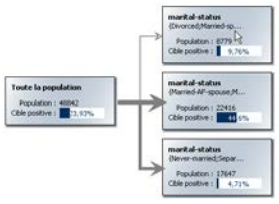
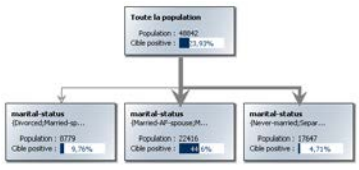


Cependant, si le noeud que vous développez correspond à une portion très faible de la population, la courbe de profit risque de ne pas être impactée. Il faut donc trouver le bon compromis entre la taille de la population et le pourcentage de cible positive.

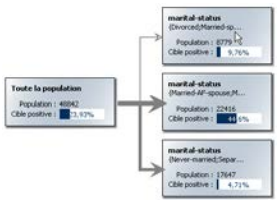
Paramétrer l'affichage

Le bouton *Paramètres d'affichage* vous permet de personnaliser l'affichage de l'arbre de décision.

- *Orientation* : cette option vous permet de définir l'orientation de l'arbre, horizontale ou verticale.

Horizontale	*Verticale*
	

- *Type d'affichage* : cette option vous permet de choisir entre un affichage standard (*Arbre de décision*) et un affichage en mode K2R (*Mode K2R*). L'affichage en *Arbre de décision* est plus condensé, mais moins lisible que l'affichage en *Mode K2R*.

Mode K2R	*Arbre de décision*
	

Une fois vos paramètres d'affichage définis, cliquez sur *Fermer*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

5.4 Etape 4 - Utiliser le modèle

Une fois généré, un modèle de classement peut être **enregistré** pour utilisation ultérieure.

Un modèle de classement peut être **appliqué sur de nouveaux jeux de données**. Le modèle vous permet alors d'effectuer des prédictions sur ces jeux de données d'application, en prédisant les valeurs d'une variable cible. Le modèle peut également être utilisé pour effectuer des **simulations** sur des observations spécifiques, au cas par cas.

Enfin, vous pouvez **affiner** un modèle de classement, en le générant à nouveau avec une liste optimisée de variables explicatives. SAP InfiniteInsight® vous permet en effet de sélectionner de manière automatique les variables explicatives les plus pertinentes par rapport à votre problématique, en fonction du taux d'information expliqué par le modèle que vous souhaitez conserver.

Pour vous permettre d'appliquer le modèle sur n'importe quelle base de données, SAP InfiniteInsight® permet de **générer les codes source** du modèle.

5.4.1 Vérification des déviations

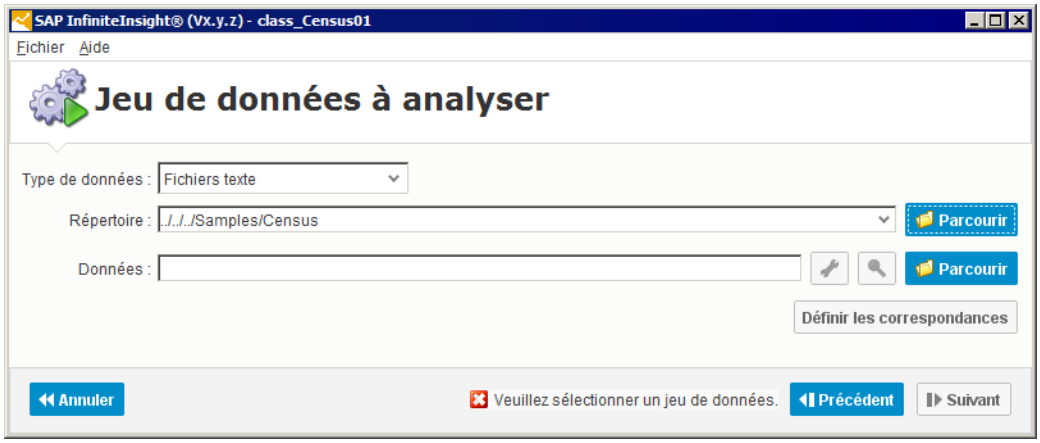
L'option *Vérification des déviations* est un outil de diagnostic des variations statistiques des variables.

Cette option peut être utilisée pour :

- comparer la distribution d'un nouveau jeu de données avec celle du jeu de données utilisé pour créer le modèle,
- vérifier la qualité de nouvelles données après les avoir chargées,
- vérifier si vos données ont évoluées au cours du temps et si nécessaire générer un modèle mieux adapté aux nouvelles données.

☑ Pour commencer la vérification des déviations

1 Dans la section *Exécution* du menu *Utilisation du modèle*, cliquez l'option *Vérification des déviations*. Le panneau de sélection du jeu de données à vérifier s'affiche.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

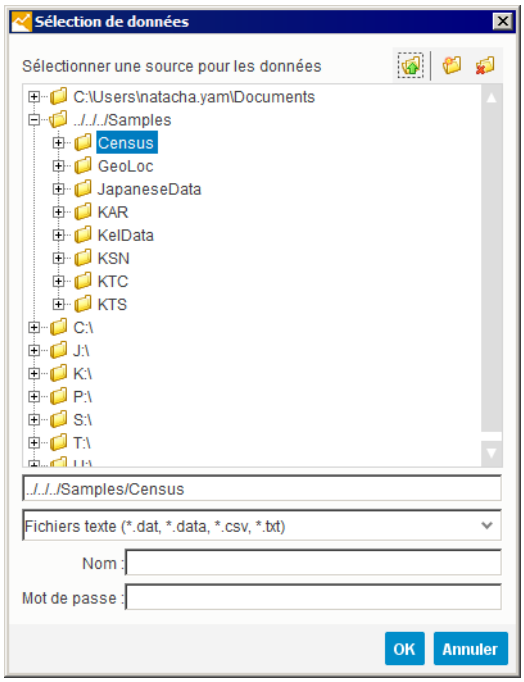
Sélectionner le jeu de données à analyser

Avant tout, vous devez sélectionner le jeu de données pour lequel vous souhaitez analyser les déviations.

Pour que les résultats soient compréhensibles, le nouveau jeu de données doit contenir les même colonnes que le jeu de données utilisé pour générer le modèle, en particulier la variable cible, qui doit être renseignée.

☑ Pour sélectionner un jeu de données

1. Dans le panneau Jeu de données à analyser, sélectionnez le format de la source de données (*Fichiers texte*, *Base de données*, ...)
2. Cliquez sur le bouton *Parcourir* à droite du champ *Répertoire*. La boîte de dialogue suivante s'affiche.

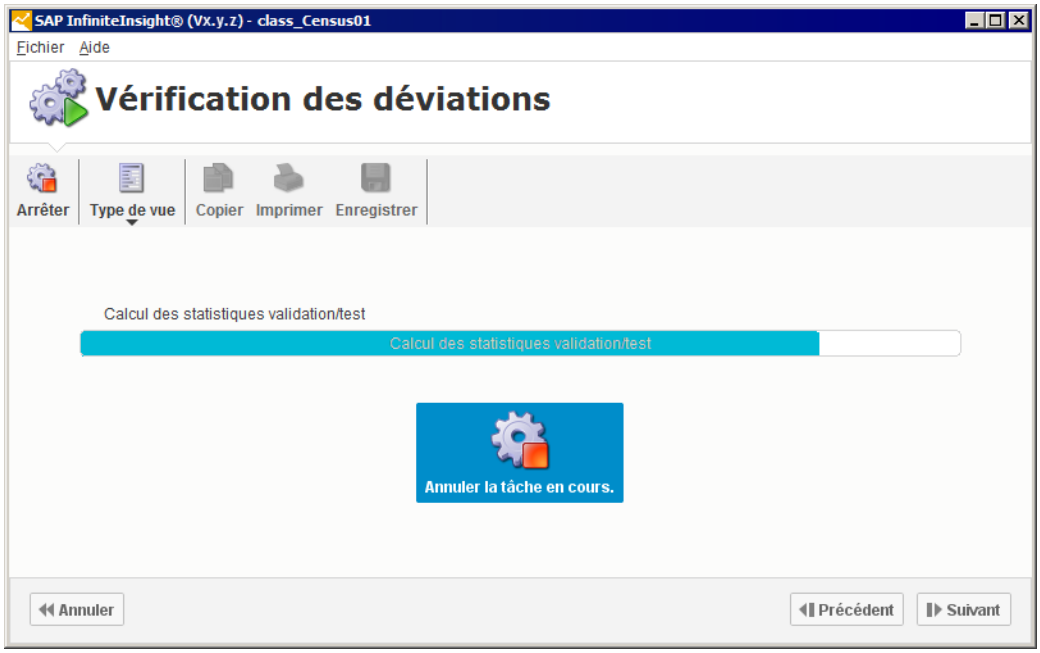


3. Ouvrez le répertoire ou la base de données contenant la source de données.
4. Sélectionnez le fichier ou la table à utiliser comme source de données.
5. Cliquez sur le bouton *OK*. La boîte de dialogue se ferme et le nom de la source de données apparaît dans le champ *Données*.
6. Cliquez sur le bouton *Suivant*. Le panneau *Vérification des déviations* s'affiche.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Suivi du processus de vérification des déviations

Le panneau *Vérification des déviations* vous permet de suivre le processus d'analyse grâce à une barre de progression.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

A la fin de la vérification, un panneau récapitulatif s'affiche. L'explication détaillée du panneau récapitulatif est fournie dans la section *Comprendre l'analyse des déviations*.

Vous pouvez utiliser la barre d'outil affichée en haut du panneau pour :

- stopper l'analyse, en cliquant sur le bouton  ,
- afficher les détails du processus, en cliquant sur le bouton  ,
- copier, imprimer ou enregistrer le panneau récapitulatif.

☑ Pour copier

1. Cliquez sur le bouton  (*Copier*).
   L'application copie le code HTML du rapport affiché.

☑ Pour imprimer

1. Cliquez sur le bouton  (*Imprimer*).
   Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
2. Sélectionnez l'imprimante et les options d'impression.
3. Cliquez sur *OK*.
   L'impression est lancée.

☑ Pour enregistrer

1. Cliquez sur le bouton  (*Enregistrer*).
   Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
2. Entrez un nom de fichier.
3. Choisissez le dossier de destination.
4. Cliquez sur *OK*.
   Le rapport est enregistré au format HTML dans le dossier sélectionné.

Comprendre l'analyse des déviations

La première chose à faire pour savoir s'il y a des déviations dans vos données est de regarder le **rapport récapitulatif** (voir à la page 151) et de comparer les performances (KI et KR) obtenues sur le jeu de données original avec celles obtenues sur le jeu de données de contrôle.

Ensuite pour visualiser quelles variables ont changé, regardez les rapports de déviations.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Rapport récapitulatif

La partie *Vérification des déviations* fournit des statistiques de base sur le *Jeu de données utilisé pour le contrôle des déviations* (ou **Jeu de données de contrôle**) telles que :

- le nom du jeu de données (*Jeu de données*),
- la source de données (*Source*),
- le nombre d'enregistrements contenus dans le jeu de données (*Nombre d'enregistrements*),
- et le nombre de variables pour lesquelles SAP InfiniteInsight® a trouvé des déviations par rapport au jeu de données utilisé pour créer le modèle (*Nombre de variables montrant des déviations*).

La deuxième et la troisième parties du rapport vous permettent de comparer les performances de votre modèle sur le jeu de données original avec ses performances sur le jeu de données de contrôle :

- la section *Indicateurs de performance* affiche pour chaque variable cible, les indicateurs *KI* et *KR* obtenus par le modèle sur le jeu de données original.
- la section *Performance sur le jeu de contrôle* affiche pour chaque variable cible, les indicateurs *KI* et *KR* obtenus par le modèle sur le jeu de données de contrôle.

Si le KI et/ou le KR du modèle sur le jeu de données de contrôle sont significativement **plus faibles** cela signifie que la relation entre les variables et la variable cible a changé, et en conséquence u**n nouveau modèle devrait être généré** sur les nouvelles données.

Si le KI et le KR n'ont pas ou peu changé, cela signifie que le relation entre les variables et la variable cible est toujours la même, mais cela ne signifie pas qu'il n'y a aucune différence de distribution entre les jeux de données.

5.4.2 Appliquer un modèle sur un nouveau jeu de données

Le modèle en cours d'utilisation peut être appliqué sur de nouveaux jeux de données. Le modèle permet alors d'effectuer des prédictions sur ces jeux de données d'application, en prédisant notamment les valeurs de la variable cible.

Pour ce scénario

Pour des contraintes d'ordre technique, un jeu de données correspondant à la base de données de 1 000 000 de clients dont il est question pour ce scénario ne peut pas vous être fourni.

Vous allez donc appliquer le modèle sur le fichier *Census01.csv*, que vous avez utilisé pour générer le modèle. Vous pourrez ainsi comparer les prédictions données par le modèle aux valeurs réelles de la variable cible *Class* de chacune des observations.

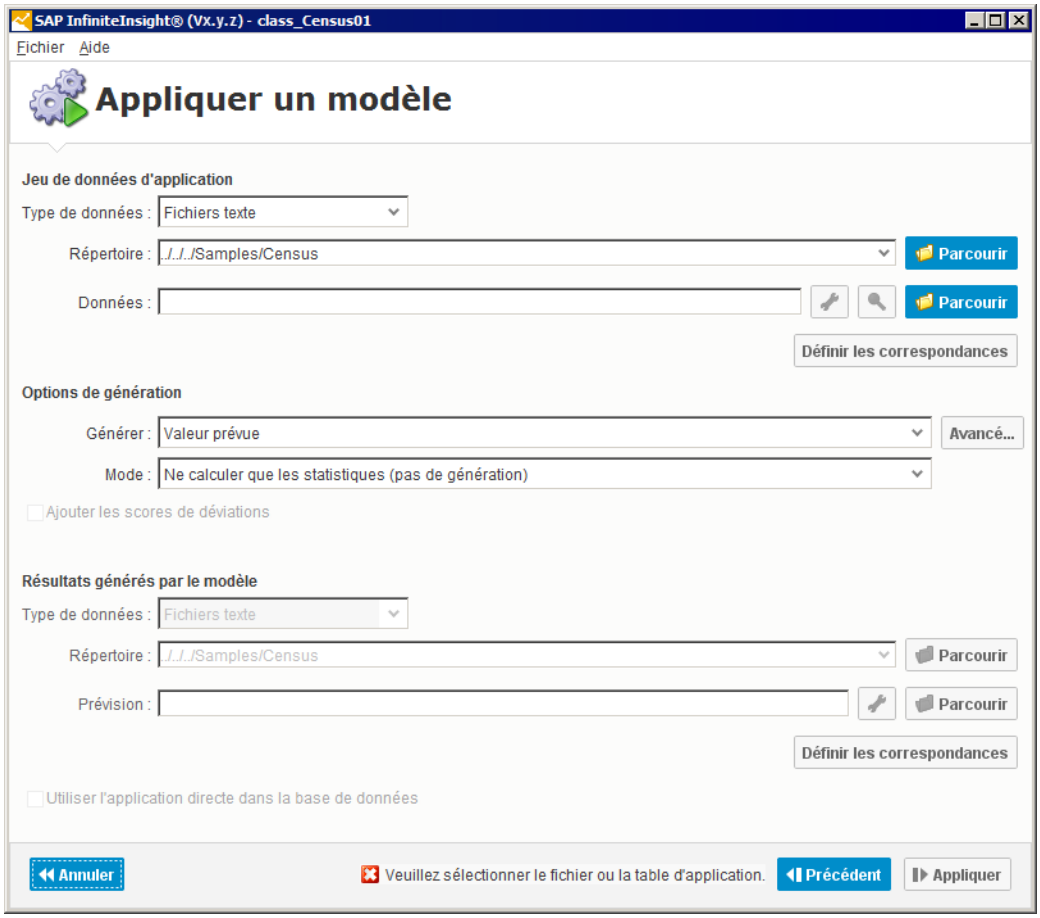
Dans la procédure *Pour appliquer le modèle sur un nouveau jeu de données* :

- Sélectionnez le format *Fichiers texte*,
- Dans le champ *Générer*, sélectionnez l'option *Contributions individuelles*.
- Sélectionnez un répertoire de votre choix pour enregistrer le fichier de résultats (*Sortie générée par le modèle*).
- Ne sélectionnez pas l'option *Conserver uniquement les observations déviantes*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Pour appliquer le modèle sur un nouveau jeu de données

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Application du modèle*.
L'écran *Appliquer un modèle* apparaît.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

2 Dans la partie *Jeu de données d'application*, sélectionnez le format de la source de données.

3 Cliquez sur les boutons *Parcourir* pour indiquer respectivement :
- dans le champ *Répertoire*, le répertoire dans lequel est stocké votre jeu de données,
- dans le champ *Données*, le nom du fichier correspondant à votre jeu de données.

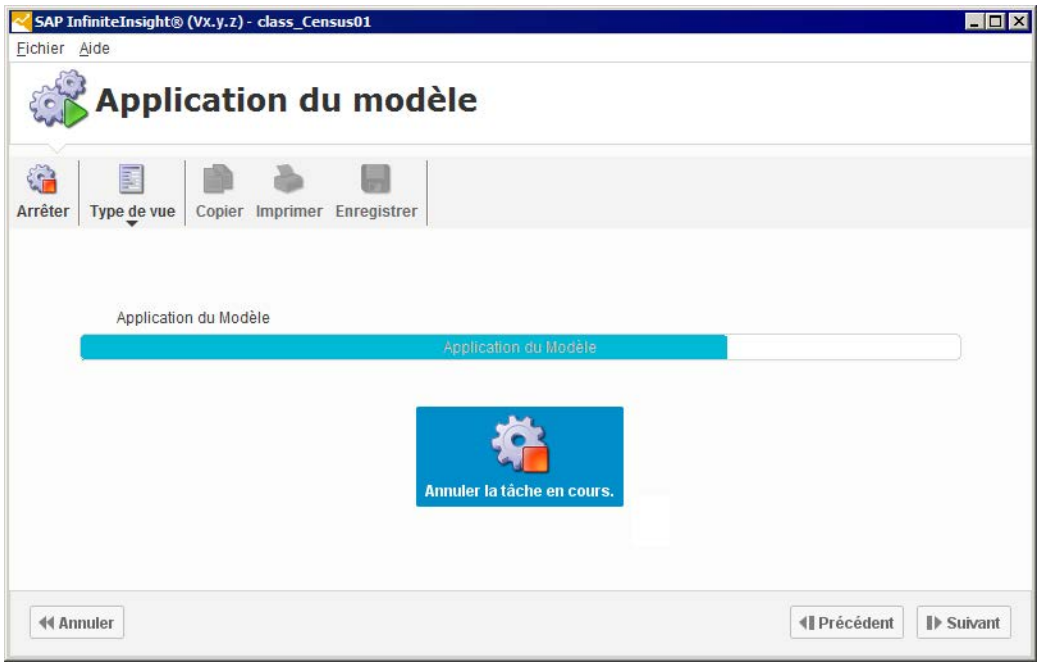
4 Dans le cadre *Options de génération*, sélectionnez dans la liste *Générer* le type de valeurs de sortie que vous souhaitez obtenir pour la variable cible.

5 Sélectionnez dans la liste *Mode*, le type de résultats voulu.

6 Dans le cadre *Résultats générés par le modèle*, sélectionnez le format du ficher de sortie

7 Cliquez sur le bouton *Appliquer*.

L'écran *Application du modèle* apparaît.
Une fois l'application du modèle terminée, le fichier de résultats de l'application est automatiquement enregistré à l'emplacement que vous avez défini sur l'écran *Appliquer le modèle*.



Contrainte d'utilisation d'un modèle

Pour qu'un modèle puisse être appliqué sur un jeu de données, le format du jeu de données d'application doit être identique à celui du jeu de données d'apprentissage utilisé pour générer le modèle. La même variable cible doit notamment être contenue dans les deux jeux de données, même si ses valeurs ne sont pas renseignées dans le jeu de données d'application.

Note

Si la variable *KxIndex* du modèle est virtuelle, l'espace de données d'application ne doit pas contenir de variable *KxIndex* physique.

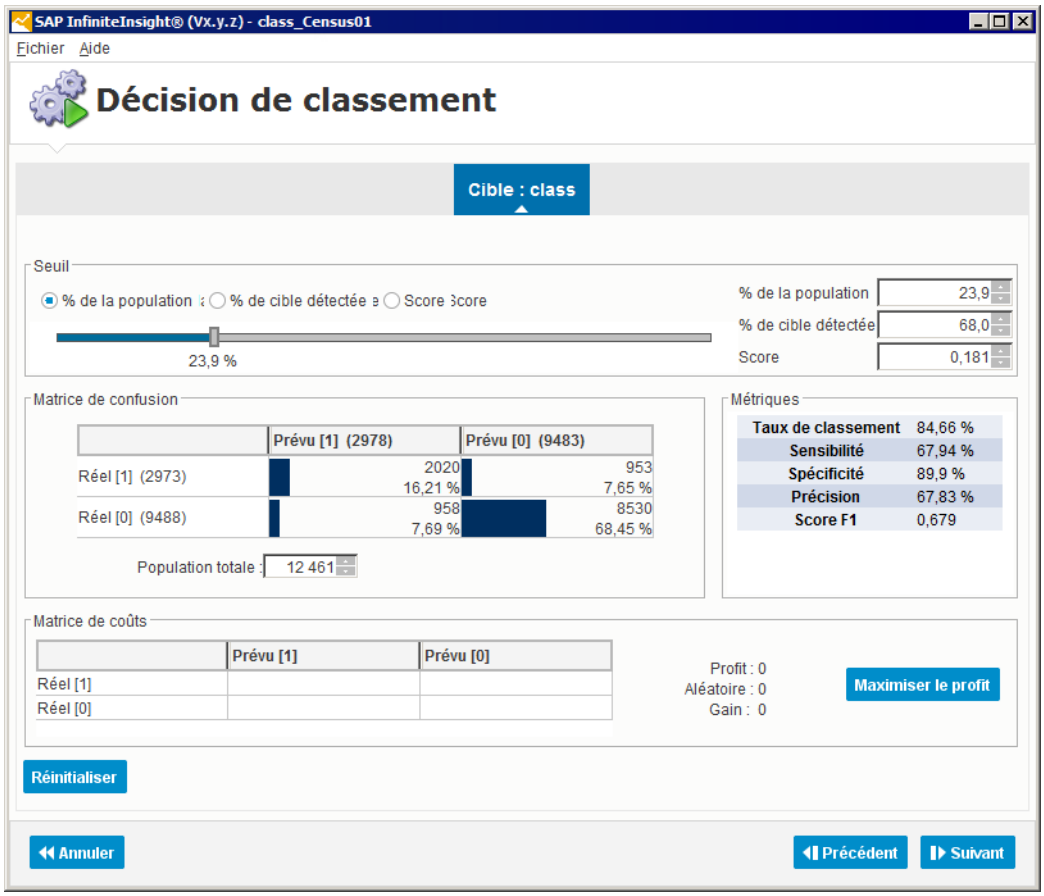
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Décision de classement

L'écran *Décision de classement* vous permet de choisir le nombre d'observations que le modèle doit détecter lors de l'application sur le nouveau jeu de données.

☑ Pour appliquer une décision de classement

1. Dans l'écran *Appliquer un modèle*, suivez les étapes de la procédure **Pour appliquer un modèle sur un nouveau jeu de données**.
2. Sélectionnez l'option *Décision* dans la liste déroulante *Générer*.
3. Cliquez sur le bouton *Appliquer*.
   L'écran *Décision de classement* s'affiche.



4. Utilisez le curseur pour choisir le pourcentage désiré. Pour plus d'information, reportez-vous à la section Matrice de confusion à la page 137.
5. Cliquez sur le bouton *Suivant*.
   Le modèle est appliqué sur le nouveau jeu de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Comprendre l'écran Décision de classement

L'écran *Décision de classement* vous permet de sélectionner un pourcentage de la population répondant positivement à votre campagne (*% de cible détectée*) ou un pourcentage de la population totale de votre jeu de données (*% de la population*).

Lorsque vous déplacez le curseur sur l'échelles, les différentes valeurs affichées sous l'échelle sont mises à jour.

Par exemple, si vous sélectionnez l'option *% de cible détectée* et placez le curseur de l'échelle sur 80%, la valeur du champ *% de la population* sera égale à *32.0*, ce qui signifie que si vous voulez que 80% des personnes qui répondront positivement à votre campagne reçoivent votre mailing, vous devrez l'envoyer à 32% de la population totale.

D'un autre côté, si vous sélectionnez l'option *% de la population* et placez le curseur de l'échelle sur 20%, la valeur du champ *% de cible détectée* sera égale à *60.4*, ce qui signifie que si votre budget ne vous permet d'envoyer votre mailing qu'à 20% de la population totale du jeu de données, vous atteindrez 60% des personnes qui répondront de façon positive.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Utiliser l'application directe dans la base de données

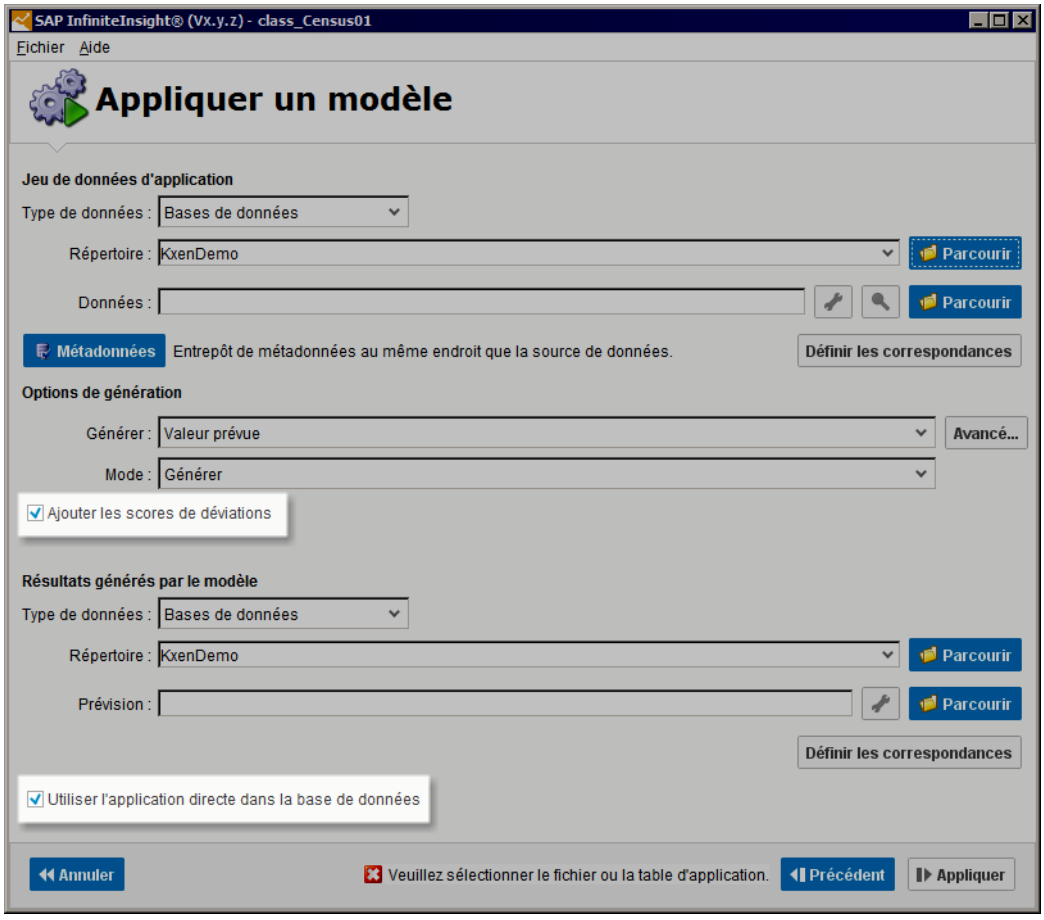
Pré-requis pour l'utilisation du mode d'application direct dans la base de données

Ce mode optimisé du score peut être utilisé si toutes les conditions suivantes sont remplies:

- le jeu de données d'application (table, vue, requête, manipulation de données) et les résultats du jeu de données sont des tables provenant de la même base de données,
- le modèle calculé contient au moins une variable avec une clé physique pré-définie dans SAP InfiniteInsight®,
- une licence InfiniteInsight® Scorer valide,
- aucune erreur apparue,
- un mode d'application dans la base de données activé,
- un accès de lecture et d'écriture (créer une table).

☑ Pour utiliser le mode d'application directe dans la base de données

Cochez l'option *Utiliser l'application directe dans la base de données*, l'option *Ajouter les scores de déviations* est automatiquement cochée.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Paramètres avancés

Sorties globales

Copier la variable de poids

Cette option vous permet d'ajouter au fichier de sortie la variable de poids si elle a été définie lors de la sélection des variables du modèle.

Copier l'identifiant de jeu de données

Cette option vous permet d'ajouter au fichier de sortie le nom du sous-jeu de données d'apprentissage auquel appartient l'enregistrement (Estimation, Validation ou Test).

Attention

Cette option n'est pas compatible avec l'application directe en base de données.

Copier les variables

Cette option vous permet d'ajouter au fichier de sortie une ou plusieurs variables du jeu de données.

☑ **Pour ajouter toutes les variables du jeu de données**

Cochez l'option *Toutes*.

☑ **Pour sélectionner uniquement les variables qui vous intéressent**

1. Sélectionnez l'option *Sélection*.
2. Cliquez sur le bouton >> pour afficher le tableau de sélection des variables.
3. Sélectionnez dans la liste *Eléments disponibles* les variables que vous voulez ajouter (utilisez la touche *Ctrl* pour sélectionner plusieurs variables à la fois).
4. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Eléments sélectionnés*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Constantes définies par l'utilisateur

Cette option vous permet d'ajouter au fichier de sortie des constantes comme par exemple la date de l'application du modèle, le nom du jeu de données utilisé, ou toute autre information utile pour l'exploitation du fichier de sortie.

Une constante est définie par les informations suivantes:

Paramètre	Description	Valeur
Générer	indique si la constante sera générée dans le jeu de données de sortie.	*coché*: la constante sera générée *décoché*: la constante ne sera pas générée
Nom	nom de la constante	1 Le nom ne peut être identique à celui d'une variable du jeu de données de référence. 2 Si le nom est identique à celui d'une constante existante, celle-ci sera remplacée par la nouvelle constante.
Format	type de la constante	*number*: nombre *string*: chaîne de caractères *integer*: entier *date*: date *datetime*: date et heure
Valeur	valeur de la constante	format des dates: `YYYY-MM-DD` format des dates avec horaire: `YYYY-MM-DD HH:MM:SS`
Clé	spécifie si la constante est une variable clé ou un identifiant de l'enregistrement. Il est possible de déclarer des clés multiples qui seront construites selon l'ordre indiqué (1-2-3-...).	*0*: la constante n'est pas un identifiant *1*: identifiant primaire *2*: identifiant secondaire ...

☑ Pour définir une constante

1. Cliquez sur le bouton *Ajouter*. Une fenêtre s'ouvre vous permettant de saisir les paramètres de la constante.
2. Dans le champ *Nom*, saisissez le nom de la constante.
3. Dans la liste *Format de sortie*, sélectionnez son type.
4. Dans le champ *Valeur de sortie*, saisissez la valeur que vous souhaitez donner à la constante.
5. Cliquez sur le bouton *OK* pour valider la création de la constante. La nouvelle constante apparaît dans la liste. Vous pouvez choisir de générer ou non les constantes définies en cochant la case *Générer* correspondante.

Table de profit

Ce panneau vous permet de calculer la table de profit pour le jeu de données d'application, c'est-à-dire de trier vos données par ordre de score décroissant et de les répartir de façon égale en quantiles (déciles, vingtiles ou centiles).

Si vous avez calculé la table de profit lors de la création de votre modèle, deux tables de profits seront calculées pendant l'application du modèle :

- Une table de profit de type Transversal qui vous permet de vérifier les déviations entre la table de profit sur les données d'apprentissage et la table de profit sur les données d'application,

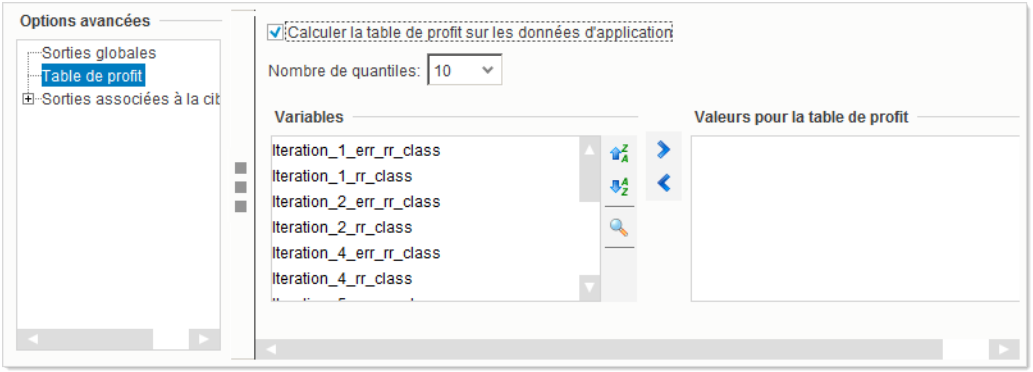
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

- Une table de profit sur les données d'application qui vous permet d'avoir le nombre exact de variables cibles dans chaque segment.

☑ Pour calculer la table de profit

6 Dans l'arbre *Options avancées*, situé à gauche du panneau, sélectionnez *Table de profit*.

7 Cochez la case *Calculer la table de profit sur les données d'application*.



8 Dans la liste, sélectionnez le *Nombre de quantiles* que vous souhaitez obtenir.

9 Vous pouvez ajouter des variables supplémentaires pour estimer le profit pour chaque segment de la population :

3. Dans la liste *Variables*, sélectionnez les variables que vous souhaitez ajouter à la table de profit. Utilisez la touche *CTRL* de votre clavier pour sélectionner plusieurs variables à la fois.
4. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Valeurs pour la table de profit*.

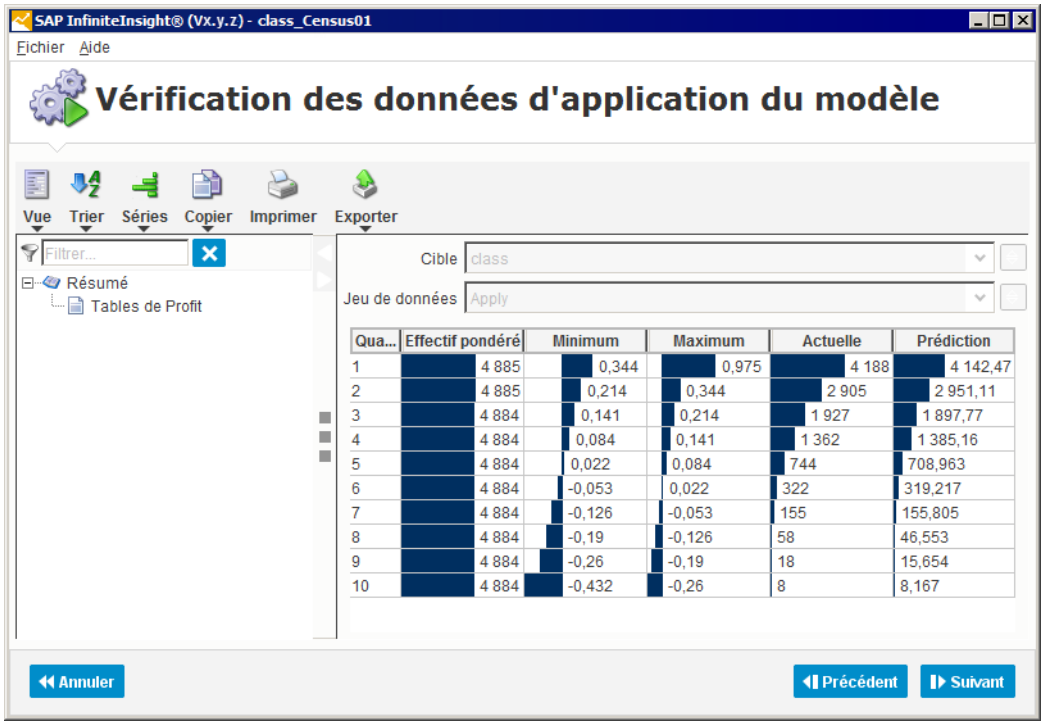
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

10 La somme de chaque variable sélectionnée sera calculée pour chaque segment de la population.

11 Cliquez sur le bouton *Valider* pour enregistrer les paramètres avancés et revenir au panneau *Appliquer un modèle*.

☑ Résultats

Le résultat du calcul de la table de profit est affiché à la fin de l'application du modèle.



Vous pouvez également le retrouver dans la section *Performance* du modèle des *Rapports de modélisation*.

Si plusieurs tables de profit ont été calculées, sélectionnez le jeu de données dans la liste proposée pour afficher la table de profit que vous souhaitez visualiser.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Sorties associées à la cible

Codes motifs

Cette fonctionnalité vous permet d'obtenir une liste des variables qui influent le plus sur une décision prise en fonction d'un score (typiquement un score de risque). Un exemple d'utilisation de ces codes motifs est de fournir à un client les raisons pour lesquelles le système de notation automatique n'a pas approuvé l'attribution d'un prêt.

☑ Pour déterminer les codes motifs

1. Dans l'arbre *Options avancées*, situé à gauche du panneau, ouvrez le noeud *Sorties associées à la cible '<Nom de la cible>'*.
2. Sélectionnez *Codes motifs*.
3. Cliquez sur le bouton + situé à droite du tableau.
4. Cliquez dans la cellule de tableau correspondant à l'option qui vous souhaitez paramétrer. Le tableau ci-dessous récapitule les options disponibles.

Options	Valeurs	Description
Nombre de codes motifs	Entier positif Valeur par défaut: *3*	Nombre de codes motifs à déterminer *Note* - Les codes motifs sont triés par ordre d'importance et seuls les plus importants sont conservés.
Seuil	▪ *Moyenne* (par défaut) ▪ Maximum ▪ Minimum	Seuil permettant de calculer les codes motifs les plus importants. Pour chaque variable la contribution correspondant au score du client est comparée à la contribution de cette variable pour l'ensemble de la population. Les codes motifs sélectionnés sont les variables dont la contribution est la plus discriminante par rapport au seuil sélectionné. Ainsi, si vous sélectionnez *Moyenne*, les contributions des variables correspondant au score du client seront comparées aux moyennes des contributions des variables de l'ensemble de la population afin de déterminer quelles variables sont les plus discriminantes.
Critère	▪ *En-dessous* (par défaut) ▪ Au-dessus	Indique si vous souhaitez générer les codes motifs quand la contribution des variables du client est inférieure ou supérieure au seuil choisi

5. Si vous souhaitez générer plusieurs types de codes motifs, répéter les étapes 3 et 4 pour chaque type.

☑ Sortie

La table fournie en sortie contient deux colonnes pour chaque code motif demandé :

- `reason_name_<critère>_<seuil>_<rang>_rr_<nom de la cible>`: contient le nom de la variable sélectionnée comme code motif.

  Par exemple, la colonne de sortie nommée `reason_name_Below_Mean_1_rr_class` contient le nom de la variable déterminée comme étant le code motif le plus important (`1`) par rapport à la variable cible `class`. Parmi les variables pour lesquelles la contribution du client est inférieure (`Below`) à la moyenne (`Mean`) des contributions de l'ensemble de la population, c'est celle qui en dévie le plus.
- `reason_value_<critère>_<seuil>_<rang>_rr_<nom de la cible>`: contient la valeur du code motif.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Cible continue

Valeur prévue

Cette option, cochée par défaut, vous permet d'ajouter au fichier de sortie la valeur prévue par le modèle pour la variable cible. Cette information apparaît dans le fichier de sortie sous le nom *rr_<variable cible>*.

Indicateur d'aberrance

Cette option vous permet d'ajouter au fichier de sortie les observations déviantes dans le fichier de sortie. Une observation est considérée comme déviante (outlier) si la différence entre sa "valeur prévue" et sa "valeur réelle" est supérieure à sa valeur de barre d'erreur. En d'autres termes, une observation est déviante quand toutes ses variables font qu'elle devrait avoir un comportement donné par rapport à la variable cible, et qu'elle a dans les faits un autre comportement.
Cette information apparaît dans le fichier de sortie sous le nom *outlier_rr_<variable cible>*. Les valeurs possibles sont 1 si l'observation est déviante et 0 sinon.

Quantiles associés à la valeur prévue

Cette option vous permet de découper le fichier de sortie, trié par ordre croissant de la cible, en quantile et d'attribuer à chaque observation le numéro du quantile dans lequel elle se trouve.

La construction des quantiles approximatifs repose sur la distribution triée et les limites des scores prédits pour l'échantillon de validation. Les limites du score sont utilisées afin de définir les quantiles approximatifs sur l'ensemble des données à appliquer.

> **Note**
>
> Le calcul exact des quantiles demanderait un tri complet des scores obtenu sur l'ensemble des données à appliquer, ce qui représente une grosse charge.
>
> L'option **Gain Chart** de la version 6.0 a pour objectif ce calcul.

Cette information apparaît dans le fichier de sortie sous le nom *quantile_rr_<variable cible>_<nombre de quantile>*, par exemple pour une variable cible nommée "class" et un nombre de quantile égal à 10 : quantile_rr_class_10.

1. Cochez l'option *Quantile associé à la valeur prévue*.
2. Saisissez le nombre de quantiles à créer dans le champs *Nombre de quantiles*.

Contributions individuelles des variables explicatives

Cette option vous permet de faire apparaître les contributions des variables explicatives de la variable cible. Vous pouvez choisir d'ajouter les contributions de toutes les variables ou bien sélectionner uniquement celles qui vous intéressent.
Cette information apparaît dans le fichier de sortie sous le nom `contrib_<variable>_rr_<variable cible>`. Ainsi si `marital-status` est une variable explicative de la cible class, la colonne du fichier sortie correspondant à la contribution de cette variable s'appellera `contrib_marital-status_rr_class`.

☑ Pour ajouter les contributions de toutes les variables

Cochez l'option *Toutes*.

☑ Pour ajouter uniquement les contributions de certaines variables

1. Cochez l'option *Sélection*.
2. Cliquez sur le bouton >> pour afficher le tableau de sélection des variables.
3. Sélectionnez dans la liste *Eléments disponibles* les variables que vous voulez ajouter (utilisez la touche *Ctrl* pour sélectionner plusieurs variables à la fois).
4. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Eléments sélectionnés*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Cible nominale

Sorties par ordre d'importance des scores

Scores

Cette option vous permet de générer dans le fichier de sortie le ou les meilleurs scores pour chaque observation. Pour chaque ligne du jeu de données d'application, SAP InfiniteInsight® compare les scores de l'observation courante obtenus pour chacune des catégories de la variable cible et affiche le meilleur score dans la colonne *best_rr_<Variable cible>_1*, puis si plusieurs scores ont été demandés par l'utilisateur il affiche le second dans la colonne *best_rr_<Variable cible>_2*, le troisième dans la colonne *best_rr_<Variable cible>_3*, et ainsi de suite... En utilisant cette option avec l'option Décision décrite ci-dessous, vous pouvez relier le meilleur score obtenu à la catégorie qui a permis l'obtention de ce score.

Décision

Cette option vous permet de générer dans le fichier de sortie la ou les meilleurs décisions pour chaque observation. Comme pour l'option précédente les scores obtenus pour chaque catégorie de la variable cible sont comparés et la catégorie ayant obtenu le meilleur score pour la ligne courante est affichée dans la colonne *decision_rr_<Variable cible>*, si plusieurs décisions ont été demandées, la catégorie ayant obtenu de second meilleur score est affichée dans la colonne *decision_rr_<Variable cible>_2*, la troisième dans la colonne *decision_rr_<Variable cible>_3*, et ainsi de suite...

Probabilités

Cette option vous permet de générer dans le fichier de sortie la probabilité des meilleurs décisions pour chaque observation. Comme pour l'option précédente, les scores obtenus pour chaque catégorie de la variable cible sont comparés et la probabilité d'apparition de la catégorie ayant obtenu le meilleur score pour la ligne courante est affichée dans la colonne *proba_decision_rr_<Variable cible>*, si plusieurs probabilités ont été demandées, la probabilité du second meilleur score est affichée dans la colonne *proba_decision_rr_<Variable cible>_2*, la troisième dans la colonne *proba_decision_rr_<Variable cible>_3*, et ainsi de suite...

Sorties par catégories de référence

Valeur prévue

Cette option vous permet de générer dans le fichier de sortie le score correspondant à chaque ligne pour les différentes catégories de la variable cible. Vous pouvez choisir d'ajouter le score pour toutes les catégories ou seulement pour certaines.
Cette information apparaît dans le fichier de sortie sous la forme *rr_<Variable cible>* pour la catégorie cible de la variable cible et *rr_<Variable cible>_<Nom de la catégorie>* pour les autres catégories de la variable cible.

☑ Pour ajouter les scores de toutes les catégories

Cochez l'option *Toutes*.

☑ Pour ajouter uniquement les scores de certaines catégories

1. Cochez l'option *Sélection*.
2. Dans la colonne *Sélection* cochez les cases correspondant aux catégories pour lesquelles vous souhaitez faire apparaître les scores dans le fichier de sortie.

Probabilité de la classe prévue

Cette option vous permet de générer dans le fichier de sortie la probabilité d'une ou plusieurs catégories de la variable cible, c'est-à-dire la probabilité
Cette information apparaît dans le fichier de sortie sous la forme *proba_rr_<Variable cible>* pour la catégorie cible de la variable cible et *proba_rr_<Variable cible>_<Nom de la catégorie>* pour les autres catégories de la variable cible.

☑ Pour ajouter les probabilités pour toutes les catégories

Cochez l'option *Toutes*.

☑ Pour ajouter uniquement les probabilités de certaines catégories

1. Cochez l'option *Sélection*.
2. Dans la colonne *Sélection*, cochez les cases correspondant aux catégories pour lesquelles vous souhaitez faire apparaître les probabilités dans le fichier de sortie.

Autres

Indicateur d'aberrance

Cette option vous permet de faire apparaître les observations déviantes dans le fichier de sortie. Une observation est considérée comme déviante (outlier) si la différence entre sa "valeur prévue" et sa "valeur réelle" est supérieure à sa valeur de barre d'erreur. En d'autres termes, une observation est déviante quand toutes ses variables font qu'elle devrait avoir un comportement donné par rapport à la variable cible, et qu'elle a dans les faits un autre comportement.
Cette information apparaît dans le fichier de sortie sous le nom *outlier_rr_<variable cible>*. Les valeurs possibles sont 1 si l'observation est déviante et 0 sinon.

Quantiles associé à la valeur prévue

Cette option vous permet de découper le fichier de sortie, trié par ordre croissant de la cible, en quantile et d'attribuer à chaque observation le numéro du quantile dans lequel elle se trouve.

La construction des quantiles approximatifs repose sur la distribution triée et les limites des scores prédits pour l'échantillon de validation. Les limites du score sont utilisées afin de définir les quantiles approximatifs sur l'ensemble des données à appliquer.

> **Note**
>
> Le calcul exact des quantiles demanderait un tri complet des scores obtenu sur l'ensemble des données à appliquer, ce qui représente une grosse charge.
>
> L'option **Gain Chart** de la version 6.0 a pour objectif ce calcul.

Cette information apparaît dans le fichier de sortie sous le nom *quantile_rr_<variable cible>_<nombre de quantile>*, par exemple pour une variable cible nommée "class" et un nombre de quantile égal à 10 : quantile_rr_class_10.

1. Cochez l'option *Quantiles associé à la valeur prévue*.
2. Saisissez le nombre de quantiles à créer dans le champs *Nombre de quantiles*.

Contributions individuelles des variables explicatives

Cette option vous permet de faire apparaître les contributions des variables explicatives de la variable cible. Vous pouvez choisir d'ajouter les contributions de toutes les variables ou bien sélectionner uniquement celles qui vous intéressent.
Cette information apparaît dans le fichier de sortie sous le nom `contrib_<variable>_rr_<variable cible>`. Ainsi si `marital-status` est une variable explicative de la cible class, la colonne du fichier sortie correspondant à la contribution de cette variable s'appellera `contrib_marital-status_rr_class`.

☑ Pour ajouter les contributions de toutes les variables

Cochez l'option *Toutes*.

☑ Pour ajouter uniquement les contributions de certaines variables

1. Cochez l'option *Sélection*.
2. Cliquez sur le bouton >> pour afficher le tableau de sélection des variables.
3. Sélectionnez dans la liste *Eléments disponibles* les variables que vous voulez ajouter (utilisez la touche *Ctrl* pour sélectionner plusieurs variables à la fois).
4. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Eléments sélectionnés*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Types de résultats proposés

L'application d'un modèle sur un jeu de données permet d'obtenir quatre types de résultats, décrit dans le tableau ci-dessous.

Type de résultat	Description
valeur prévue, ou score	Pour une variable continue, la valeur prévue Correspond à la valeur prévue par le modèle pour la variable cible de chaque observation. Les "valeurs prévues" correspondent aux valeurs présentées sur l'axe des abscisses du graphique des courbes de profit. La "valeur prévue" d'une observation est calculée en remplaçant les paramètres du polynôme représentant le modèle par les valeurs de chacune des variables de cette observation. Dans le cas d'une variable binaire, le modèle donne en sortie un score.
probabilité	Correspond à la probabilité de chaque observation d'appartenir ou non à la catégorie visée de la variable cible, c'est-à-dire la catégorie la moins fréquente sur l'ensemble des valeurs de la variable cible.
intervalle de prédiction, ou erreur maximale	L'intervalle de prédiction permet de détecter sur le jeu de données les observations ayant un comportement déviant. Une observation est considérée comme déviante (outlier) si la différence entre sa "valeur prévue" et sa "valeur réelle" est supérieure à sa valeur de l'intervalle de prédiction. En d'autres termes, une observation est déviante quand toutes ses variables font qu'elle devrait avoir un comportement donné par rapport à la variable cible, et qu'elle a dans les faits un autre comportement.
contributions individuelles	Correspondent aux contributions individuelles des variables contenues dans le jeu de données par rapport à la variable cible. La somme de toutes ces contributions individuelles correspond à la valeur prévue (score), à la constante près.
décision	L'option "décision" n'est utilisable que pour les modèles de classement, c'est-à-dire lorsque la variable cible est nominale. Elle permet de générer une décision de classement à partir des "valeurs prévues" (ou scores) générées par le modèle. Le fichier de résultat obtenu comporte une colonne dans laquelle une catégorie de la variable cible est affectée à chaque observation. La décision s'effectue en appliquant un seuil sur les "valeurs prévues" générées lors de l'application du modèle. Les observations dont la valeur prévue est supérieure au seuil défini se voient affecter la catégorie cible de la variable cible. Le seuil par défaut (calculé par lors de la phase de génération, ou d'apprentissage, du modèle) est choisi tel que l'affection de chaque catégorie de la variable cible aux observations soit représentatif de la répartition observée dans le jeu de données d'apprentissage.

En fonction du niveau d'information souhaité, vous pouvez choisir de générer différents fichiers de résultats, décrits dans le tableau ci-dessous.

En sélectionnant l'option...	Vous obtiendrez un fichier de résultats contenant pour chaque observation les informations...
valeur prévue	uniquement la valeur prévue (rr_TargetVariableName)
Probabilité	▪ la valeur prévue ▪ la probabilité (proba_rr_TargetVariableName) ▪ l'intervalle de prédiction (bar_rr_TargetVariableName)
Contributions individuelles	▪ la valeur prévue ▪ la probabilité ▪ l'intervalle de prédiction ▪ les contributions individuelles des variables (contrib_VariableName_rr_TargetVariableName)

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Decision	▪ la valeur prévue ▪ la décision (decision_rr_TargetVariableName) ▪ la probabilité de la décision (proba_decision_rr_TargetVariableName) ▪ la probabilité

Analyser les résultats de l'application

Pour ce scénario

Dans Microsoft Excel, ouvrez le fichier de résultats au format texte que vous avez obtenu suite à l'application du modèle sur le fichier *Census01.csv*.

Pour ouvrir le fichier de résultats de l'application d'un modèle

1 En fonction du format du fichier de résultats généré, utilisez *Microsoft Excel* ou toute autre application pour ouvrir ce fichier.
La figure ci-dessous présente les premières lignes et les colonnes du fichier de résultats obtenu pour le scénario.

	A	B	C	D	E	F	G
1	KxIndex	class	rr_class	proba_rr_class	bar_rr_class	contrib_age	contrib_workclass
2	1	0	0,02628554	0,177313641	1,17675817	-0,00014918	0,004575992
3	2	0	0,268730283	0,567965508	1,496288657	-0,00604592	0,004575992
4	3	0	-0,192986876	0,023777327	0,457313478	0,00038689	-0,002218456
5	4	0	-0,027590154	0,094321787	0,946476877	-0,00765412	-0,002218456
6	5	0	0,230136439	0,548656046	1,521923184	0,00574755	-0,002218456
7	6	0	0,35434714	0,737949133	1,401365995	0,00092295	-0,002218456
8	7	0	-0,358695984	0,00095057	0,467965275	-0,00550985	-0,002218456
9	8	1	0,188632727	0,435952127	1,483268023	-0,00711805	0,004575992
10	9	1	0,634126425	0,962523282	0,575068474	0,00413935	-0,002218456
11	10	1	0,540619075	0,916772068	0,933868706	-0,00175738	-0,002218456
12	11	1	0,235637605	0,551528811	1,519493461	0,00092295	-0,002218456
13	12	1	0,317775369	0,676707566	1,426785707	0,00467542	0,004575992
14	13	0	-0,192047745	0,024259994	0,463258654	0,00842789	-0,002218456
15	14	0	-0,11635045	0,058997665	0,731234848	0,00360329	-0,002218456
16	15	1	0,164729878	0,406741321	1,466872215	-0,00068525	-0,002218456

2 Vous pouvez maintenant analyser les résultats obtenus et utiliser les résultats de vos analyses pour prendre les bonnes décisions.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Description du fichier de résultats

En fonction des options que vous avez sélectionnées, le fichier de résultats contient une partie ou la totalité des informations suivantes, dans l'ordre dans lequel elles sont présentées ci-dessous :

- la **variable clé** définie lors de la description des données à l'étape de définition des paramètres de modélisation.
- éventuellement la **variable cible** renseignée par des valeurs connues si celles-ci figuraient dans le jeu de données d'application, comme c'est le cas pour ce scénario.
- la **valeur prévue** (*score*) par le modèle pour la variable cible de chaque observation. Le nom de cette colonne correspond au nom de la variable cible préfixé par *rr_*, soit pour ce scénario *rr_Class*.
- la **décision** se base sur la valeur prévue ou score. Par exemple, sa valeur peut être de 1 si l'observation est considérée comme intéressante ou de 0 si elle est considérée comme inintéressante pour le modèle. Le nom de cette colonne correspond au nom de la variable cible préfixé par *decision_rr*, soit pour ce scénario *decision_rr_class.*
- la **probabilité de la décision** se base également sur la valeur prévue ou score et donne la probabilité de la décision. Plus la probabilité est forte, plus on est sûr que la décision est bonne. Le nom de cette colonne correspond au nom de la variable cible préfixé par *proba_decision_rr_*, soit pour ce scénario *proba_decision_rr_class*.
- la **probabilité** de chaque observation d'appartenir ou non à la catégorie visée de la variable cible. Le nom de cette colonne correspond au nom de la variable cible préfixé par *proba_rr_*, soit pour ce scénario *proba_rr_Class*.
- l'**intervalle de prédiction**, ou "erreur maximale". Le nom de cette colonne correspond au nom de la variable cible préfixé par *bar_rr_*, soit pour ce scénario *bar_rr_Class*.
- les **contributions individuelles** des variables contenues dans le jeu de données par rapport à la variable cible. Les noms des colonnes des contributions individuelles correspondent aux noms de chacune des variables, préfixés par *contrib_*, soit pour ce scénario *contrib_age*, *contrib_workclass*, etc.

5.4.3 Effectuer une simulation

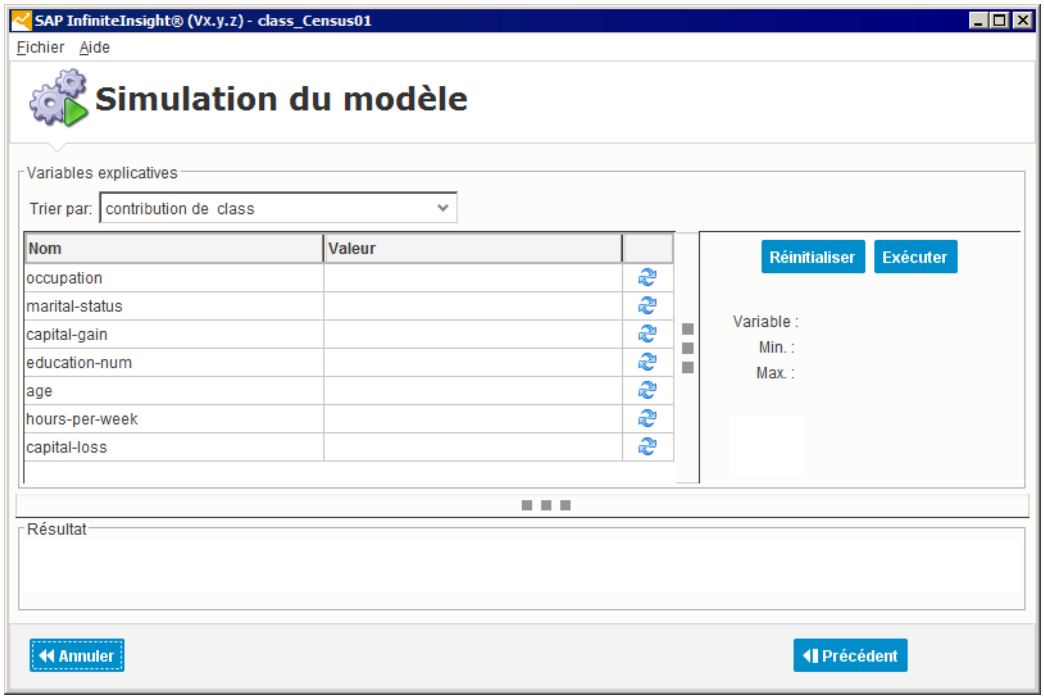
Le modèle en cours d'utilisation peut être utilisé pour effectuer des simulations sur des observations spécifiques, au cas par cas. Pour définir l'observation à analyser, vous renseignez les variables de votre choix, par exemple les variables *occupation* (profession) et *workclass* (catégorie socioprofessionnelle). Lors de l'exécution de la simulation, SAP InfiniteInsight® renseigne automatiquement certaines variables dans les valeurs sont manquantes, et essentielles au bon déroulement de la simulation.

Suite à la simulation, vous obtenez les résultats suivants :

- la **valeur prévue** (*score*),
- la **probabilité** de cette observation d'appartenir à la catégorie cible de la variable cible.

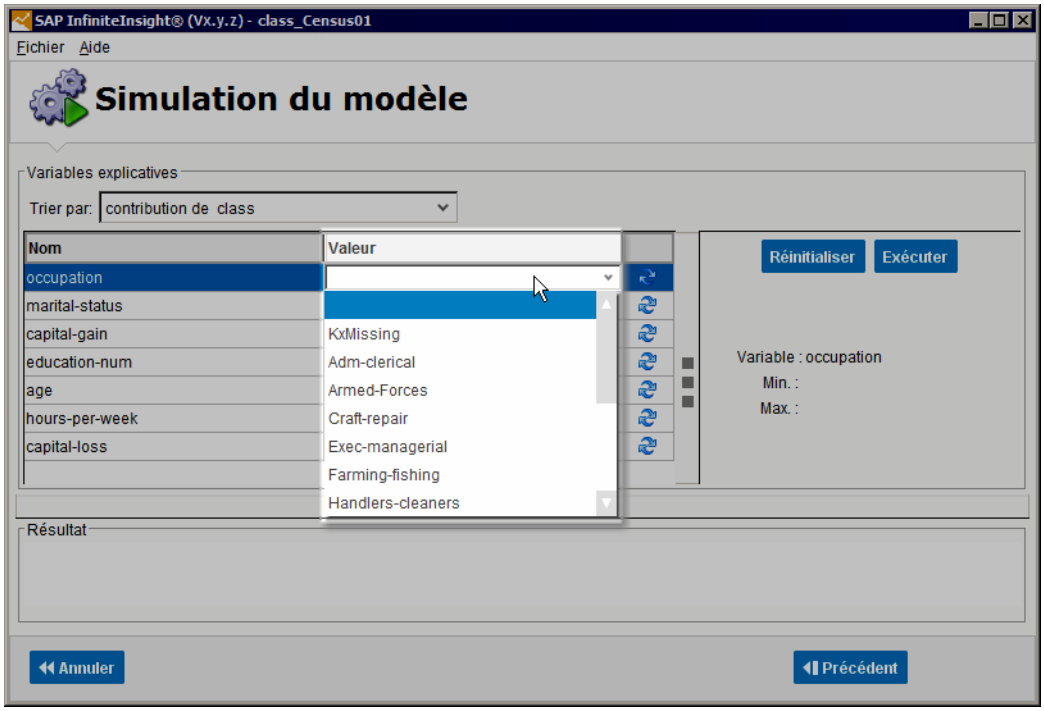
☑ Pour simuler un modèle

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Simulation*.
L'écran *Simulation du modèle* apparaît.

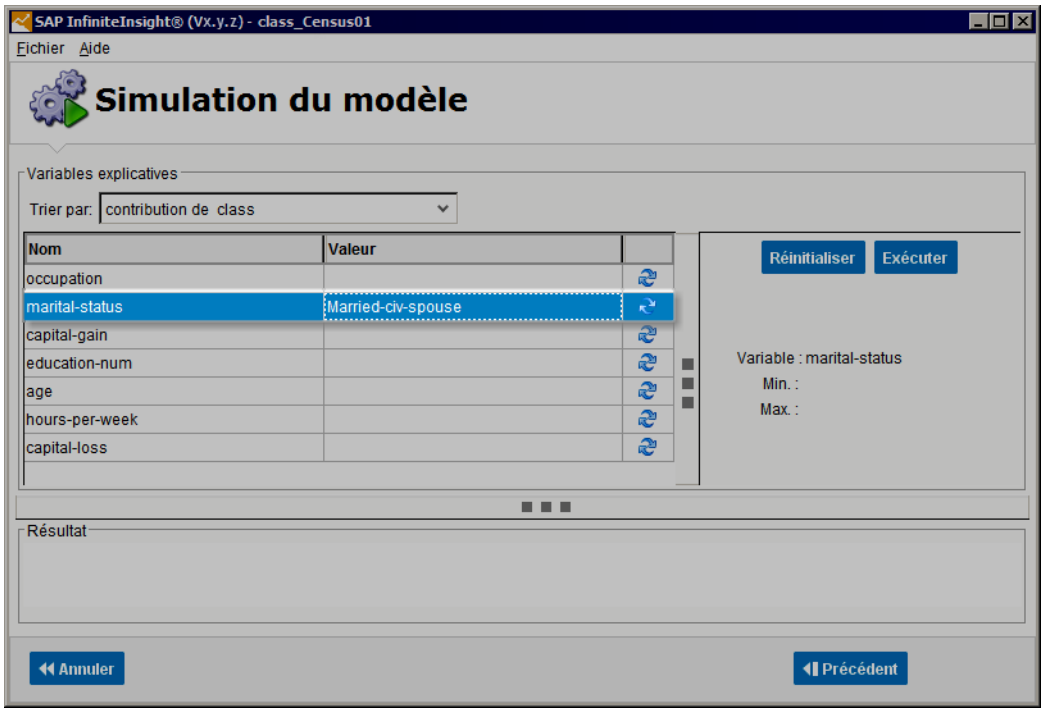


© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

2 Dans la partie de gauche (*Variables explicatives*), sélectionnez une variable, par exemple la variable *marital-status*.
Ses valeurs apparaissent dans la partie *Modification des valeurs*, dans la partie droite de l'écran.



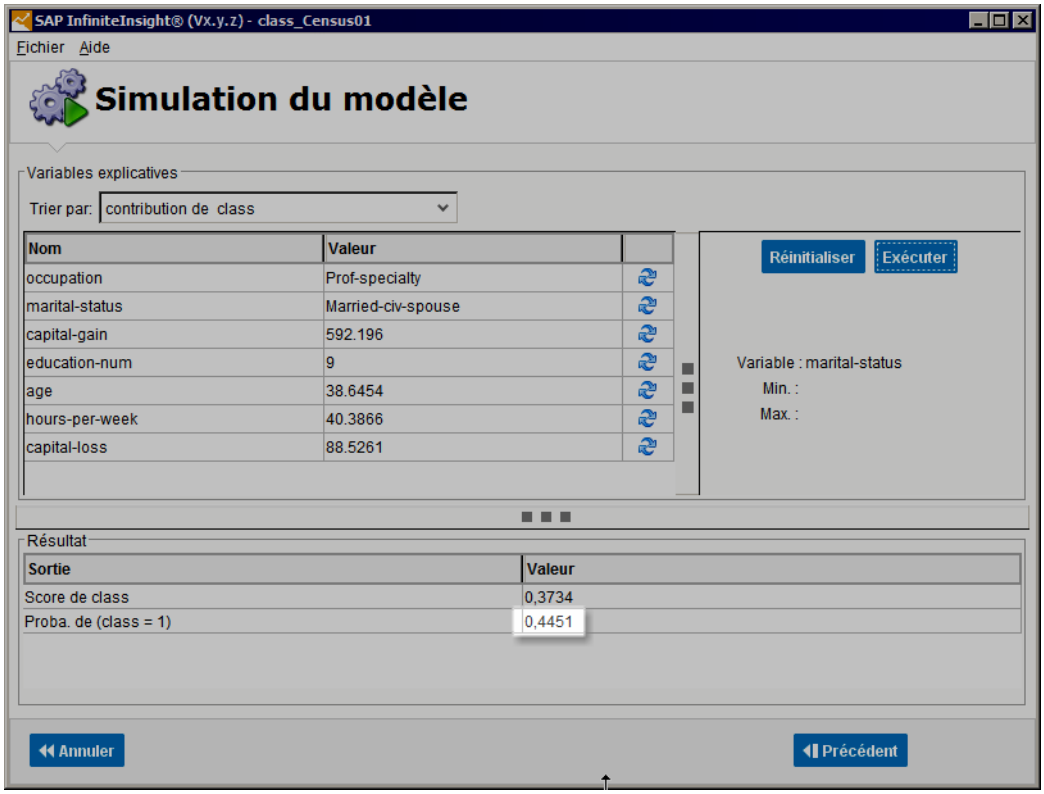
3 Dans la partie *Modification des valeurs*, dans le champ *Valeur*, sélectionnez ou entrez une valeur, par exemple *Married-civ-spouse* (marié).
La valeur apparaît dans le tableau des *Variables explicatives*, en face de la variable sélectionnée.



4 Si vous souhaitez sélectionner d'autres variables explicatives, retournez à l'étape *2*.
Sinon, passez à l'étape *5*.

5 Cliquez sur le bouton *Exécuter* pour effectuer une simulation du modèle. Les résultats de la simulation apparaissent dans la section *Résultat*. Vous obtenez la Valeur prévue (score) de l'observation décrite dans le tableau des Variables explicatives, ainsi que la probabilité de cette observation d'appartenir à la catégorie cible de la variable cible. Dans notre exemple, une seule variable (marital-status) a été initialement renseignée. La probabilité que cette observation appartienne à la catégorie cible de la variable cible est de 0,1120. Vous remarquez que certaines variables du tableau des Variables explicatives ont été automatiquement renseignées suite à l'exécution de la simulation. Le modèle complète en effet automatiquement certaines valeurs manquantes, essentielles à la simulation.
Ces valeurs sont indiquées dans le tableau ci-dessous.

Type de variable	Valeur par défaut
Variable continue	Valeur moyenne
Variable nominale	Valeur la plus fréquente
Variable ordinale	Valeur la plus fréquente



6 Vous pouvez modifier la valeur d'une variable explicative et exécuter à nouveau la simulation pour mesurer l'impact d'un tel changement par rapport à la variable cible. Par exemple :

1. Assignez à la variable *marital-status* la valeur *Widowed* (veuf) à la place de la valeur *Married-civ-spouse*.
2. Exécuter la simulation.
   La probabilité obtenue est maintenant de 0,0040.

7 Cliquez sur le bouton *Réinitialiser* pour effectuer une nouvelle simulation du modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

5.4.4 Affiner un modèle

SAP InfiniteInsight® vous permet d'affiner un modèle en cours d'utilisation. Par exemple, vous pouvez :

- essayer de **réduire le nombre de variables explicatives** utilisées pour le modèle, tout en conservant ses indicateurs de qualité KI et de robustesse KR initiaux,
- **générer un modèle de degré 2** à partir des variables les plus importantes d'un modèle de degré 1.

La *Sélection intelligente* vous permet de laisser SAP InfiniteInsight® choisir les variables ayant les plus fortes contributions selon la quantité d'information que vous souhaitez conserver.

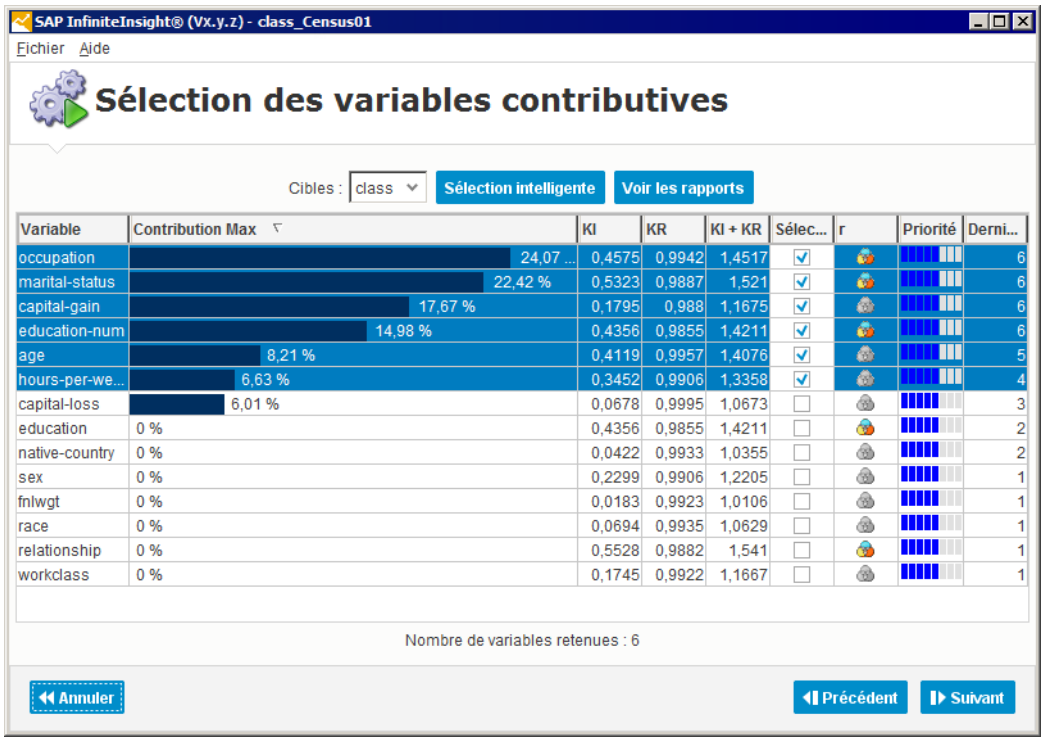
Pour chaque variable, les informations suivantes sont fournies:

- l'indice de la variable (*Index*)
- le nom de la variable (*Variable*)
- la contribution maximale de la variable (*Max Contribution*)
- le KI individuel de la variable (*KI*), qui représente la capacité de cette variable seule de prédire la variable cible.
- le KR individuel de la variable (*KR*)
- la présence de corrélations pour cette variable (*r*). Si d'autres variables sont corrélées à cette variable, l'indicateur de corrélations est allumé.

Par défaut, les variables sont triées par contributions maximales décroissantes.

☑ Pour affiner un modèle

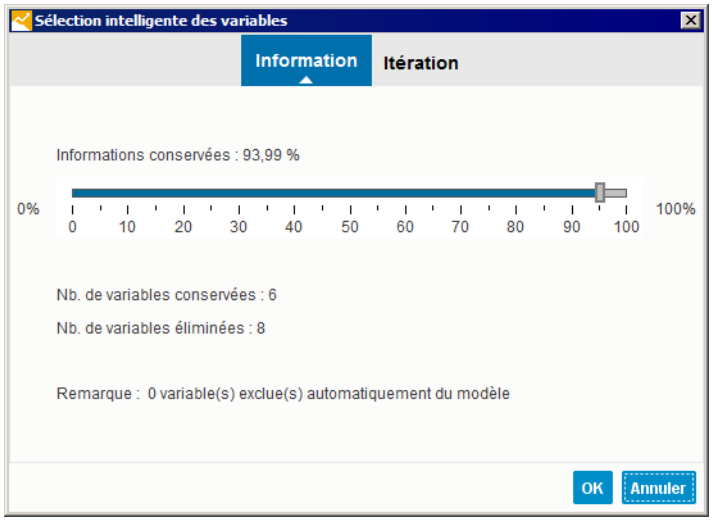
1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Sélectionner les variables à forte contribution*. L'écran *Sélection des variables contributives* apparaît.



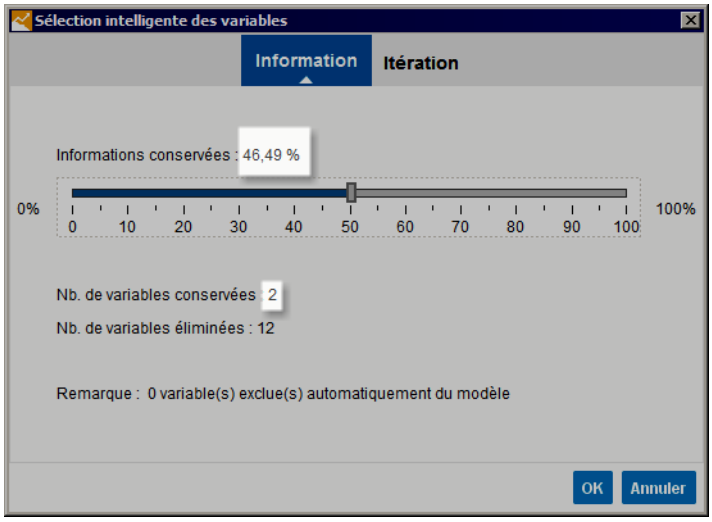
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 Dans la liste *Cibles*, sélectionnez la variable cible du modèle que vous souhaitez affiner.

3 Cliquez sur le bouton *Sélection intelligente*. La fenêtre *Sélection intelligente des variables* s'ouvre.



4 Sur la barre *Pourcentage de l'information conservée*, déplacez le curseur pour sélectionner la quantité d'information à conserver. Le nombre de variables sélectionnées est modifié en fonction de la quantité d'information.
Plus vous déplacez le curseur vers la gauche, plus vous excluez des variables. Les variables exclues sont automatiquement sélectionnées en fonction de leur importance vis à vis du modèle.
Par exemple, la figure ci-dessous montre qu'en ne conservant que **deux variables** sur les douze variables initiales, 45,2% de l'information apportée par le modèle est conservée.

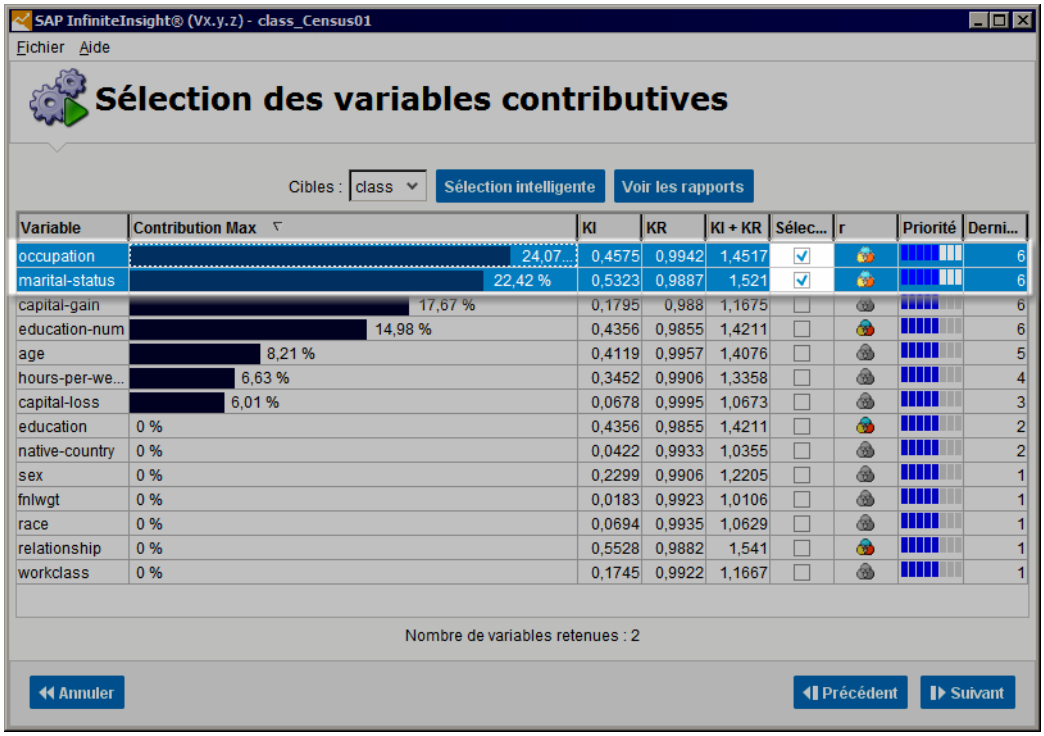


Remarque

Certaines variables du jeu de données d'apprentissage peuvent n'apporter aucune information, telles que les variables à valeur constante. Ces variables sont alors automatiquement exclues du modèle lors de la phase d'apprentissage. Le nombre de ces variables exclues est affiché sous forme de *Remarque*. Dans la figure ci-dessus, ce nombre est égal à "0".

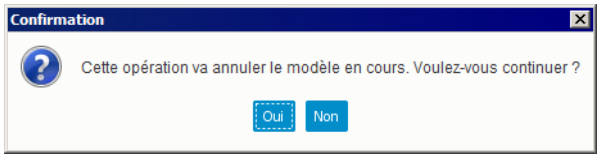
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

5 Cliquez sur le bouton *OK*.
La fenêtre se ferme et l'écran *Sélection des variables explicatives* est mis à jour avec les variables sélectionnées, vous permettant ainsi de visualiser les variables conservées et des variables exclues. Pour notre exemple, SAP InfiniteInsight® a automatiquement déterminé que les deux variables explicatives qui apportait le plus d'information pour expliquer la variable cible sont les variables *marital-status* et *capital-gain*.

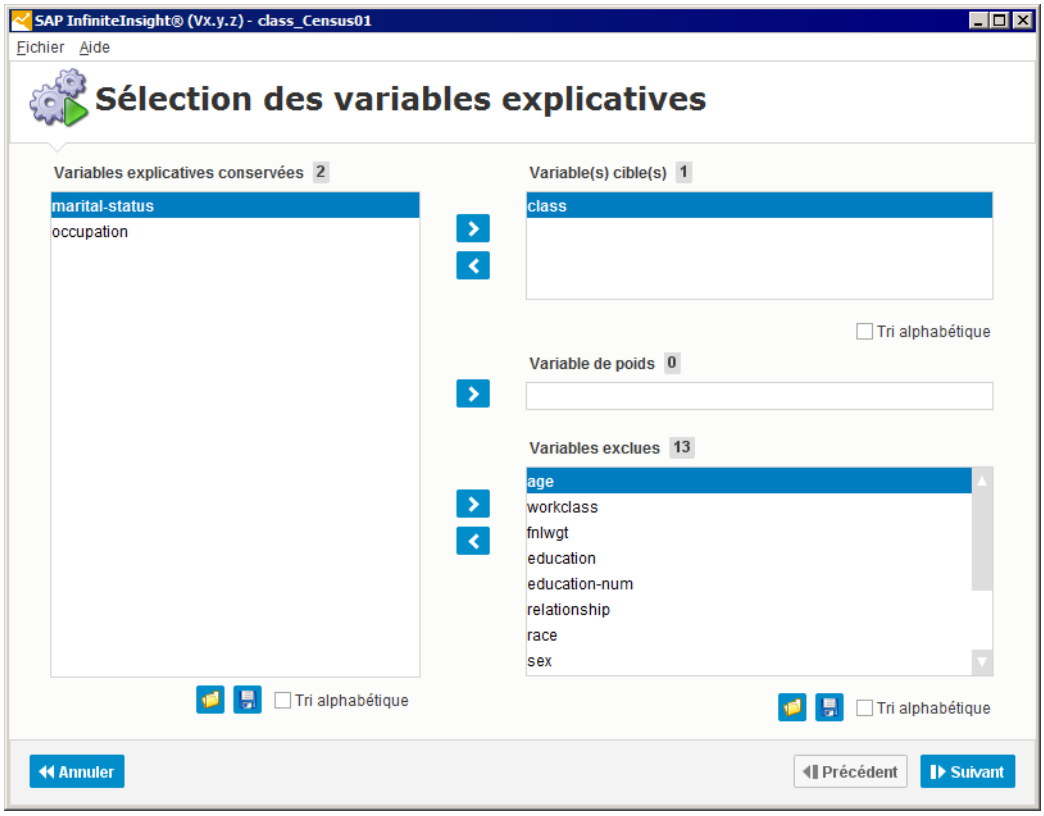


Variable	Contribution Max	KI	KR	KI + KR	Sélec...	r	Priorité	Derni...
occupation	24,07...	0,4575	0,9942	1,4517	☑			6
marital-status	22,42 %	0,5323	0,9887	1,521	☑			6
capital-gain	17,67 %	0,1795	0,988	1,1675	☐			6
education-num	14,98 %	0,4356	0,9855	1,4211	☐			6
age	8,21 %	0,4119	0,9957	1,4076	☐			5
hours-per-we...	6,63 %	0,3452	0,9906	1,3358	☐			4
capital-loss	6,01 %	0,0678	0,9995	1,0673	☐			3
education	0 %	0,4356	0,9855	1,4211	☐			2
native-country	0 %	0,0422	0,9933	1,0355	☐			2
sex	0 %	0,2299	0,9906	1,2205	☐			1
fnlwgt	0 %	0,0183	0,9923	1,0106	☐			1
race	0 %	0,0694	0,9935	1,0629	☐			1
relationship	0 %	0,5528	0,9882	1,541	☐			1
workclass	0 %	0,1745	0,9922	1,1667	☐			1

6 Cliquez sur le bouton *Suivant*. Une boîte de dialogue de confirmation apparaît.



7 Cliquez sur *Oui* pour valider la sélection des variables et réentraîner le modèle sur ces variables. L'écran *Sélection des variables explicatives* apparaît.



8 Reprenez le paramétrage du modèle à partir de l'étape de sélection des variables (voir à la page 82).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

5.4.5 Générer le code source d'un modèle

La fonctionnalité InfiniteInsight® Scorer permet d'exporter des modèles SAP InfiniteInsight® de segmentation et de regression vers différents langages de programmation. Le code ainsi généré permet d'appliquer les modèles hors de SAP InfiniteInsight®. Les codes générés permettent d'intégrer les modèles SAP InfiniteInsight® au sein d'applications ou progiciels, ou de les appliquer sur des données sans nécessiter la présence de SAP InfiniteInsight®. Ils permettent notamment d'utiliser les modèles sur des plate-formes techniques différentes de celle sur laquelle ils ont été générés.

Cette fonctionnalité nécessite l'achat d'une licence spécifique. Selon votre licence, vous pouvez générer les codes sources dans les langages suivants :

Le fichier de code généré par SAP InfiniteInsight® contiendra toute information nécessaire pour le modèle, comme l'encodage des variables, les valeurs de remplacement des valeurs manquantes, les compressions et les paramètres du modèle.

☑ Pour générer le code correspondant au modèle

1. Dans la liste *Cible à utiliser*, sélectionnez la cible du modèle.
2. Dans la section *Options de génération*, sélectionnez l'option désirée :

Option choisie	Résultats du modèle généré
Score/Estimations	le score (classement) ou l'estimation (régression)
Probabilité	le score et la probabilité (sauf pour HTML et tous les codes SQL, pour lesquels seule la probabilité est donnée)
Bar	le score et la barre d'erreur (sauf pour HTML et tous les codes SQL, pour lesquels seule la barre d'erreur est donnée)

Attention

Les options *Probabilités* et *Bar* sont disponible seulement pour les modèles InfiniteInsight® Modeler / Régression ou Classement avec cible nominale.

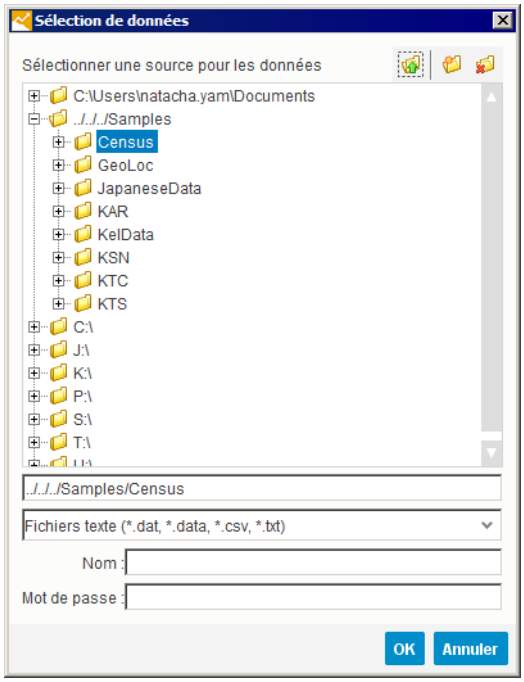
Remarque

Dans le cas d'une variable continue, le code généré comprend toujours un nombre de catégories supérieur à celui de la structure utilisateur définie ou du paramètre de *nombre de segments* si aucune structure utilisateur n'a été définie. En effet, l'encodage des variables introduit des points de continuité pour augmenter la précision de codage par rapport au jeu de données d'apprentissage. Ces points de continuité scindent certaines catégories existantes et augmentent donc le nombre de catégories dans le code généré.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Dans la liste *Choix du type de code*, sélectionnez le type de code que vous voulez générer (Liste de codes générés).

4 Dans la section *Génération*, utilisez le bouton *Parcourir* situé à droite du champ *Répertoire* pour sélectionner où le fichier sera enregistré.
Une fenêtre de sélection apparaît.

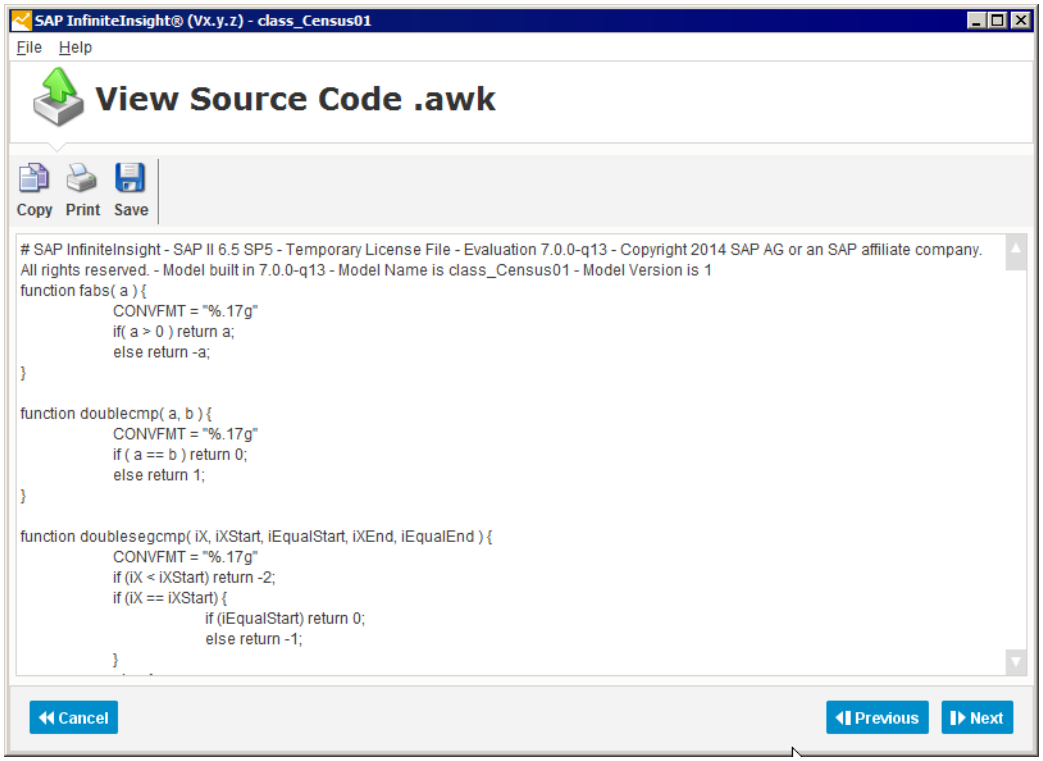


© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

5 Saisissez dans le champ *Fichier généré* le nom à donner au fichier exporté. Si vous souhaitez remplacer un fichier existant, utilisez le bouton *Parcourir* pour le sélectionner.

6 Si vous avez sélectionné l'option *Visualiser le code généré*, celui-ci s'affiche à la fin de la génération.

7 Cliquez sur le bouton *Générer*. Si le fichier existe déjà, un message de demande de confirmation s'affiche. La figure ci-dessous représente le début d'un exemple de code source C d'un modèle SAP InfiniteInsight®.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Liste des codes générés

Le tableau ci-dessous récapitule les codes proposés ainsi que leurs particularités.

Code généré	Commentaire
AWK Code	
C Code	se référer à la documentation C Code Generator (en anglais)
PMML 3.0	
PMML 3.1	
PMML 3.2	
Cpp	
DB2 UDF (SQL)	
HTML (Javascript)	contient un formulaire permettant de reproduire le modèle SAP InfiniteInsight®
JAVA Code	le fichier KxJRT.jar est nécessaire à sa compilation et son exécution
Oracle UDF (SQL)	
PMML2	
SAS Code	
SQL Code (ANSI)	
SQL Code for MySQL	
SQL Code for NEOVIEW	
SQL Code for Oracle	
SQL Code for SQLServer	entoure les nom de variables avec [ ]
SQL Code for SYBASE ASE	
SQL Code for Sybase IQ	
SQL Code for Teradata	
SQL Code for WX2	
SQLServer 2000 UDF (SQL)	
SQLTeradata	Teradata databases
SQLNetezza	Netezza databases
SQLVertica	Vertica databases
ScoreCard	seulement disponible dans InfiniteInsight® Modeler / Régression ou Classement
Teradata V2R5.1 UDF	
UDF Code for MySQL	
UDF Code for Sybase IQ	

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Code généré	Commentaire
VB Code	

> **Remarque**
>
> Lorsque vous générez du code SQL, SAS ou SQL pour MySQL, il vous sera demandé de fournir les noms de la colonne clé et du jeu de données utilisés.

Paramètres avancés

Mode UNICODE

Le *Mode Unicode* vous permet de générer le code choisi en Unicode pour qu'il puisse supporter les langues non-latines telles que le japonais, le russe, etc.

> **Note**
>
> Cette option s'applique en particulier aux codes SQL.

Options SQL/UDF

- L'option *Ne pas générer le code pour les variables non contributives* vous permet d'exclure du code toutes les variables ayant une contribution de 0 puisqu'elles n'influencent pas le résultat. Dans certains cas, ceci peut réduire d'une façon significative la taille du code généré.
- Vous pouvez soit *Utiliser le séparateur par défaut* ("GO"), soit *Utiliser un séparateur personnalisé*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

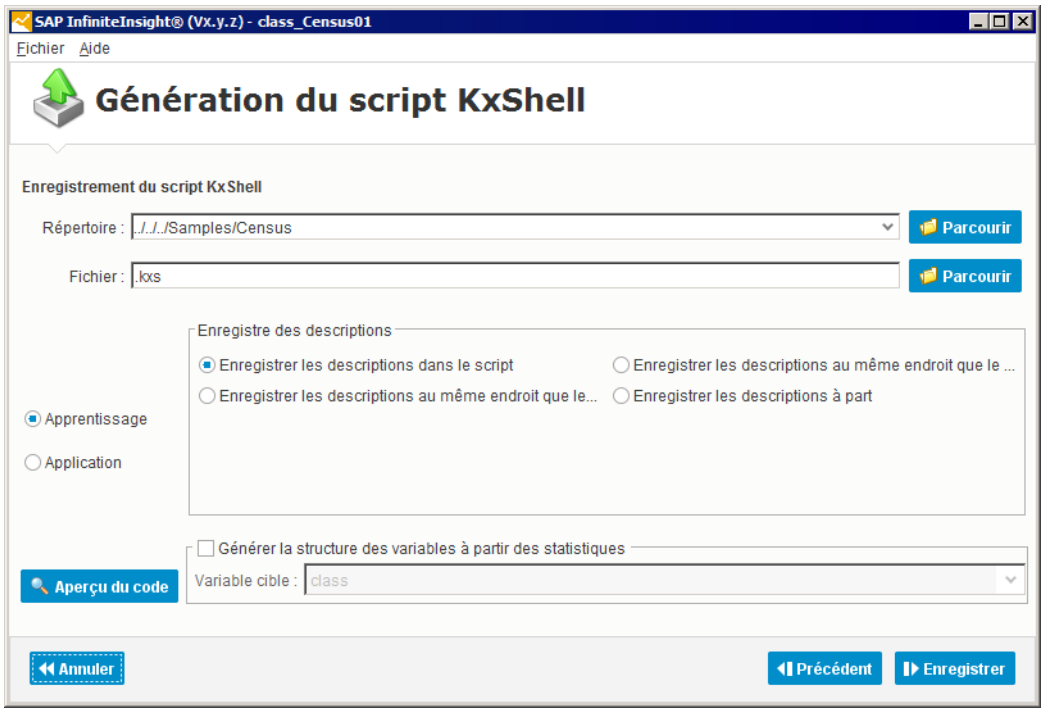
5.4.6 Exporter le script KxShell

L'export de script KxShell vous permet de générer un script reproduisant le modèle en cours. Ce script peut être ensuite utilisé pour entraîner des modèles par lots.

Lorsque vous souhaitez ajouter au script exporté des paramétrages spécifiques, tel que la sélection automatique des variables par exemple, le moyen le plus simple est d'effectuer les opérations correspondantes dans l'interface graphique avant de générer le code. Ainsi, si vous faites une sélection automatique des variables avant l'export du script shell, celui-ci contiendra le code nécessaire à cette opération.

☑ Pour enregistrer le script KxShell

1 Dans le menu *Enregistrement/Export* du panneau d'*Utilisation du modèle*, double-cliquez l'option *Exporter le script KxShell*. Le panneau *Génération de script KxShell* s'affiche.



2 Cliquez sur le bouton *Parcourir* situé à droite du champ *Répertoire* pour sélectionner le répertoire dans lequel le script sera sauvegardé.

3 Dans le champ *Fichier*, saisissez le nom du script ou s'il existe déjà, sélectionnez le avec le bouton *Parcourir*.

4 Dans le cadre *Sauvegarde des descriptions*, Sélectionnez comment vous souhaitez enregistrer la description des données de votre modèle. Les quatre options suivantes sont disponibles :

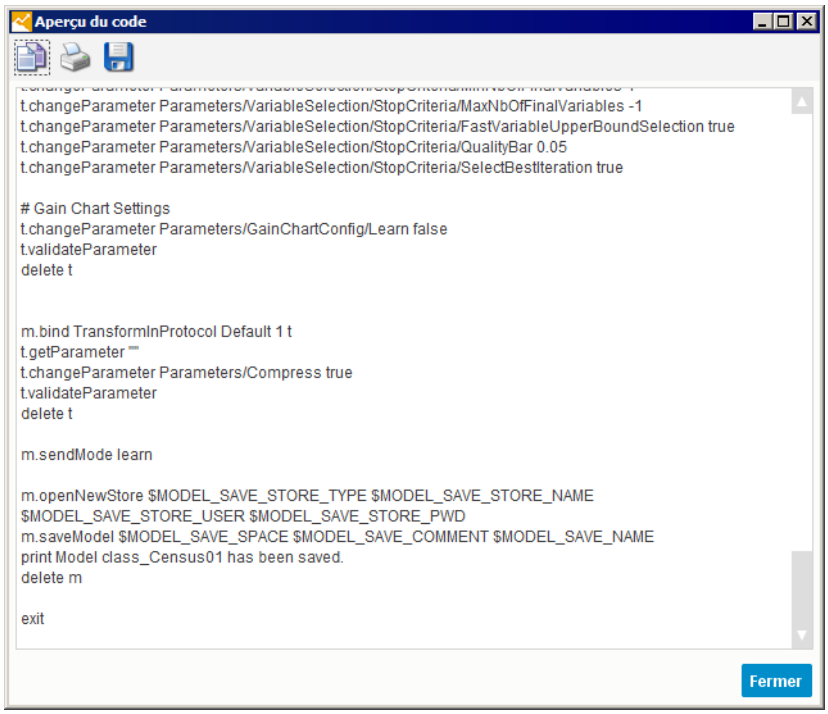
- *Sauvegarder les descriptions dans le script*
  la description des données est ajoutée dans le script KxShell. Un seul fichier est généré.
- *Sauvegarder les descriptions là où est le script*
  La description des données est enregistrée dans un nouveau fichier situé dans le même répertoire que le script KxShell.
- *Sauvegarder les descriptions là où sont les données*
  La description des données est enregistrée dans un nouveau fichier situé dans le même répertoire que les données utilisées pour créer le modèle.
- *Sauvegarder les descriptions à part*
  L'utilisateur choisit sous quel format et où sera enregistré la description des données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

> **Note**
>
> Lorsque la description est sauvegardée dans un fichier séparé, ce fichier est nommé sur le modèle suivant : `KxDesc_<Role du jeu de données>_<Nom du jeu de données>.` Par exemple, pour un jeu de données d'apprentissage nommé *Census.csv*, le nom du fichier de description sera *KxDesc_Training_Census.csv*.

5 De plus vous pouvez exporter la structure des variables qui dépend de la variable cible en sélectionnant l'option *Exporter la structure des variables dans le script*. Cette option vous permet de forcer les groupements des catégories lors de l'utilisation du modèle sur de nouveaux jeux de données.

6 Avant de générer le code, vous pouvez en voir un aperçu en cliquant sur le bouton *Aperçu du code*. Le code s'affiche dans une nouvelle fenêtre. Il peut alors être copié, imprimé ou sauvegardé.



7 Dans la fenêtre principale, cliquez sur le bouton *Suivant* pour lancer la génération du script.

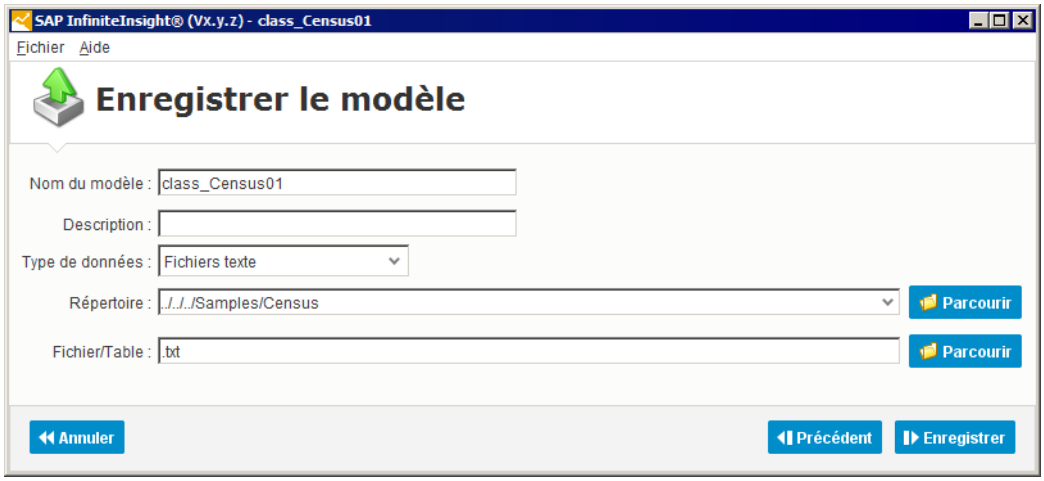
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.4.7 Enregistrer un modèle

Une fois un modèle généré, vous pouvez l'enregistrer. L'enregistrement conserve la totalité des informations qui sont relatives au modèle, c'est-à-dire ses paramètres de modélisation, ses courbes de profits, etc.

☑ Pour enregistrer un modèle

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Enregistrement*.
L'écran *Enregistrer le modèle* apparaît.



2 Renseignez les champs suivants :

- *Nom du modèle* : Ce champ vous permet d'associer un nom au modèle. Ce nom est utilisé dans la liste des modèles qui vous est proposée quand vous chargez un modèle existant.
- *Description* : Ce champ vous permet d'entrer des informations de votre choix, telles que le nom du jeu de données d'apprentissage utilisé, l'ordre du polynôme ou la capacité prédictive et la reproductibilité obtenus pour ce modèle. Ces informations peuvent vous être utiles ultérieurement pour identifier le modèle.
- *Type de données* : Cette liste vous permet de sélectionner dans quel format votre modèle sera enregistré. Les formats suivants sont proposés :

- *Fichiers texte*, pour enregistrer le modèle dans un fichier texte,
- *Bases de données*, pour enregistrer le modèle dans une table ODBC,
- *Espace de stockage mémoire*, pour enregistrer le modèle en mémoire. Le modèle sera conservé jusqu'à la fermeture de l'interface graphique de SAP InfiniteInsight®. Notez que selon votre licence d'autres formats peuvent être disponible (comme SAS, par exemple).
  - *Répertoire* : En fonction de l'option que vous avez sélectionnée, ce champ vous permet de spécifier la source ODBC ou le répertoire dans lequel vous souhaitez enregistrer le modèle .
  - *Fichier/Table* : Ce champ vous permet d'entrer le nom du fichier ou de la table qui contiendra le modèle. Le nom de fichier doit contenir l'une des deux extensions de format *.txt* (fichier texte dans lequel les données sont séparées par des tabulations) ou *.csv* (fichier texte dans lequel les données sont séparées par des virgules).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

Fichiers créés lors de l'enregistrement d'un modèle

Lorsque vous enregistrez un modèle, SAP InfiniteInsight® crée un certain nombre de fichiers à l'emplacement spécifié. Le tableau ci-dessous liste les fichiers ou tables créés lors de l'enregistrement d'un modèle et pour quel type de modèle.

Nom du fichier	Description	Utilisé par
`KxAdmin`	liste tous les modèles contenus dans le répertoire ou la base de données ainsi que leurs informations de base (date, version, nom du modèle, commentaires)	tous les modèles InfiniteInsight
`<Model_name>`	fichier nommé d'après le modèle et contenant toutes les données à l'exception des informations des graphiques. Ces dernières sont stockées dans des tables ou fichiers supplémentaires (voir ci-dessous)	tous les modèles InfiniteInsight
`KxInfos`	indique quelles tables additionnelles sont utilisées par le modèle	tous les modèles InfiniteInsight
`KxOlapCube`	contient les informations du cube OLAP utilisé par l'arbre de décision, lorsque l'option Arbre de décision a été activée	les modèles de régression ou de classement avec arbre de décision
`KxLinks`	contient les liens des graphiques du modèle	les modèles de réseaux sociaux uniquement
`KxNodes`	liste l'ensemble des noeuds de tous les graphiques et leurs attributs	les modèles de réseaux sociaux uniquement
`KxCommunities`	contient les correspondances entre les noeuds et leur communauté lorsque la détection des communautés a été activée	les modèles de réseaux sociaux uniquement

Attention

Lorsque vous partagez ou envoyez un modèle, *tous les fichiers créés lors de la sauvegarde du modèle doivent être joints*, sinon le destinataire ne pourra pas ouvrir le modèle.

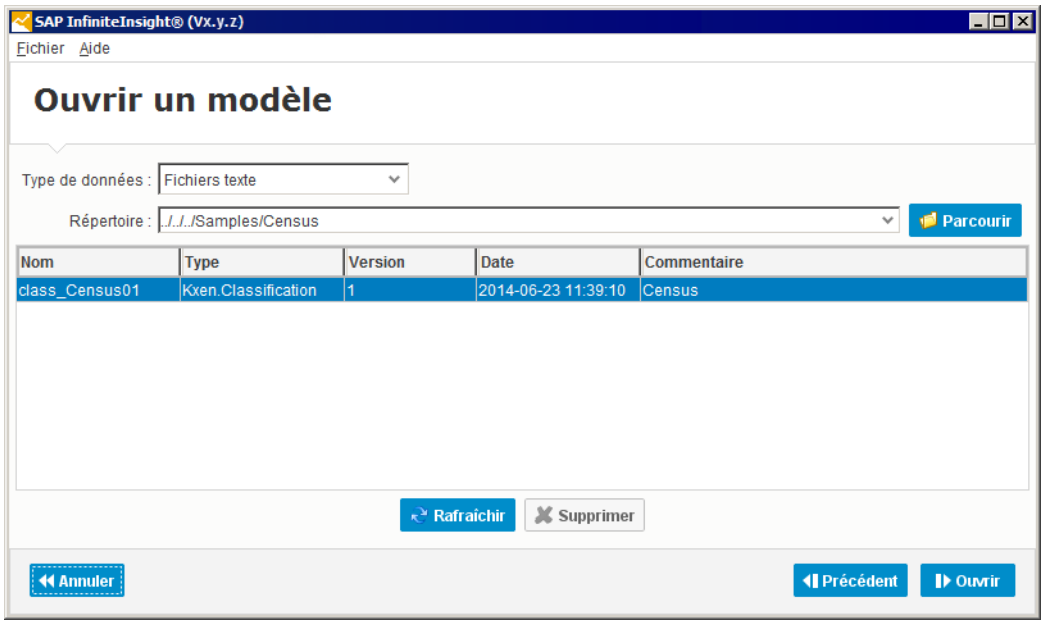
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

5.4.8 Ouvrir un modèle existant

Une fois enregistrés, les modèles peuvent être ouverts et réutilisés dans SAP InfiniteInsight®.

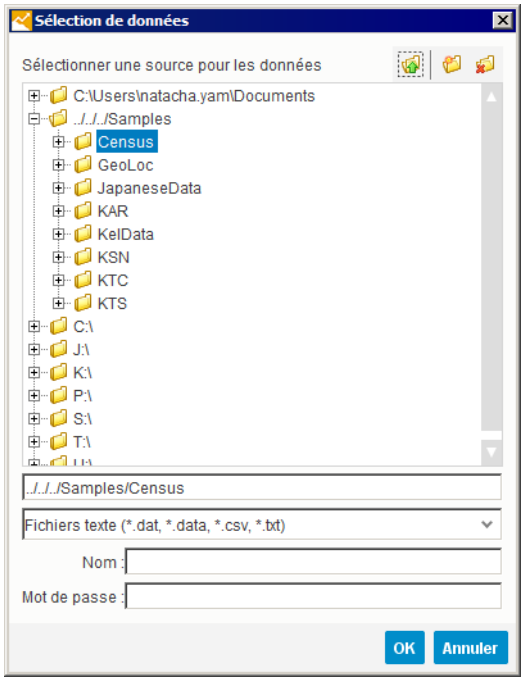
☑ Pour ouvrir un modèle

1 Sur la page d'accueil de l'assistant de modélisation, sélectionnez *Ouvrir un modèle*, puis cliquez sur le bouton *Suite*.
L'écran *Ouvrir un modèle* apparaît.



2 Dans la liste *Type de données*, sélectionnez le format du modèle que vous souhaitez ouvrir.

3 Cliquez sur le bouton *Parcourir*.
Une fenêtre de sélection apparaît.



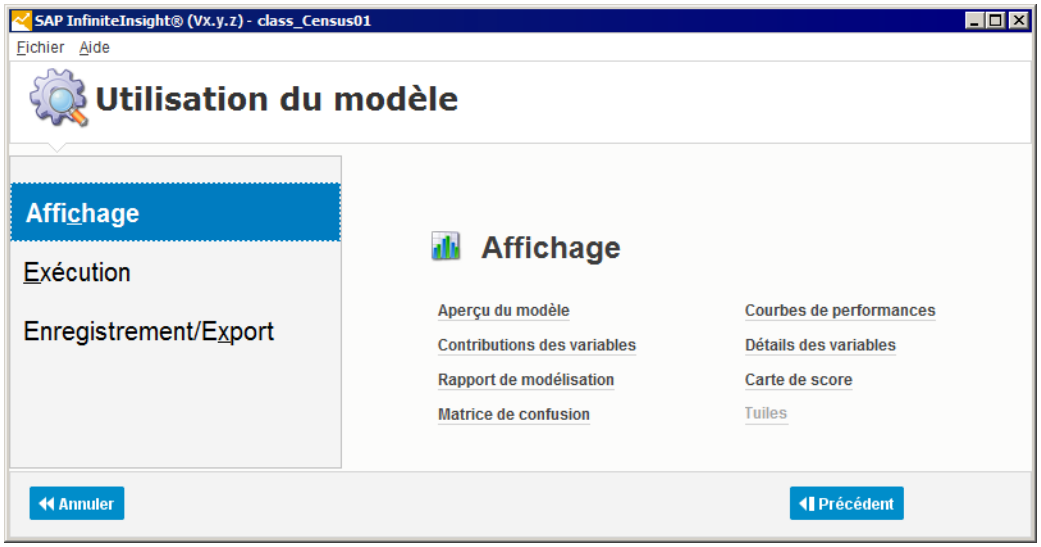
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de classement ou de régression avec InfiniteInsight® Mo

4 Sélectionnez le répertoire dans lequel est stocké le modèle que vous souhaitez ouvrir.
La liste des modèles stockés dans ce répertoire apparaît. Le tableau ci-dessous décrit les informations fournies pour chaque modèle et permettant d'identifier plus facilement le modèle recherché.

Colonne	Description	Valeurs
Nom	Nom sous lequel le modèle a été enregistré	Chaîne de caractères
Type	Type du modèle	▪ `Kxen.Classification` : classement (cible nominale) ▪ `Kxen.Regression` : régression (cible continue) ▪ `Kxen.Segmentation` : segmentation ou regroupement en mode SQL ▪ `Kxen.Clustering` : segmentation sans mode SQL ▪ `Kxen.TimeSeries` : séries temporelles ▪ `Kxen.AssociationRules` : règles d'association ▪ `Kxen.Social` : réseaux sociaux ▪ `Kxen.SimpleModel` : modèles multi-cibles, regroupement sans mode SQL et tous les autres types de modèles
Version	Numéro de version du modèle lorsque celui-ci a été sauvegardé plusieurs fois	Entier commençant à 1
Date	Date de sauvegarde du modèle	Date et heure au format `aaaa-mm-jj hh:mm:ss`
Commentaire	Commentaire facultatif saisi par l'utilisateur pour faciliter l'identification du modèle	Chaîne de caractères

5 Sélectionnez un modèle dans la liste.

6 Cliquez sur le bouton *Ouvrir*.
Le menu d'utilisation du modèle apparaît.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

6 Scénario d'utilisation : Personnalisez votre communication grâce à la modélisation de données

Pour un simple résumé du scénario d'utilisation de InfiniteInsight® Modeler / Segmentation, voir Scénario 2 : InfiniteInsight® Modeler / Segmentation à la page 186.

6.1 Présentation

Ce scénario constitue la suite logique du scénario 1.

Lors du scénario 1, grâce à InfiniteInsight® Modeler / Régression ou Classement de SAP InfiniteInsight®, vous avez atteint tous les objectifs de votre première campagne marketing, en respectant les délais et le budget qui vous étaient impartis.

Afin de **personnaliser les messages marketing de la banque et d'améliorer la communication** avec les différents clients et de prospects de ce nouveau produit, la Direction Générale vous demande maintenant d'**établir une segmentation précise des clients de ce produit**.

Grâce à InfiniteInsight® Modeler / Segmentation, vous construisez un modèle descriptif dans les meilleurs délais et à moindre coût. Ce modèle vous permet de connaître les **profils caractéristiques** des clients qui sont intéressés par votre nouveau produit, et ainsi, de répondre à votre problématique et de remplir vos objectifs.

6.2 Votre objectif

Imaginons le cas suivant.

Grâce à la fonctionnalité de régression / classement de SAP InfiniteInsight®, vous avez atteint tous les objectifs de votre dernière campagne marketing, en respectant les délais et le budget qui vous étaient impartis (voir scénario 1 à la page 54).

Pour **améliorer le taux de retour de votre campagne**, la Direction Générale vous demande :

- d'**établir une segmentation de votre clientèle**,
- d'**analyser les caractéristiques des segments identifiés**,
- de **définir une communication adaptée à chaque segment**.

La segmentation doit vous permettre en particulier de distinguer les segments de clients **en fonction de leur propension à acheter le nouveau produit** d'épargne haut de gamme proposé par votre entreprise. Vous pouvez ainsi comprendre au mieux le profil de vos clients.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Scénario d'utilisation : Personnalisez votre communication grâce à la mo

6.3 Votre approche

Pour des raisons pratiques d'organisation, vous souhaiter définir **cinq groupes de clients**, ou segments, et décrire les profils des clients appartenant à chacun de ces groupes.

Pour ce projet, vous utilisez le jeu de données que constitue l'échantillon des 50 000 personnes ayant répondu à votre premier test, lors de la campagne précédente.

Ce fichier correspond au fichier exemple *Census01.csv*, livré avec SAP InfiniteInsight® et décrit dans la section Présentation du fichier exemple (voir à la page 63).

6.4 Votre problématique

Dans votre base de données marketing, vous possédez :

- une liste de 1 000 000 prospects,
- une liste de 50 000 prospects (personnes sélectionnées lors de la phase de test de votre campagne), dont vous connaissez la réponse à la campagne. Cet échantillon constitue donc un jeu de données d'apprentissage. Cet échantillon, issu de la base de données globale, comporte également des valeurs manquantes.

Votre problématique consiste donc à :

- créer rapidement une segmentation sur le jeu de données d'apprentissage que constitue l'échantillon, utilisé en l'état. Les segments obtenus vous permettront de mieux comprendre le profil des individus de votre base de données en fonction de leur propension à acheter.
- appliquer ensuite le modèle de segmentation obtenu sur la totalité de votre base de données, pour déterminer à quel segment appartient chaque individu référencé dans cette base de données.

6.5 Vos solutions

Pour sélectionner les individus à qui envoyer un courrier, vous avez plusieurs solutions. Vous pouvez utiliser :

- une **méthode intuitive**,
- une **méthode statistique classique** (nuées dynamiques, K-means, segmentations hiérarchiques ascendantes et descendantes),
- la **méthode InfiniteInsight**.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

6.5.1 Méthode intuitive

Cette méthode consiste à utiliser la connaissance que vous avez des différents profils de vos clients. Grâce à la connaissance métier que vous avez de votre clientèle, vous déterminez vous-même quels sont les critères de segmentation déterminants et créez ainsi les segments.

Le principal inconvénient de cette méthode est que le nombre d'informations disponibles pour chaque client référencé dans votre base de données croît avec le temps. Au fur et à mesure de l'enrichissement de votre base de données, il vous est donc de plus en plus difficile de créer des segments qui prennent en compte toutes les données disponibles et répondent en même temps à votre problématique. De plus, alors que ce volume d'informations croissant vous impose de créer des segmentations de plus en plus fréquemment, le temps nécessaire à la création de ces segmentations devient de plus en plus important.

Enfin, votre hiérarchie souhaite que vous utilisiez une méthode rationnelle, et ne reposant pas simplement sur votre intuition, pour effectuer vos segmentations.

6.5.2 Méthode statistique classique

Sur la base des informations que vous possédez, des experts en Data Mining peuvent construire une segmentation. En d'autres mots, vous allez demander à l'un de vos expert statisticien de créer un modèle mathématique qui permette de créer des segments basés sur les profils de vos clients.

Afin de mettre en place cette méthode le statisticien doit :

- analyser en détails votre base de données.
- préparer minutieusement votre base de données, notamment en encodant les variables en fonction de leur type (nominal, ordinal ou continue) de manière à ce qu'ils soient exploitables par les outils d'analyse à utiliser. La stratégie d'encodage utilisée déterminera la nature de la segmentation obtenue. A cette étape, le statisticien oriente donc de manière plus ou moins consciente les résultats.
- tester différents types d'algorithmes (nuées dynamiques, K-means, segmentations hiérarchiques ascendantes et descendantes) et sélectionner le plus adapté à votre problématique.
- évaluer la pertinence des segments obtenus, notamment en fonction de votre problématique métier.

Après quelques semaines, l'expert statisticien est en mesure de fournir un certain nombre de segments, ou groupes homogènes, dans lesquels sont assignés chacun des individus de votre base.

Cette méthode présente des contraintes importantes. Vous devez :

- vous assurer que l'expert statisticien, externe au Département Marketing, est disponible selon le planning fixé,
- vous assurer que le montant de ses honoraires entre bien dans votre budget,
- passer du temps à lui expliquer votre problématique métier,
- passer du temps à comprendre les résultats qu'il vous fournit,
- demander à un programmeur d'écrire un programme permettant de déterminer à quel segment appartient tout nouvel individu ajouté à votre base de données.

De plus, cette méthode n'est pas systématique. En effet, deux statisticiens réalisant cette segmentation, sur le même jeu de données, obtiendront des résultats différents.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Scénario d'utilisation : Personnalisez votre communication grâce à la mo

6.5.3 Méthode InfiniteInsight

InfiniteInsight® Modeler / Segmentation vous permet de générer en quelques minutes un modèle de segmentation de vos clients, en prenant en compte l'intérêt de vos clients pour votre nouveau produit.

InfiniteInsight® Modeler / Segmentation détecte automatiquement les interactions entre les variables de votre jeu de données de manière a construire des sous-jeux de données homogènes, ou segments. Chaque segment est homogène vis-à-vis de l'ensemble des variables, et particulièrement vis-à-vis de la variable cible "a répondu favorablement à mon test".

Vous découvrez ainsi les caractéristiques des différents segments, c'est-à-dire des segments qui ont un fort taux de réponse et de ceux qui ont un mauvais taux de réponse. De plus, si votre base de données clients contient les dépenses de vos clients sur vos autres produits, vous obtenez en même temps les synergies de ventes de produits par segment.

Grâce à InfiniteInsight® Modeler / Segmentation, vous possédez tous les éléments d'analyse pour définir le type de message à envoyer à chaque segment de clients. Vous disposez de segments homogènes et vous permettant de répondre à votre problématique. Surtout, cette segmentation est systématique : les résultats obtenus ne représentent pas une vue particulière de vos données mais sont robustes. En d'autres mots, deux personnes réalisant cette segmentation avec la méthode InfiniteInsight obtiendront les mêmes résultats.

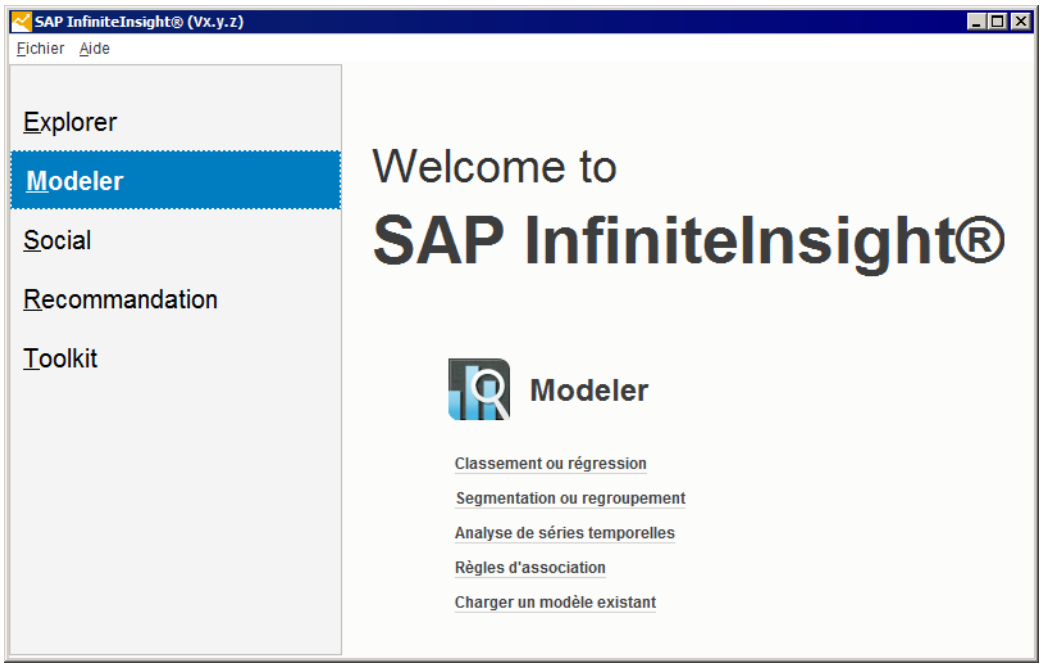
6.6 L'assistant de modélisation

Pour réaliser les deux scénarios, vous utilisez l'assistant de modélisation SAP InfiniteInsight®. Cet assistant vous permet de sélectionner la fonctionnalité avec laquelle vous souhaitez travailler, et vous assiste dans toutes les étapes de la modélisation.

Pour voir plus d'informations sur les fonctionnalités de InfiniteInsight® Modeler, voir la section Architecture et fonctionnement à la page 11.

☑ Pour démarrer l'assistant de modélisation

1 Sélectionnez *Démarrer* > *Programmes* > *SAP Business Intelligence* > *SAP SAP InfiniteInsight®* > *SAP InfiniteInsight®*
L'assistant de modélisation apparaît.



2 Cliquez sur l'action que vous souhaitez réaliser (création de modèle, exploration de données, préparation de données...).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Scénario d'utilisation : Personnalisez votre communication grâce à la mo

6.6.1 Editer les options

☑ Editer les options de l'assistant de modélisation

1 Dans le menu *Fichier*, cliquez sur *Préférences...*.
Une fenêtre *Editer les options...* s'ouvre.

Les options suivantes peuvent être modifiées :

Catégorie	Options
Général	▪ Pays ▪ Langage ▪ Niveau de message ▪ Taille maximum du fichier log ▪ Niveau de message pour les valeurs aberrantes ▪ Afficher l'arbre des paramètres ▪ Taille de l'historique des répertoires ▪ Toujours quitter sans confirmer ▪ Inclure test dans la stratégie de découpage par défaut
Emplacements	▪ Emplacement par défaut pour les données d'application en entrée ▪ Emplacement par défaut pour les données d'application en sortie ▪ Emplacement par défaut pour l'enregistrement des modèles
Entrepôt de métadonnées	▪ Activer un espace de stockage unique pour les métadonnées ▪ Editer le contenu de la bibliothèque de variables
Graphique	▪ Nombre de points de la courbe de performance ▪ Nombre de barres affichées ▪ Désactiver le Look and feel SAP InfiniteInsight® ▪ Afficher les diagrammes en 3D ▪ Désactiver le double tampon ▪ Optimiser pour les affichages distants ▪ Se souvenir de la position et de la taille en quittant
Rapport	▪ Nombre de variables intéressantes ▪ Feuille de style active ▪ Personnalisez vos feuilles style
Géolocalisation	▪ Protocol du système d'information géographique

Personnaliser les feuilles de style

SAP InfiniteInsight® vous offre la possibilité de personnaliser les rapports. La feuille de style par défaut, appelée *Feuille de style SAP InfiniteInsight® (par défaut)*, ne peut être modifiée. Vous devez créer vos propres feuilles de styles pour changer la configuration.

Note

Pour créer, charger et enregistrer une feuille de style, vous devez préciser le répertoire des feuilles de style dans le panneau *Editer les options...* avant d'ouvrir la fenêtre *Editeur de feuilles de style SAP InfiniteInsight®*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Créer une nouvelle feuille de style

1. Dans le champ *Répertoire*, cliquez sur le bouton (*Parcourir*).
2. Sélectionnez un dossier qui contiendra vos feuilles de style.
3. Cliquez sur le bouton (*Ajouter*).
   Une nouvelle feuille de style a été créée.
4. Cliquez sur le bouton  .
   La fenêtre *Editeur de feuilles de style* s'ouvre.
5. Dans le champ *Nom de la feuille de style*, entrez un nom pour la nouvelle feuille de style.
   L'extension `.krs` est automatiquement ajoutée.

> **Note**
>
> Vous pouvez dupliquer une feuille de style en modifiant le nom de votre feuille. La feuille de style précédente n'est pas supprimée.

☑ Supprimer une feuille de style

1. Sélectionnez une des feuilles de styles proposées.
2. Cliquez sur le bouton (*Retirer*).

> **Note**
>
> La feuille de style n'est pas seulement supprimée de la liste, mais également du répertoire.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Scénario d'utilisation : Personnalisez votre communication grâce à la mo

☑ Modifier la configuration générale

Configuration...	Options...	Note...
Couleur de fond	▪ choisir la couleur ▪ rendre transparent	Uniquement les formats PDF et HTML peuvent afficher une couleur de fond.
Editer la configuration	▪ taille des polices ▪ style ▪ couleurs de fond ▪ configuration de tableaux	Cochez l'option *Rendre dynamiquement les changements* ou cliquez sur *Appliquer* pour visualiser les modifications.

Les options sélectionnées s'appliquent à l'assistant de modélisation et aux rapports générés.

☑ Modifier les paramètres des graphiques

Configuration...	Options...	Note...
Couleurs des graphiques	▪ modifier les couleurs	
Histogrammes	▪ horizontal ▪ vertical	Il est possible de choisir une orientation différente que celle définie par défaut pour une section spécifique.

☑ Modifier des sections de rapport

1. Sélectionnez les propriétés de votre choix.
2. Cliquez sur *Enregistrer*.
   Une fenêtre s'ouvre, indiquant que votre feuille de style a bien été sauvegardée.
3. Cliquez sur *OK*.

Configuration...	Options...	Note...
Type de vue	choisissez entre tabulaire, HTML et graphique. La dernière option n'est disponible que si la section peut être affichée comme graphique.	
Type de graphique	choisissez un des types proposés.	Cette option n'est disponible que pour les sections du type *graphique*.
Basculer l'orientation	cette option vous permet de choisir une orientation différente que celle définie par défaut pour une section de rapport	
Trier	vous pouvez choisir la colonne à utiliser pour le tri et l'ordre de tri	
Visibilité	vous pouvez cacher une colonne d'une section ou même toute une section de rapport	Au moins une colonne d'une section doit rester visible.

☑ Appliquer la nouvelle feuille de style aux rapports générés

1. Dans la fenêtre *Rapport*, sélectionnez la feuille de style que vous souhaitez appliquer à vos rapports.
2. Cliquez sur *OK*.
   Une fenêtre s'ouvre, indiquant que vous devez redémarrer l'assistant de modélisation pour prendre en compte les options modifiées.
3. Cliquez sur *OK*.
   Lorsque vous exécutez un modèle, tous les rapports générés (rapport de modélisation, rapport excel et rapport statistique) sont personnalisés selon votre feuille de style.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

7 Créer un modèle de segmentation ou de regroupement avec InfiniteInsight® Modeler

La modélisation de données avec InfiniteInsight® Modeler / Segmentation se subdivise en quatre grandes étapes:

- Etape 1 - **Définition** des paramètres de modélisation
- Etape 2 - **Génération et validation** du modèle
- Etape 3 - **Analyse et compréhension** des résultats d'analyse
- Etape 4 - **Utilisation** du modèle généré

7.1 Etape 1 - Définir les paramètres de modélisation

Pour répondre à votre problématique, vous cherchez à :

- **décomposer** l'échantillon des 50000 prospects ayant répondu à la phase de test de votre campagne marketing (voir Scénario 1 (voir "Scénario d'utilisation : Gagnez en efficacité et maîtrisez votre budget grâce à la modélisation" à la page 54)) en groupes homogènes.
- **décrire** chacun de ces groupes et assurer une communication personnalisée vers ces différentes cibles.

InfiniteInsight® Modeler / Segmentation vous permet de créer des modèles descriptifs.

La première étape du processus de modélisation consiste à définir les paramètres de modélisation, c'est-à-dire à :

1. Sélectionner une **source de données** à utiliser comme jeu de données d'apprentissage.
2. **Décrire le jeu de données** sélectionné.
3. **Sélectionner les variables**.
4. Vérifier les **paramètres du modèle**.
5. Définir le **nombre de segments**.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

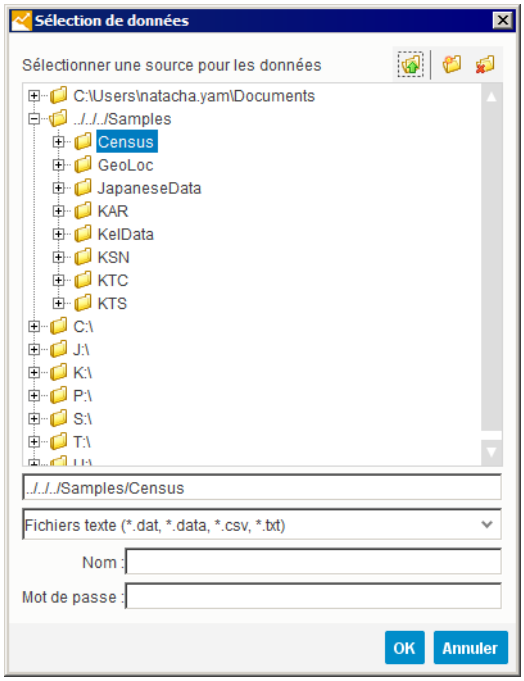
7.1.1 Sélectionner une source de données

Pour sélectionner une source de données

☑ Pour sélectionner une source de données

1 Dans l'écran *Données à modéliser*, sélectionnez l'option *Fichiers texte* pour sélectionner le format de la source de données à utiliser.

2 Cliquez sur le bouton *Parcourir*.
La fenêtre de sélection suivante apparaît.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Double-cliquez sur le répertoire *Samples*, puis sur le répertoire *Census*.

4 Sélectionnez le fichier *Census01.csv*, puis cliquez sur *OK*.
Le nom du fichier apparaît dans le champ *Estimation*.

5 Cliquez sur le bouton *Suivant*.
L'écran *Description des données* apparaît.

6 Passez à la section **Décrire les données**.

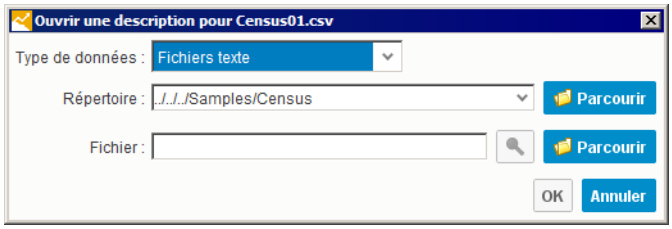
7.1.2 Décrire les données sélectionnées

Pour ce scénario

- Sélectionnez *Fichiers texte* comme type de source de données.
- Utilisez le fichier de description existant `desc_Census01.csv`, correspondant au fichier de données `Census01.csv`.

Pour utiliser un fichier de description existant

1 Dans l'écran *Description des données*, cliquez sur le bouton *Ouvrir*. La fenêtre *Ouvrir une description* s'affiche.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

2 Sélectionnez le type de votre source de données dans la liste en haut à droite.

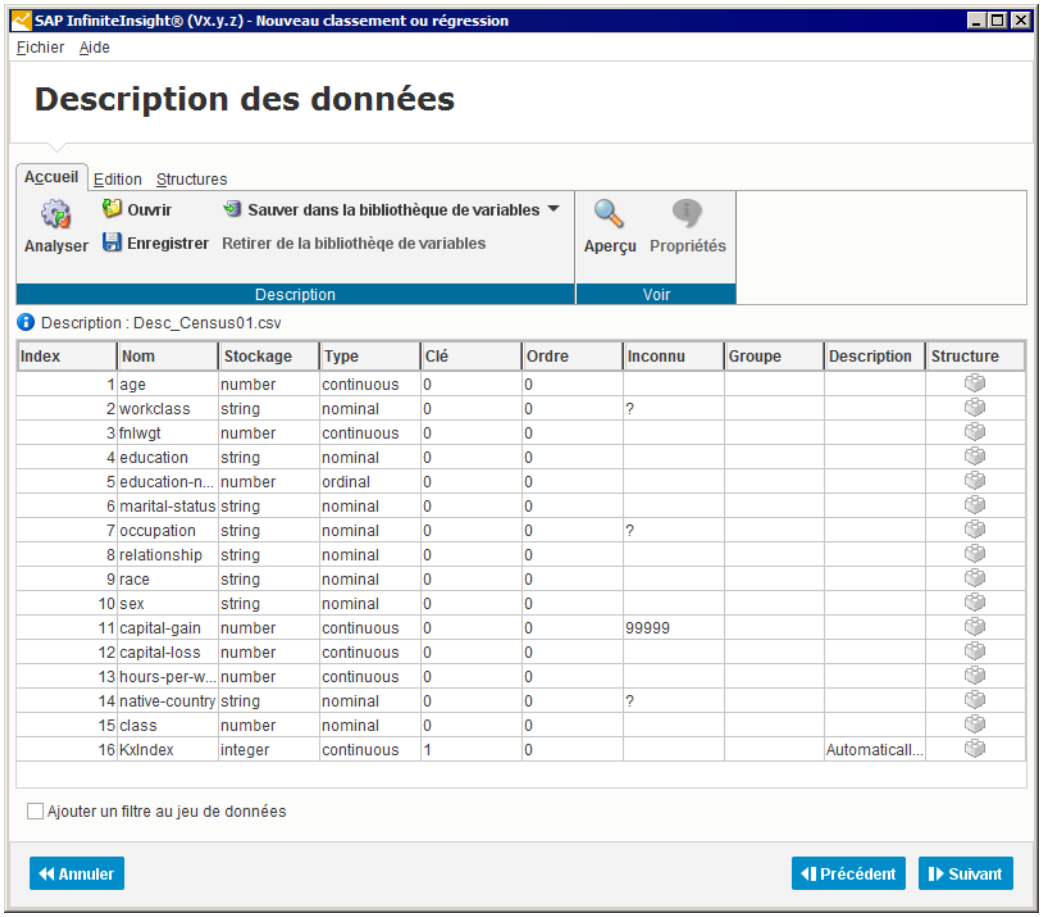
3 Utilisez le bouton *Parcourir* du champ *Répertoire* pour sélectionner le répertoire ou la base de données contenant la source de données.

> **Note**
>
> Le répertoire sélectionné par défaut est le même que celui sélectionné à l'étape précédente.

4 Utilisez le bouton *Parcourir* du champ *Fichier* pour sélectionner le fichier ou la table contenant les données.

> **Attention**
>
> Quand l'espace de données utilisé pour la construction du modèle contient une variable physique appelée *KxIndex*, il n'est pas possible d'utiliser un fichier de description ne comportant aucune clé pour l'espace de données courant.
>
> Quand l'espace de données utilisé pour la construction du modèle ne contient pas de variable nommée *KxIndex*, il n'est pas possible d'utiliser un fichier de description incluant une description à propos d'une variable *KxIndex* car cette variable n'existe pas dans l'espace de donnée courant.

5 Cliquez sur le bouton *OK*. La fenêtre *Ouvrir une description* se ferme et la description des données s'affiche dans la fenêtre principale.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

6 Cliquez sur le bouton *Suivant*.
L' écran *Sélection des variables explicatives* apparaît.

7 Passez à la section Sélectionner les variables explicatives.

☑ Pour créer un fichier de description

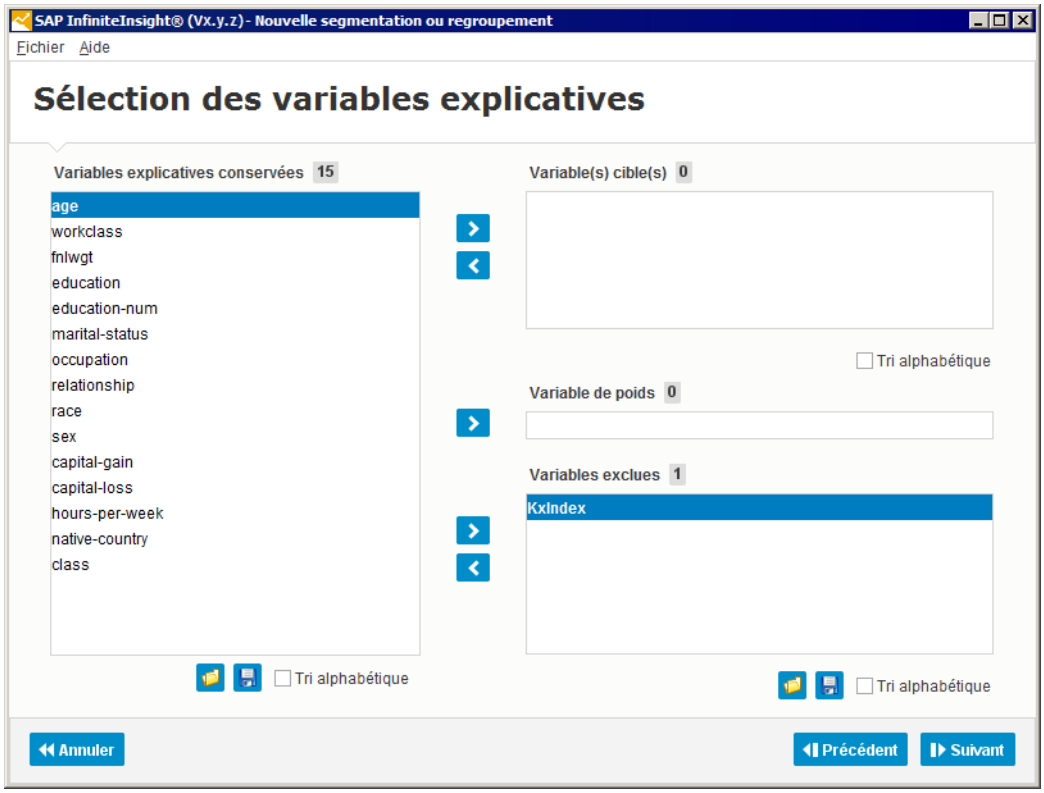
1 Dans l' écran *Description des données*, cliquez sur le bouton *Analyser*.
La description des données apparaît.

2 Vérifiez l'exactitude de la description obtenue.
Si votre fichier de données initial contient des variables qui ont fonction de clés, elles ne sont pas reconnues automatiquement. Décrivez-les manuellement.

> **Attention**
>
> L'espace de données source utilisé, qu'il s'agisse d'une fichier texte ou d'une base de données ODBC, doit contenir au minimum une variable clé.

3 Une fois la description des données validée, vous pouvez :
- la sauvegarder en cliquant sur le bouton *Enregistrer*.
- cliquer sur le bouton *Suivant* pour passer à l'étape suivante.

L'écran *Sélection des variables explicatives* apparaît.

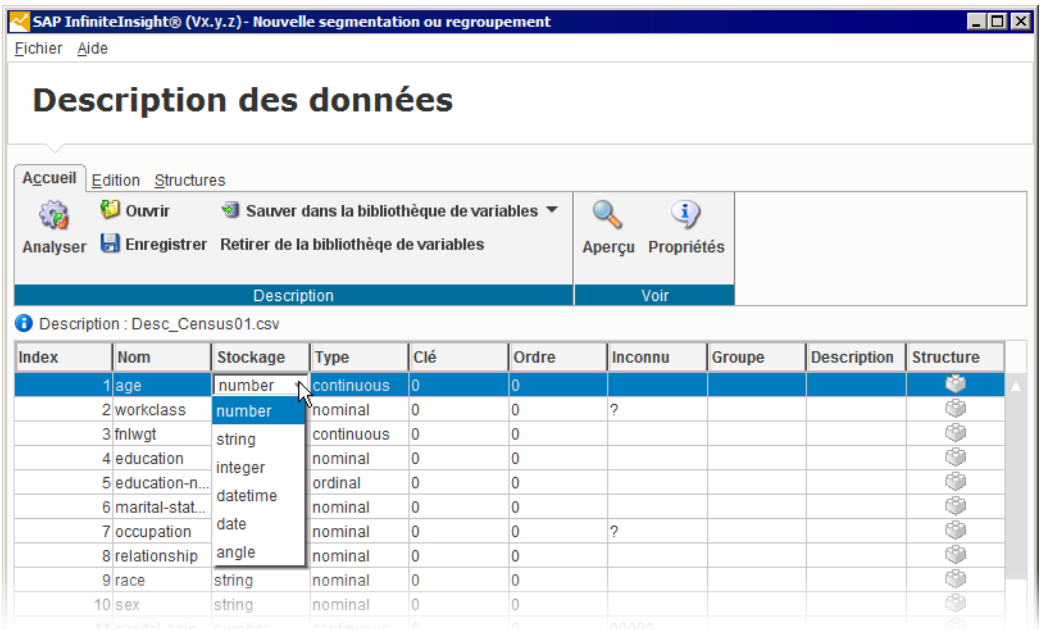


© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

4 Passez à la section Sélectionner les variables explicatives.

☑ Pour modifier la description des données

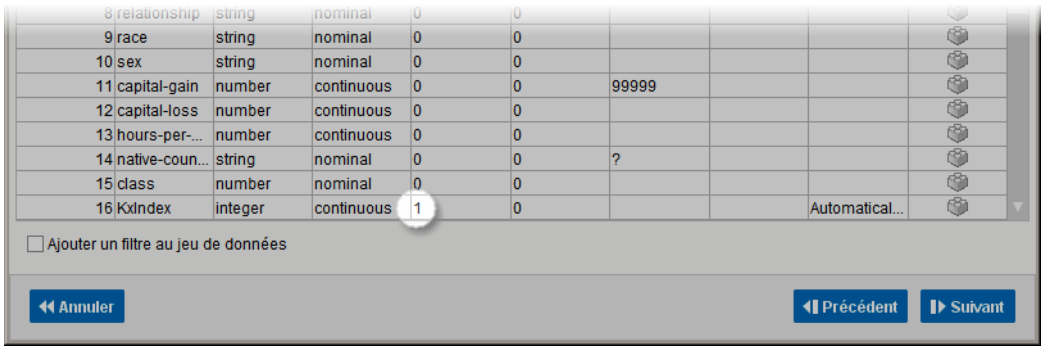
1 Dans la colonne de votre choix, par exemple la colonne *Stockage*, cliquez sur la case que vous souhaitez modifier.
La liste des valeurs possibles apparaît.



2 Sélectionnez la valeur souhaitée dans la liste.

☑ Pour spécifier qu'une variable est une clé

1 Dans la colonne *Clé*, cliquez sur la case correspondant à la ligne de la variable clé.
2 Entrez la valeur "*1*" pour définir cette variable comme clé.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

> **Note**
>
> Chaque modèle doit contenir une clé, c'est-à-dire qu'une ou plusieurs variables avec un champ clé ayant une valeur de clé différente de zéro. Si aucune clé n'a été détectée pendant le processus d'analyse et qu'aucune variable physique nommée *KxIndex* n'existe dans l'espace de données source, il est impossible d'ajouter une variable appelée *KxIndex* avec sa description. Une variable virtuelle ne peut pas être décrite.
>
> Dans ce cas particulier, en effet, les composants applicatifs de SAP InfiniteInsight® génèrent une variable-clé virtuelle nommée *KxIndex* et une description est ajoutée par les composants applicatifs InfiniteInsight® dans la colonne *Description* : 'Automatically added'.

Pourquoi décrire les données sélectionnées

Pour que vos données soient interprétables et analysables par les fonctionnalités SAP InfiniteInsight®, elles doivent être décrites. En d'autres mots, le fichier de description spécifie la nature de chaque variable en déterminant leur :

- format de **stockage** : nombre (*number*), chaînes de caractère (*string*), date et heure (*datetime*) ou date (*date*).

> **Note**
>
> Lorsqu'une variable est déclarée comme date (*date* ou *datetime*), la fonctionnalité <FR_KDC> (*KDC*) en extrait automatiquement des informations spécifiques telles que le jour du mois, l'année, le trimestre, etc. Des variables contenant ces informations sont créées lors de la génération du modèle et sont utilisées comme variables d'entrée. KDC est activé pour toutes les fonctionnalités SAP InfiniteInsight® à l'exception de InfiniteInsight® Modeler / Séries temporelles (*KTS*).

- **type** : variables continues (*continuous*), nominales (*nominal*) ordinales (*ordinal*) ou textuelle (*textual*).

> **Note**
>
> Toutes les variables décrites doivent se trouver dans la source de données utilisée pour l'apprentissage. Dans le cas où une variable physique décrite n'existe pas dans la source de données, il n'est pas possible de générer un modèle.

Pour plus d'informations sur la description des données, Types de variables à la page 27 et Formats de stockage à la page 30.

> **Note**
>
> La traduction des catégories d'une variable n'a pas d'influence sur sa structure qui doit être définie en fonction des valeurs initiales de la variable.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Comment décrire les données sélectionnées

Pour décrire vos données, vous pouvez :

- soit **utiliser un fichier de description existant**, c'est-à-dire issu de votre système d'information ou d'une précédente utilisation des fonctionnalités SAP InfiniteInsight®,
- soit **créer un fichier de description grâce à l'option** *Analyser*, mise à votre disposition dans l'assistant de modélisation SAP InfiniteInsight®. Dans ce cas, vous devez valider le fichier de description obtenu. Vous pouvez sauvegarder ce fichier pour une utilisation ultérieure.

Attention

Le fichier de description obtenu avec l'option *Analyser* résulte de l'analyse des 100 premières lignes du fichier de données initial. Afin d'éviter tout biais, n'hésitez pas à brasser votre jeu données avant de l'analyser.

Le scénario d'utilisation standard [ouverture d'un espace de donnée ODBC - description en utilisant la fonction d'*Analyse* - génération du modèle] ne peut pas être mis en oeuvre lorsque l'espace de données source contient une variable nommée *KxIndex* mais aucune variable ODBC ayant le statut de clé.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

La description d'une variable est composée des champs décrits dans le tableau ci-dessous :

Le champ...	contient...
Nom	le nom de la variable (celui-ci ne peut être modifié)
Stockage	le type de valeurs stockées dans cette variable : ▪ *Number* : la variable contient uniquement des nombres "caculables" (attention : les numéros de téléphone, codes postaux, numéros de compte ne doivent pas être considérés comme des nombres) ▪ *String* : la variable contient des chaînes de caractères. ▪ *Datetime* : la variable contient des dates et des heures ▪ *Date* : la variable contient des dates
Type	le type de la variable : ▪ *Continuous* : une variable numérique pour laquelle la moyenne, la variance, etc. peuvent être calculées. ▪ *Nominal* : variable catégorique, seul type possible pour une chaîne de caractère (les codes postaux, numéros de téléphone, etc. sont généralement de ce type). ▪ *Ordinal* : variable numérique discrète pour laquelle l'ordre est important ▪ *Textual* : variable textuelle contenant des mots, des phrases ou des textes complets. Attention - lors de la création d'un modèle d'analyse textuelle, si aucune variable textuelle n'est définie le bouton *Suivant* est désactivé et il est impossible de passer à l'étape suivante.
Clé	indique si cette variable est une clé ou un identifiant pour l'observation : ▪ *0* la variable l'est pas un identifiant; ▪ *1* clé primaire; ▪ *2* clé secondaire...
Ordre	indique si la variable représente un ordre naturel. Dans un jeu de données d'évènements il doit y avoir au moins une variable marquée comme ordonnée. Attention - si la source de données est un fichier et que la variable marquée comme représentant un ordre naturel n'est pas effectivement ordonnée, un message d'erreur s'affichera au moment de la vérification ou de la génération du modèle.
Inconnu	la chaîne de caractères utilisée dans le fichier de description pour représenter les valeurs manquantes (par exemple "999" ou "#Vide" - sans les guillemets)
Groupe	le nom du groupe auquel appartient la variable. les variables appartenant à un même groupe sont considérées comme apportant la même information et ne seront donc pas croisées dans les modèles d'ordre supérieur à 1. Ce paramètre sera activé dans une future version.
Description	une éventuelle description supplémentaire de la variable
Structure	structure de la variable, c'est-à-dire les groupements des catégories des variables.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Un mot sur les clés de base de données

Pour des raisons de gestion des données et de performance, le jeu de données à analyser doit comporter une variable ayant fonction de clé. Deux cas se présentent :

- Si le jeu de données initial ne contient **pas de variable clé**, une variable index *KxIndex* est automatiquement créée par les fonctionnalités SAP InfiniteInsight®. Elle correspondra au numéro de la ligne de données traitée.

  > **Note**
  >
  > Il n'est pas possible de forcer l'indice de clé (Key Level) à 0 pour une clé virtuelle si aucune autre clé n'a été définie.

- Si le fichier contient **une ou plusieurs variables clés**, ces dernières ne sont pas automatiquement reconnues. Vous devez alors le spécifier manuellement dans la description des données en renseignant l'indice de clé à la valeur appropriée. Se reporter à la procédure **Pour spécifier qu'une variable est une clé**. Par ailleurs, si vos données sont stockées dans une base de données, elles seront automatiquement reconnues.

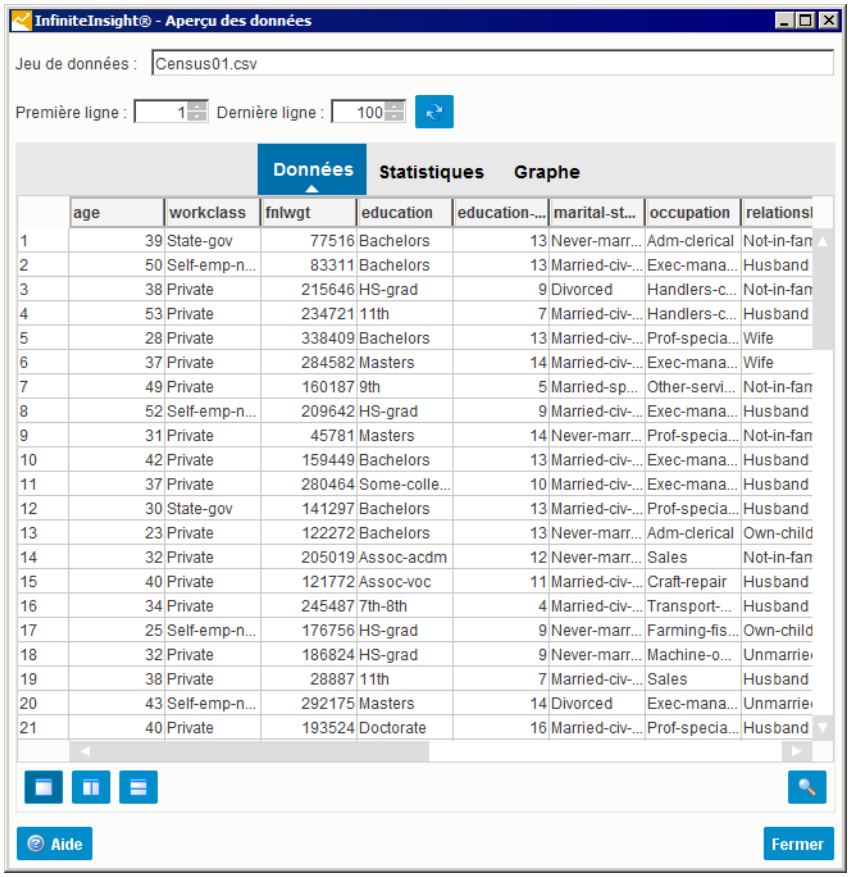
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Voir les données

Pour vous aider à valider la description obtenue par analyse, vous pouvez afficher le contenu de votre jeu de données.

☑ Pour voir les données

1 Cliquez sur le bouton *Aperçu*. Une nouvelle fenêtre s'ouvre affichant les cent premières lignes du jeu de données.



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	23	Private	122272	Bachelors	13	Never-marr...	Adm-clerical	Own-child
14	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
15	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
16	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
17	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
18	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
19	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
20	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
21	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband

2 Dans le champ *Première ligne*, saisissez le numéro de la première ligne à afficher.

3 Dans le champ *Dernière ligne*, saisissez le numéro de la dernière ligne à afficher.

4 Cliquez sur le bouton  (*Rafraîchir)* pour afficher les lignes sélectionnées.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

7.1.3 Ajouter un filtre au jeu de données

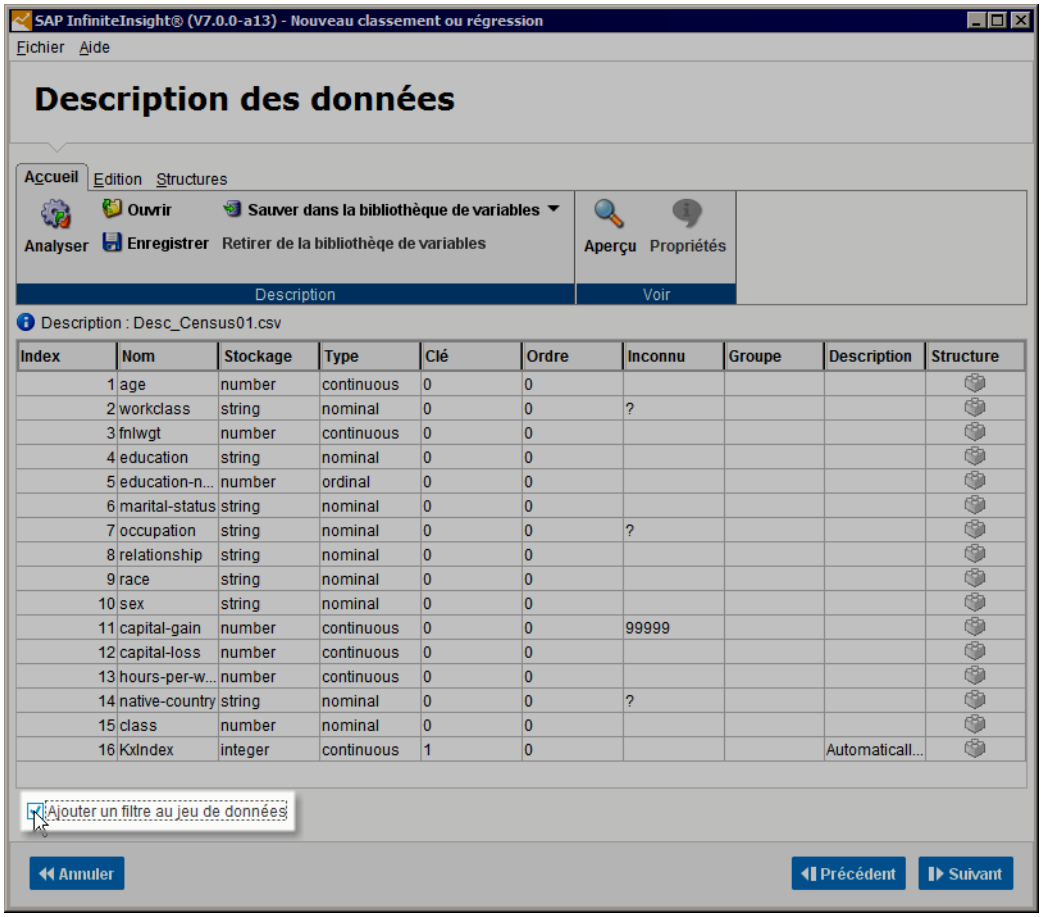
Vous avez la possibilité d'appliquer un filtre à votre jeu de données afin d'accélérer le processus d'apprentissage et d'optimiser le modèle qui en résulte.

Pour ce scénario

N'utilisez pas de filtre pour votre jeu de données.

Ajouter un filtre

1 Cochez la case *Ajouter un filtre au jeu de données*.

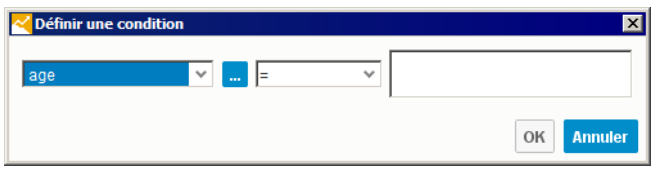


Index	Nom	Stockage	Type	Clé	Ordre	Inconnu	Groupe	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

2 Cliquez sur *Suivant*.

Ajouter une condition

1 Cliquez sur le bouton *Ajouter une condition*.
La fenêtre *Définir une condition* s'ouvre.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2. Choisissez une variable dans la première liste déroulante.
3. Choisissez un opérateur dans la deuxième liste.
4. Indiquez une valeur dans la troisième liste :
   - Pour une variable du type *Number* entrez une valeur.
   - Pour une variable du type *String* choisissez une variable dans la liste. Si cette liste est vide, cliquez sur le bouton  pour extraire les catégories.
5. Cliquez sur *OK*.

> **Note**
>
> Vous pouvez modifier une condition en double-cliquant dessus.

☑ Ajouter une conjonction logique

Cliquez sur le bouton *Ajouter un "ET" logique* ou sur le bouton *Ajouter un "OU" logique*.

> **Note**
>
> Vous pouvez modifiez le type de conjonction en double-cliquant dessus.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

☑ Changer l'ordre

Vous pouvez changer l'ordre des noeuds pour accélérer l'application du filtre en mettant les conditions, qui ont une grande probabilité de s'avérer fausse, en haut de la liste.

1 Sélectionnez le noeud que vous voulez déplacer vers le haut ou vers le bas.

2 Utilisez les boutons et pour changer sa position dans le filtre.

☑ Supprimer un noeud

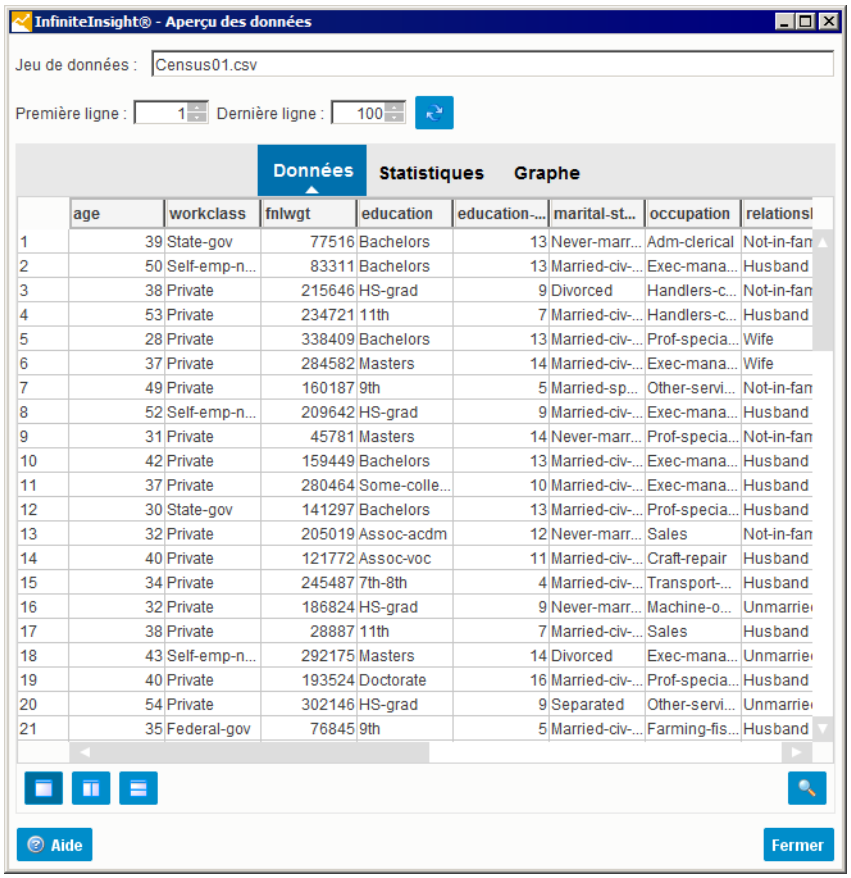
1 Sélectionnez le noeud que vous voulez supprimer.

2 Cliquer sur le bouton *Supprimer le noeud sélectionné*.

☑ Afficher le jeu de données filtré

Vous pouvez visualiser le jeu de données qui vous obtiendrez en appliquant le filtre.

Cliquez sur le bouton *Aperçu*.
Une nouvelle fenêtre s'ouvre.



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
14	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
15	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
16	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
17	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
18	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
19	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband
20	54	Private	302146	HS-grad	9	Separated	Other-servi...	Unmarrie
21	35	Federal-gov	76845	9th	5	Married-civ-...	Farming-fis...	Husband

☑ Enregistrer un filtre

Vous pouvez enregistrer le filtre créer pour le réutiliser ultérieurement sans être obligé de recréer un filtre avec les mêmes conditions.

1. Cliquez sur le bouton *Enregistrer ce filtre*.
   La fenêtre *Enregistrer ce filtre* s'ouvre.
2. Dans la liste *Type de données*, sélectionnez le format de l'enregistrement.
3. Utilisez le bouton *Parcourir* à droite du champ *Répertoire* pour choisir un répertoire ou une base de données pour l'enregistrement.
4. Dans le champ *Description*, entrez le nom du fichier ou de la table.
5. Cliquez sur *OK*.

☑ Charger un filtre existant

Pour filtrer un jeu de donnée, vous pouvez utiliser un filtre préalablement créé avec SAP InfiniteInsight® pour ce jeu de données.

1. Cliquez sur le bouton *Charger un filtre existant*.
   La fenêtre *Charger un filtre existant* s'ouvre.
2. Utilisez la liste déroulant Type de données pour sélectionner le format du filtre.
3. Utilisez le bouton *Parcourir* à droite du champ *Répertoire* pour choisir le répertoire ou la base de données où se trouve le filtre.
4. Utilisez le bouton *Parcourir* à droite du champ *Description* pour choisir le fichier ou la table contenant le filtre.
5. Cliquez sur *OK*.

7.1.4 Traduire les catégories de variables

Vous pouvez traduire les catégories des variables nominales, enregistrer la traduction ou charger une traduction existante. Cette traduction n'influence pas la structure de la variable, qui doit être définie en fonction des valeurs originales de la variable.

> **Note**
>
> La variable "Catégorie cible", utilisée par exemple dans les paramètres avancés, ne prend pas en compte une éventuelle traduction quand les valeurs possibles de cette variable sont affichées. Pour cette raison des valeurs entrées manuellement ne peuvent pas être traitées correctement, si elles ne correspondent pas aux valeurs d'origine.

☑ Traduire les catégories de variables

1. Faites un clique droit sur la variable nominale dont vous souhaitez traduire les catégories. Un menu contextuel est affiché.
2. Sélectionnez l'option *Traduire les catégories de <nom_de_la_variable>*.
3. Choisissez dans quelles langues vous voulez traduire. Par défaut, la langue de l'interface utilisateur est affichée comme colonne.
4. Cliquez sur le bouton  pour extraire les catégories de variables du jeu de données.
5. Traduisez les catégories.

> **Note**
>
> Vous n'êtes pas obligé de renseigner tous les champs.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

6 Cliquez sur *OK*.

☑ Enregistrer la traduction des catégories

1 Traduisez les catégories de variables comme expliqué ci-dessus.
2 Cliquez sur le bouton *Enregistrer*.
3 Choisissez un *Type de données*.
4 Sélectionnez un *Répertoire*.
5 Entrez un *Nom* pour le fichier ou la table.
6 Cliquez sur *OK*.

☑ Charger une traduction existante

1 Faites un clique droit sur une variable nominale. Un menu contextuel est affiché.
2 Sélectionnez l'option *Traduire les catégories de <nom_de_la_variable>*.
3 Cliquez sur le bouton *Charger*.
4 Sélectionnez le format de la traduction dans la liste *Type de données*.
5 Utilisez le bouton *Parcourir* situé à droite du champ *Répertoire* pour choisir le répertoire ou la base de données contenant la traduction.
6 Utilisez le bouton *Parcourir* situé à droite du champ *Table ou fichier* pour choisir la traduction des catégories de variables.
7 Cliquez sur le bouton *OK*.
8 Cliquez sur le bouton  *Rafraîchir* pour actualiser l'affichage des catégories.
9 Si les colonnes ne sont pas nommées correctement, utilisez les Paramètres avancés  (voir paragraphe suivant) pour choisir la ligne d'en-tête et actualisez à nouveau.
10 Mettez les noms des langues en correspondance avec les langues de la traduction chargée en cliquant sur les catégories et en choisissant la langue qui correspond dans le menu contextuel.
11 Cliquez sur le bouton *OK*.

7.1.5 Sélectionner les variables

Une fois le jeu de données d'apprentissage et sa description chargés, vous devez sélectionner :

- la ou les variables à utiliser comme variables cibles si vous le souhaitez,
- éventuellement une variable de poids,
- les variables explicatives.

Sélectionner les variables cibles

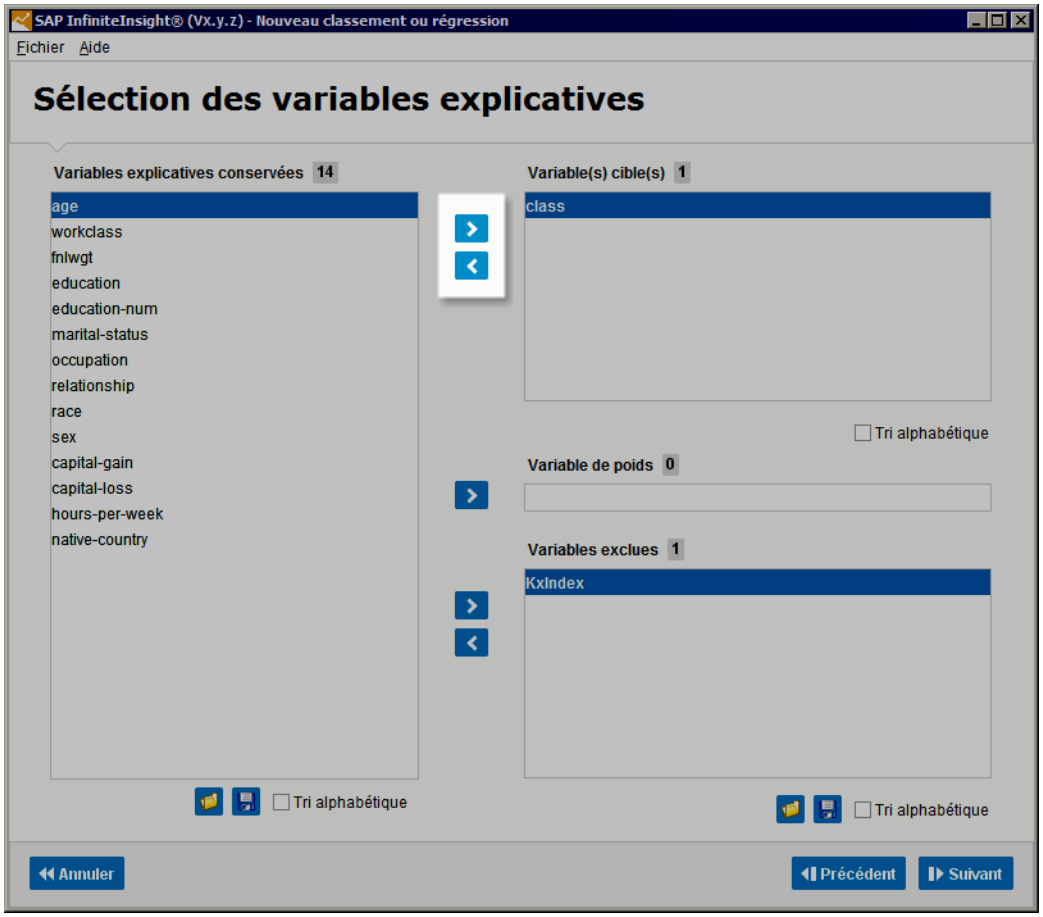
Une fois le jeu de données d'apprentissage et sa description chargés, vous pouvez sélectionner une variable à utiliser comme variable cible. InfiniteInsight® Modeler / Segmentation est capable de segmenter un jeu de données de manière absolue, c'est-à-dire sans qu'aucune variable cible ne soit sélectionnée. Même si elle n'est pas obligatoire, la sélection d'une variable cible est cependant fortement conseillée. En effet, la segmentation d'un jeu de données prend tout son sens quand elle est réalisée en fonction d'une problématique métier, exprimée par une variable cible.

Pour ce scénario

Sélectionnez pour **variable cible** la variable *Class*, c'est-à-dire la variable indiquant la probabilité d'un individu à répondre de manière positive ou négative à votre campagne.

Pour sélectionner la variable cible

1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez la ou les variables choisies comme cibles.



Remarque

Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

2 Cliquez sur le bouton > situé gauche du champ *Variable(s) cible(s)*.
Les variables sélectionnées passent dans la partie *Variable(s) cible(s)*.

3 Pour retirer une ou plusieurs variables de la liste des variables cibles, sélectionnez celles-ci dans la liste puis cliquez sur le bouton <.

4 Passez à la section Sélectionner la variable de poids (à la page 211).

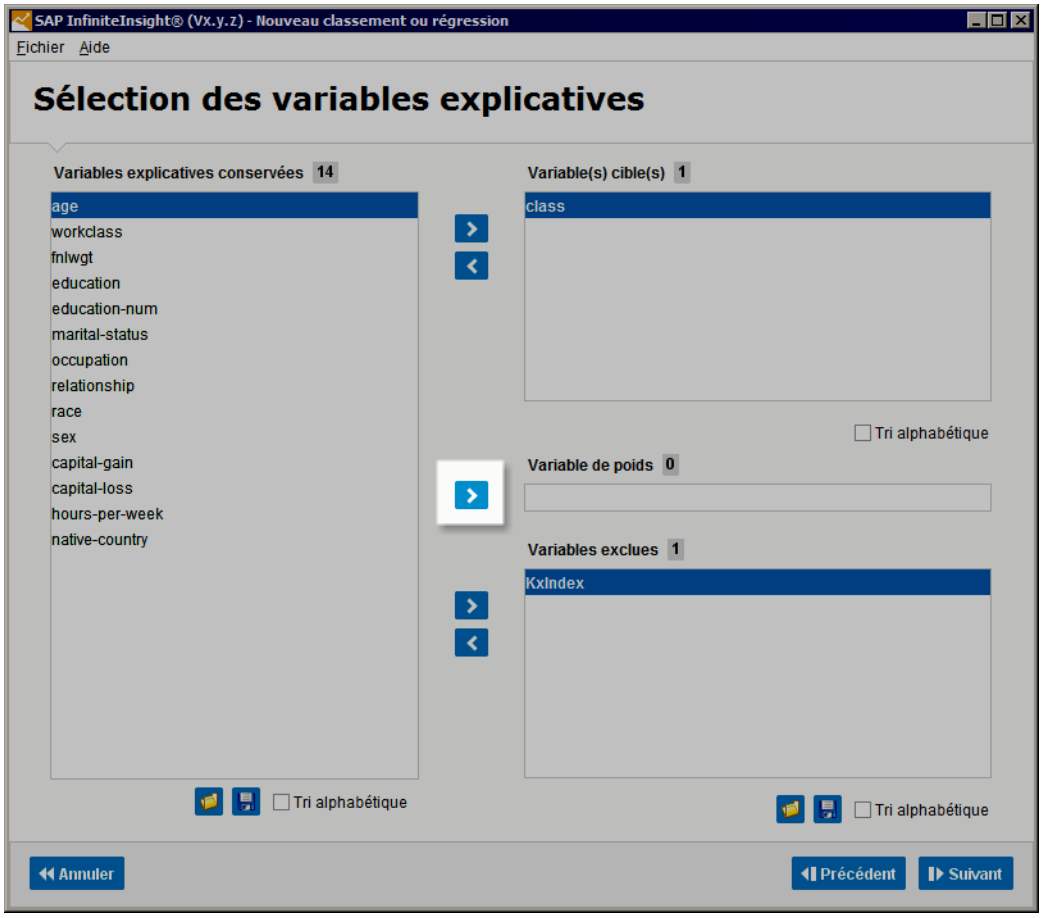
Sélectionner la variable de poids

Pour ce scénario

Ne sélectionnez **aucune** variable de poids.

Pour sélectionner une variable de poids

1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez la variable à utiliser comme variable de poids.



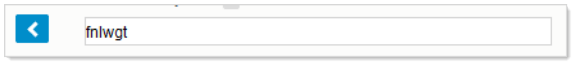
Remarque

Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 Cliquez sur le bouton > situé gauche du champ *Variable de poids*.
La variable passe dans le champ *Variable de poids*.

3 Pour supprimer la variable de poids, cliquez sur le bouton <.



4

5 Passez à la section Sélectionner les variables explicatives (à la page 213).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Sélectionner les variables explicatives

Par défaut, et à l'exception des variables clés, toutes les variables contenues dans votre jeu de données sont prises en compte pour la génération du modèle. Vous pouvez exclure certaines de ces variables.

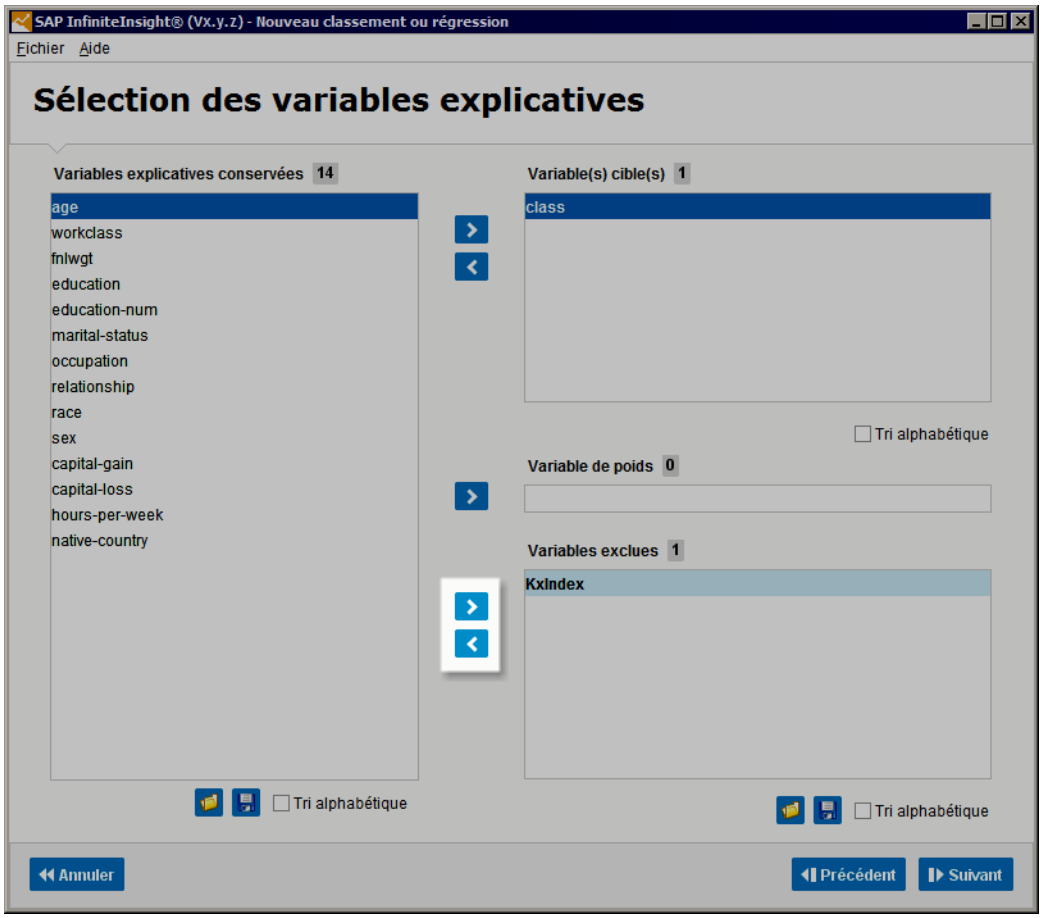
Le choix d'exclure ou d'inclure une variable dans la génération d'un modèle de segmentation dépend de considérations métiers. Votre connaissance métier vous permet de déterminer quelles sont les variables les plus intéressantes pour la description du jeu de données en groupes homogènes. Un modèle de régression généré avec *InfiniteInsight® Modeler* constitue également un outil pour déterminer les variables les plus explicatives d'un phénomène.

Pour ce scénario

- Laissez la variable *KxIndex* exclue. Cette variable est une variable clé. Le jeu de données initial ne contenant pas de variable clé, les composants SAP InfiniteInsight® ont généré automatiquement la variable *KxIndex*.
- Conservez toutes les autres variables.

Pour exclure des variables de l'analyse des données

1 Dans l'écran *Sélection des variables explicatives*, dans la partie *Variables explicatives conservées* (partie de gauche), sélectionnez les variables à exclure.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

2 Cliquez sur le bouton > situé gauche du champ *Variables exclues*.
Les variables sélectionnées passent dans la partie *Variables exclues*.

3 Pour retirer une ou plusieurs variables de la liste des variables exclues, sélectionnez celles-ci dans la liste des variables exclues, puis cliquez sur le bouton <.

> **Note**
>
> Par défaut, toute variable définie comme clé est exclue automatiquement : elle figure dans la section *Variables Exclues*. Cependant, l'utilisateur a la possibilité de déplacer une variable clé dans la section *Variables Explicatives Conservées* s'il veut que cette variable joue un tel rôle.

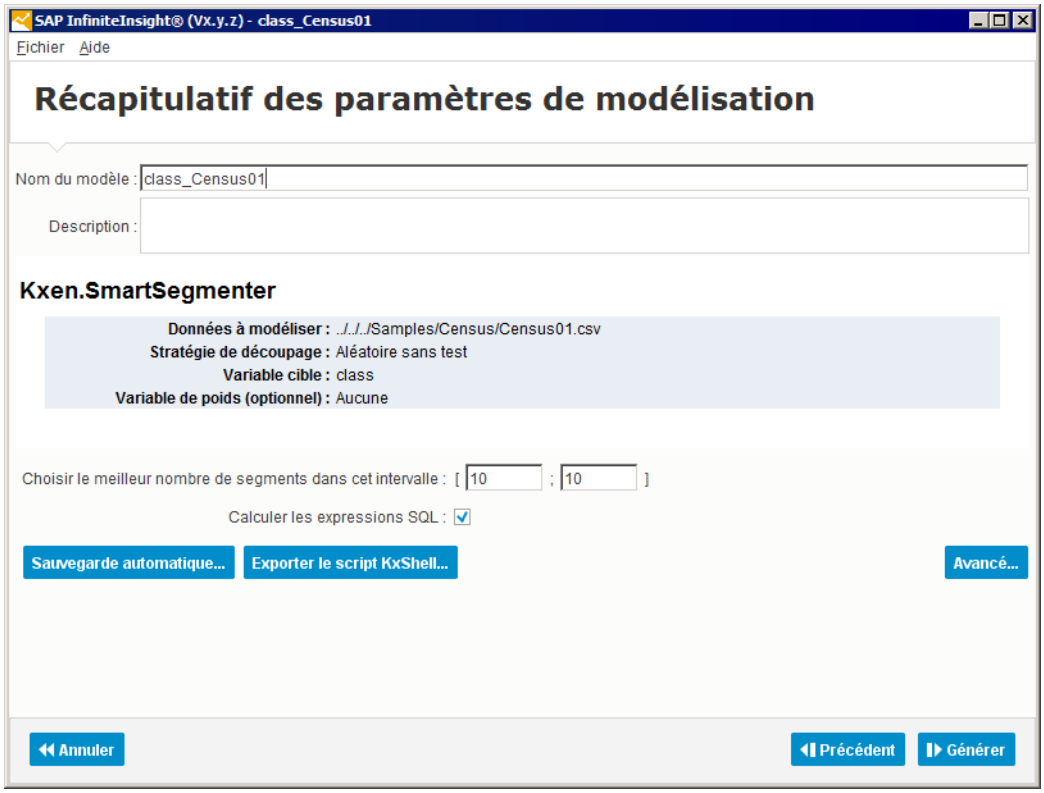
4 Cliquez sur le bouton *Suivant*.
L'écran *Récapitulatif des paramètres de modèlisation* apparaît.

5 Passez à la section Vérifier les paramètres de modélisation.

> **Remarque**
>
> Dans l'écran *Sélection des variables explicatives*, les variables sont présentées dans le même ordre que celui dans lequel elles sont présentées dans la table de données. Pour les trier de manière alphabétique, sélectionnez l'option *Tri alphabétique*, présentée sous chacune des parties de l'écran.

7.1.6 Vérifier les paramètres de modélisation

L'écran *Récapitulatif des paramètres de modélisation* vous permet d'effectuer une dernière vérification des paramètres de modélisation avant de générer le modèle.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

> **Note**
>
> L'écran *Récapitulatif des paramètres de modélisation* présente également un bouton *Avancé*. Ce bouton vous permet d'accéder à l'écran *Paramètres spécifiques du modèle* dans lequel vous pouvez choisir de calculer les statistiques croisées pour le modèle à générer. Pour plus d'informations, voir la section Paramètres spécifiques du modèle (voir à la page 217).

- Le **nom du modèle** est renseigné automatiquement. Il correspond au nom de la variable cible (`class` pour notre scénario), suivi du signe underscore ("_") et du nom de la source de données, sans son extension de fichier (`Census01` pour notre scénario).
- Le bouton *Sauvegarde automatique* vous permet de spécifier que le modèle doit être automatiquement enregistré dès la fin de la génération du modèle. Les informations d'enregistrement sont paramétrables dans le panneau *Sauvegarde automatique*. Lorsque la sauvegarde automatique est activée, une coche verte s'affiche sur le bouton.

> **Note**
>
> Pour plus de détails, reportez-vous à la section Activation de la sauvegarde automatique (à la page 90).

Avant de générer le modèle, vous pouvez :

- activer la **sauvegarde automatique** du modèle,
- définir le **nombre de segments** que vous souhaitez obtenir,
- choisir de calculer les **expressions SQL** définissant les segments trouvés par le modèle,
- spécifier les **paramètres spécifiques** du modèle.

Définir le nombre de segments

D'un point de vue méthodologique, vous pouvez retenir que plus le nombre de segments est élevé :

- plus il est possible de trouver des segments très différents les uns des autres,
- plus le nombre d'observations nécessaires pour assurer la robustesse de la segmentation est élevé.

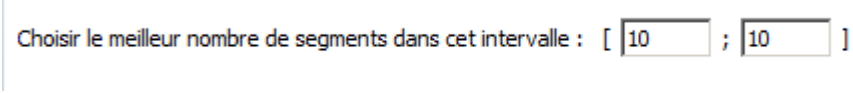
Il est conseillé d'effectuer plusieurs segmentations, en modifiant à chaque fois le nombre segments, jusqu'à obtenir une décomposition particulièrement intéressante du jeu de données.

Pour ce scénario

Définir un nombre de segments dont l'intervalle est égal à 1.

Pour définir le nombre de segments

Sur l'écran *Récapitulatif des paramètres de modélisation*, dans le champ *Choisir le meilleur nombre de segments dans cet intervalle*, entrez le **nombre de segments** que vous souhaitez obtenir.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Pour une **segmentation non supervisée** (c'est-à-dire sans variable cible), l'utilisateur choisit le meilleur nombre de segments, par exemple [5;10] signifiant que l'utilisateur souhaite avoir entre 5 et 10 segments. Le moteur SAP InfiniteInsight® choisit le meilleur nombre de segments en se basant sur le calcul capacité prédictive (KI) + reproductibilité (KR), par exemple 7 segments.

Pour une **segmentation supervisée** (c'est-à-dire avec variable cible), le moteur SAP InfiniteInsight® calcule le nombre minimum de segments, par exemple [10;10], soit 10 segments.

Attention

Lorsque l'option *Calculer les expressions SQL* est activée, SAP InfiniteInsight® crée un segment supplémentaire contenant les observations non assignées (pour plus de détails sur les expressions SQL et les observations non assignées, **Différence entre statistiques croisées classiques et expressions SQL** (à la page 249)).

Calculer les expressions SQL

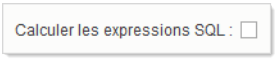
Les expressions SQL permettent de visualiser les requêtes SQL correspondant à chaque segment créé lors de la génération du modèle. Le calcul des expressions SQL est activé par défaut.

Pour ce scénario

Sélectionnez l'option *Calcul des expressions SQL*.

Pour désélectionner le calcul des expressions SQL,

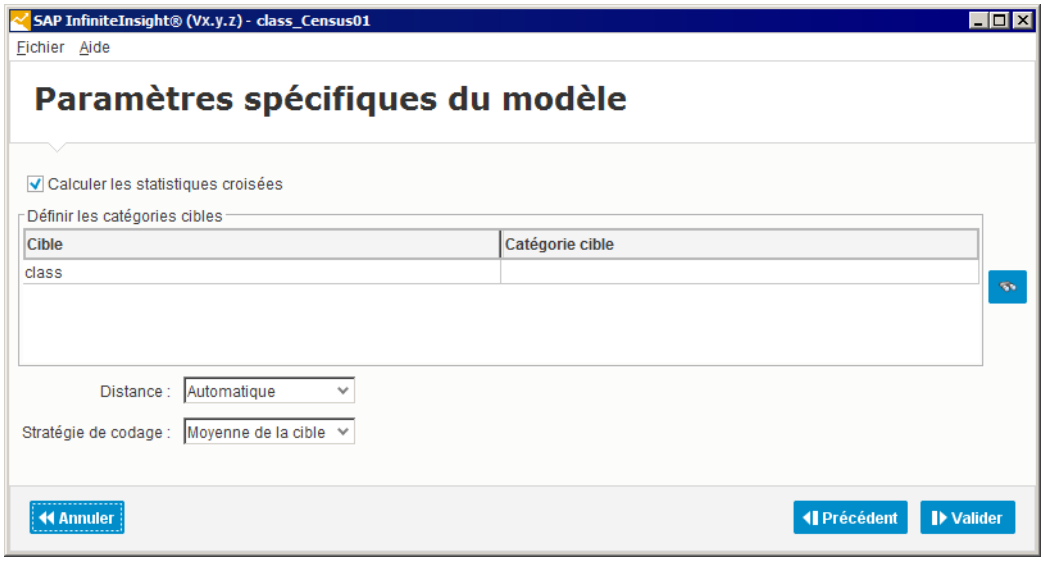
Décochez la case *Calculer les expressions SQL*.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Paramètres spécifiques du modèle

En cliquant sur le bouton *Avancé...* de l'écran *Récapitulatif des paramètres de modélisation*, vous accédez à un écran vous permettant de sélectionner les paramètres spécifiques du modèle.



Parmi les paramètres à sélectionner, vous pouvez :

- activer le **calcul des statistiques croisées**,
- sélectionner le mode de calcul de la **distance**,
- sélectionner la **stratégie de codage**,
- définir la valeur clé des **catégories cibles**.

Ces options sont détaillées ci-dessous.

Activer le calcul des statistiques croisées

Cette option vous permet de visualiser le profil de chaque variable explicative pour chaque segment, comparé à son profil pour l'ensemble du jeu de données.

Pour ce scénario

- Sélectionnez l'option *Calcul des statistiques croisées*.

Pour sélectionner le calcul des statistiques croisées

Cochez la case *Calculer les statistiques croisées*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Choisir la distance à utiliser

La liste **Distance** vous permet de spécifier la distance à utiliser pour comparer les données d'entrée une fois codées par le codeur analytique d'SAP InfiniteInsight®.

Ce paramètre peut prendre les valeurs suivantes :

- *"Chessboard"* : la somme des valeurs absolues des différences entre les coordonnées (LInf).
- *Euclidienne* : racine carrée de la somme des carrés des différences entre les coordonnées (L2).
- *"City Block"* : maximum de la valeur absolue des différences entre les coordonnées (L1).
- *Automatique* (valeur par défaut) : le système sélectionne la distance la plus appropriée selon les paramètres du modèle.

> **Note**
>
> La politique actuelle est d'utiliser LInf en mode non supervisé ou lorsque les expressions SQL ont été demandées et L2 dans tous les autres cas.

Pour ce scénario

- Gardez la valeur par défaut.

Pour sélectionner la distance à utiliser

Dans la liste *Distance*, sélectionnez l'option choisie.

Stratégie de codage

L'option *Stratégie de codage* permet de définir le type de codage que le moteur de segmentation attend de l'encodeur analytique de *InfiniteInsight® Modeler*.

Pour sélectionner une stratégie de codage :

Dans la liste déroulante, choisissez une option parmi celles décrites ci-dessous :

Option	Description
Automatique	Laisse le système sélectionner le meilleur codage d'après les paramètres du modèle. Le codage *Moyenne de la cible* est utilisé pour les modèles supervisés. Pour les modèles non-supervisés, c'est l'option *Non supervisé* qui sera utilisée.
Moyenne de la cible	*Valeur par défaut pour la segmentation supervisée* Chaque valeur d'une variable continue est remplacée par la moyenne de la variable cible sur le segment auquel la valeur appartient. Chaque catégorie d'une variable nominale est remplacée par la moyenne de la variable cible pour cette catégorie. Dans le cas d'une variable cible nominale, la moyenne de la variable cible correspond au pourcentage de cas positifs de la variable cible pour cette catégorie.
Uniforme	Chaque segment de variable est codé dans l'intervalle [-1;+1] afin que la distribution des variables soit uniforme.
Non supervisé	*Valeur par défaut pour la segmentation supervisée* Une stratégie sans cible. Seule la fréquence des segments est utlisée pour coder les variables.

Les options suivantes ne sont disponbiles que **lorsque toutes les variables sont continues** :

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Option	Description
Natural	Aucune transformation n'est appliquée aux données d'entrée.
Min Max	Les variables sont codées dans l'intervalle [0,1], où 0 correspond à la valeur minimale de la variable et 1 à sa valeur maximale.
Normalisation de l'écart-type	Cette option applique une normalisation reposant sur la moyenne de la variable et l'écart-type. $\frac{x - Mean}{StdDev}$

7.2 Etape 2 - Générer et valider le modèle

Une fois les paramètres de modélisation définis, vous pouvez générer le modèle. Vous devez ensuite valider ses performances grâce aux indicateurs de qualité KI et de robustesse KR :

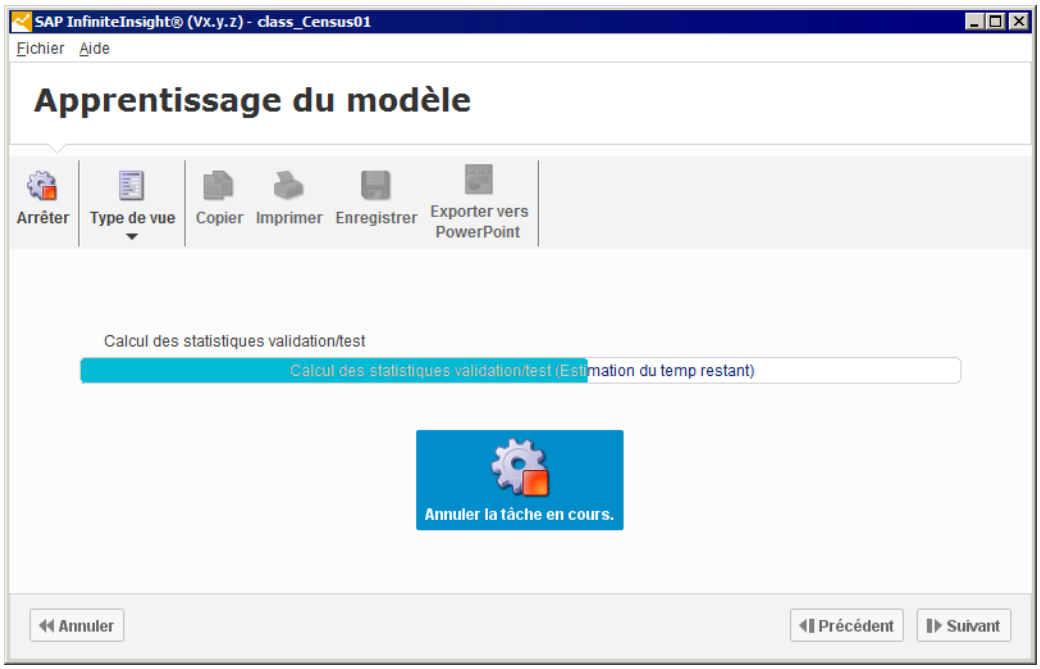
- Si le modèle est suffisamment performant, vous pouvez analyser les réponses qu'il apporte par rapport à votre problématique (étape 3 à la page 108, à la page 225), puis l'appliquer sur de nouveaux jeux de données (étape 4).
- Sinon, vous pouvez modifier les paramètres de modélisation de manière à ce qu'ils soient plus adaptés à votre jeu de données et à votre problématique, et générer ainsi de nouveaux modèles plus performants.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

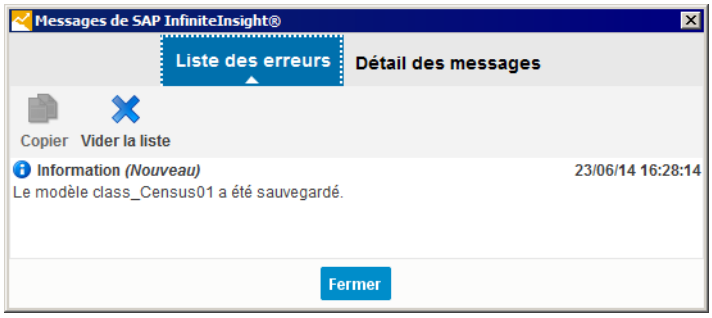
7.2.1 Générer le modèle

☑ Pour générer le modèle

1 Dans l'écran *Récapitulatif des paramètres du modèle*, cliquez sur le bouton *Générer*.
L'écran *Apprentissage du modèle* apparaît. La génération du modèle est en cours. Une barre de progression vous permet de suivre le déroulement des différentes étapes.



2 Si l'option *Sauvegarde automatique* a été activée dans le panneau *Récapitulatif des paramètres de modélisation*, un message d'alerte s'affiche à la fin du processus de génération du modèle indiquant que celui-ci a bien été enregistré.



Cliquez sur le bouton *Fermer*.

3 Une fois le modèle généré, passez à la section Valider le modèle généré (voir à la page 70).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

7.2.2 Suivi du processus de génération

Il existe deux manières de suivre la progression du processus de génération du modèle :

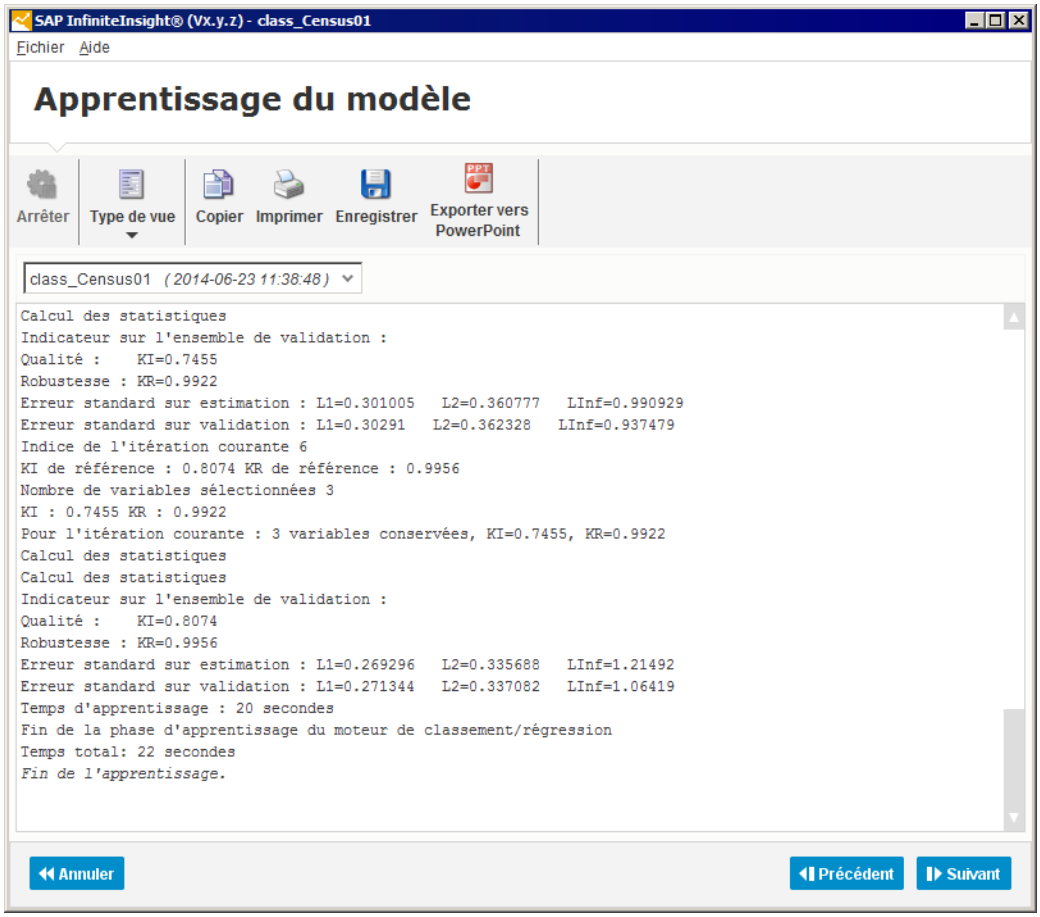
- La **Barre de progression** affiche la progression de chaque étape. C'est l'écran par défaut..
- Le **Détail du processus** affiche des messages détaillés pour chaque étape.

☑ Pour afficher la barre de progression

Cliquez sur le bouton  (*Affiche la progression*).
La barre de progression s'affiche.

☑ Pour afficher le détail du processus

Cliquez sur *Type de Vue* >  *Détails des messages*.
L'écran ci-dessous s'affiche.



☑ Pour arrêter le processus d'apprentissage

1 Cliquez sur le bouton  (*Arrêter*).
Une boîte de dialogue de confirmation s'affiche.

2 Cliquez sur le bouton *Précédent*.
L'écran *Récapitulatif des paramètres de modélisation* s'affiche.

3 Reportez-vous à la section Vérifier les paramètres de modélisation.

CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

7.2.3 Valider le modèle généré

Une fois le modèle généré, vous devez vérifier sa validité en observant les indicateurs de performance :

- la **capacité prédictive** vous permet de connaître le **pouvoir explicatif** du modèle, c'est-à-dire sa capacité à expliquer les valeurs de la variable cible sur le jeu de données d'apprentissage. Un modèle parfait possède une capacité prédictive égale à 1 et un modèle purement aléatoire possède une capacité prédictive égale à 0.
- la **reproductibilité** vous permet de connaître le **degré de robustesse** du modèle, c'est-à-dire sa capacité à conserver le même pouvoir explicatif sur un nouveau jeu de données. En d'autres mots, le degré de robustesse correspond à la capacité prédictive du modèle sur un jeu de données d'application.

Pour savoir comment sont calculés la capacité prédictive et la reproductibilité, voir Capacité prédictive, reproductibilité et courbes de profit à la page 232.

Remarque

La validation du modèle est une phase primordiale dans le processus global de Data Mining. Accordez toujours une importance majeure aux valeurs obtenues pour la capacité prédictive et la reproductibilité d'un modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Pour valider un modèle de segmentation, vous pouvez également observer les valeurs des indicateurs "fréquence" et "moyenne de la cible" de chacun des segments identifiés. En effet, les segments les plus intéressants d'une segmentation possèdent une "fréquence" élevée et une "moyenne de la cible" différente de la "moyenne de la cible" calculée sur la totalité du jeu de données. Or, un modèle de segmentation dont la capacité prédictive est faible peut receler de tels types de segments.

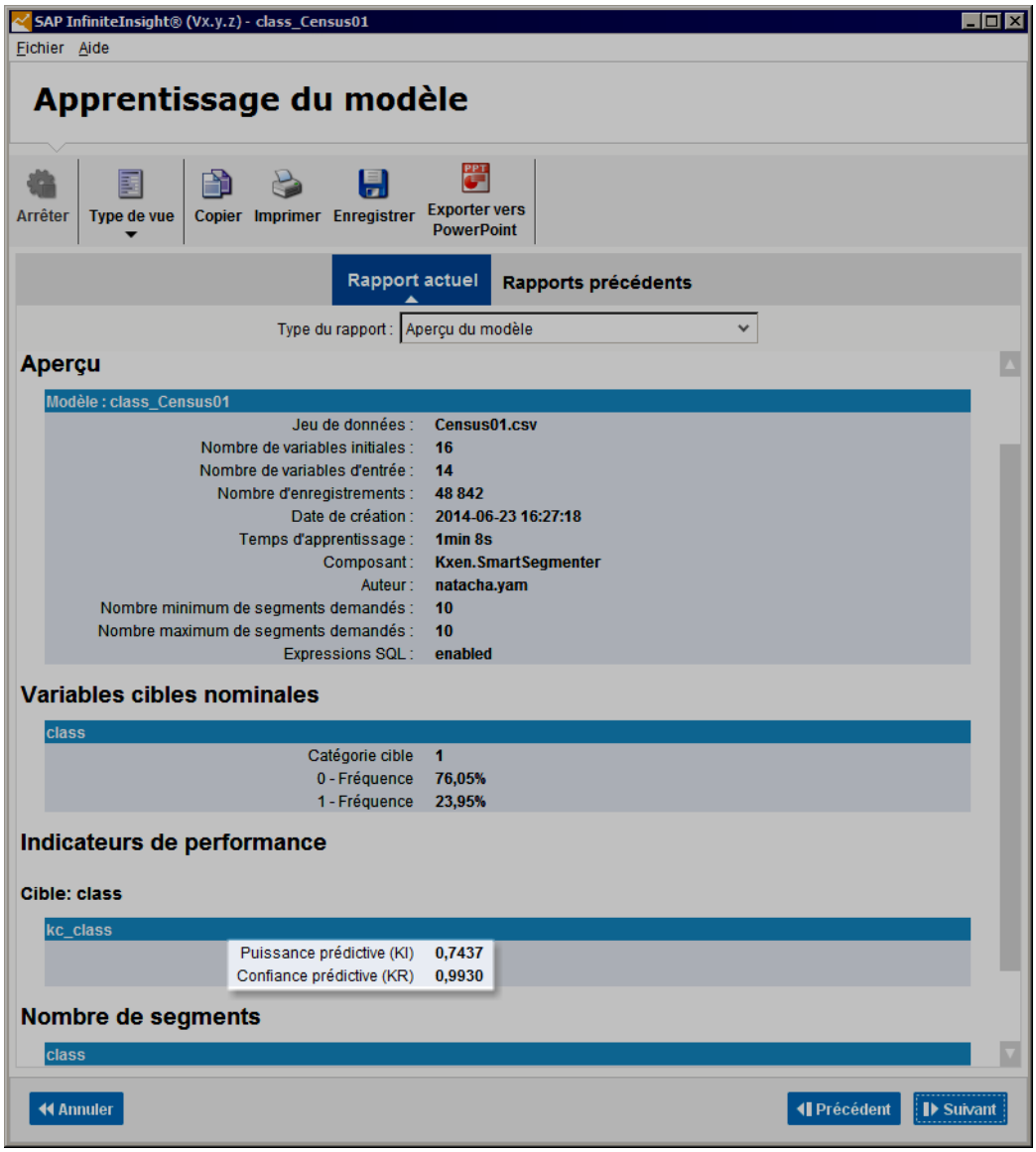
Pour ce scénario

Le modèle généré possède :

- une capacité prédictive égale à 0,7437,
- une reproductibilté égale à 0,9930.

Pour valider le modèle généré

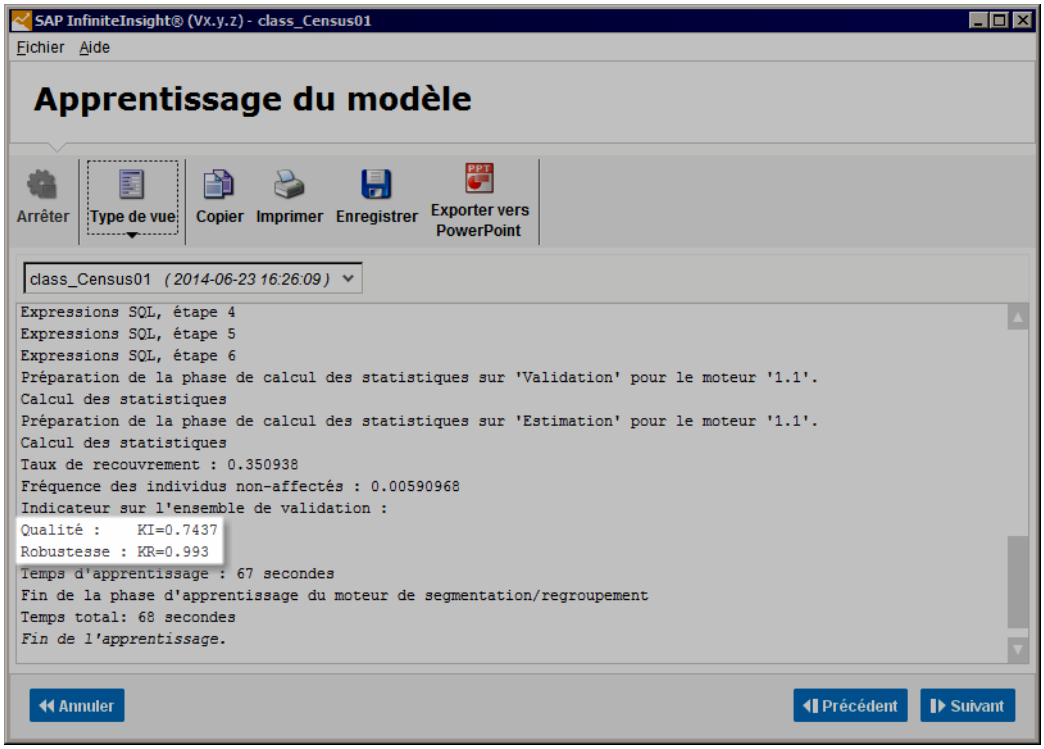
1 Vérifiez la *Capacité prédictive (KI)* et la *Confiance prédictive (KR)* du modèle. Ces indicateurs sont mis en évidence sur la figure suivante.



> **Remarque**
>
> A titre indicatif, d'autres indicateurs que la capacité prédictive (KI) et la reproductibilité (KR) sont indiqués lors de la génération du modèle. Vous pouvez par exemple visualiser le temps total requis pour générer le modèle (encadré en bleu dans la figure ci-dessus).

Vous pouvez également vérifier les indicateurs dans le journal détaillé du processus.

2 Cliquez sur *Type de vue*, puis sur  (*Détail des messages*). L'écran suivant s'affiche.



3 a) Si les performances du modèle vous conviennent, passez à l'étape 3 "Analyser et comprendre le modèle généré" (voir à la page 70)"
b) Sinon, passez à la procédure Pour générer un nouveau modèle (voir à la page 70).

☑ Pour générer un nouveau modèle

Vous avez deux options. Dans l'écran *Apprentissage du modèle*, vous pouvez :

- soit cliquer sur le bouton *Précédent* pour revenir sur les paramètres de modélisation initialement définis. Vous pouvez alors modifier les paramètres un à un.
- soit cliquer sur le bouton *Annuler* pour revenir à la page d'accueil de l'assistant de modélisation. Vous devez alors redéfinir tous les paramètres de modélisation.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

7.3 Etape 3 - Analyser et comprendre le modèle généré

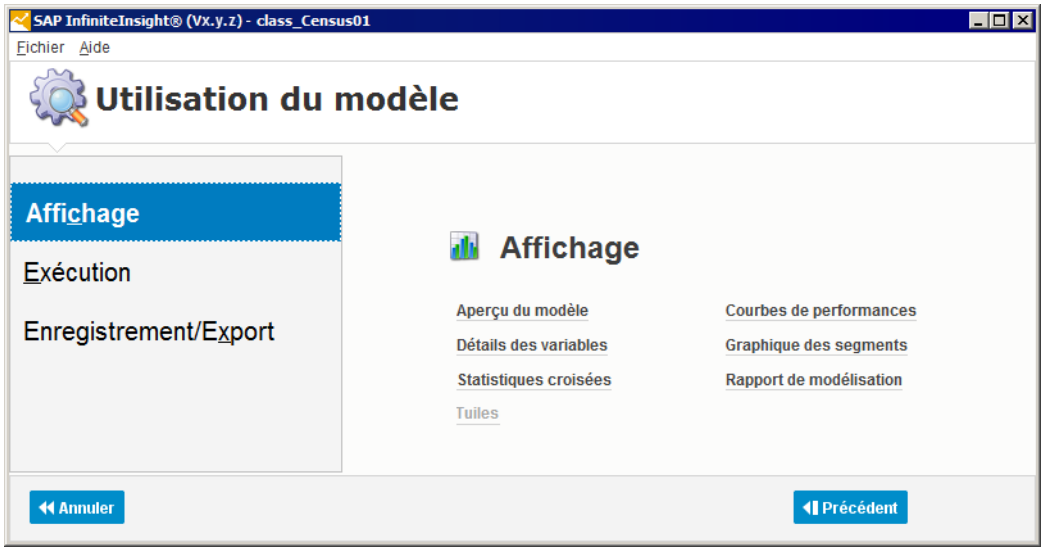
Un ensemble d'outils graphiques vous permet d'analyser le modèle généré et de connaître :

- la **performance du modèle** par rapport à un hypothétique modèle parfait et un modèle de type aléatoire,
- les **caractéristiques de chacun des segments**,
- l'**importance des différentes catégories de chaque variable d'un segment** par rapport à la variable cible (statistiques croisées).

Si vous avez choisi une variable cible pour votre modèle, la segmentation sera dite supervisée, c'est-à-dire que les segments seront créés en fonction de leur comportement vis-à-vis de la cible.

7.3.1 Menu d'utilisation

Une fois le modèle généré, cliquez sur le bouton *Suivant*. L'écran *Utilisation du modèle* apparaît.



L'écran *Utilisation du modèle* présente les différentes options d'utilisation du modèle, qui vous permettent :

- d'afficher les informations relatives au modèle généré, c'est-à-dire les graphiques des courbes de profit, la description détaillée des segments, les graphiques des segments et le profil des variables de chaque segment (groupe *Affichage*).
- d'appliquer le modèle généré sur de nouvelles données (groupe *Exécution*).
- d'enregistrer le modèle, l'exporter sous forme de script KxShell ou générer son code source dans un autre langage (groupe *Enregistrement/Export*).

7.3.2 Aperçu du modèle

L'*aperçu du modèle* reprend les informations récapitulée à la fin du processus de génération.

Ces informations sont détaillées dans les sections ci-dessous.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Aperçu

Modèle: <Nom>	Nom du modèle, créé à partir du nom de la variable cible et du nom du jeu de données
Jeu de données	Nom du jeu de données
Nombre de variables initiales	Nombre de variables explicatives dans le jeu de données
Nombre de variables d'entrée	Nombre de variables explicatives utilisées par le modèle
Nombre d'enregistrements	Nombre d'enregistrements du jeu de données
Date de création	Date et heure de la création du modèle
Temps d'apprentissage	Temps total pour l'apprentissage du modèle
Fonctionnalité	*Kxen.KMeans* (InfiniteInsight® Modeler / Segmentation)
Nombre de segments demandés	Nombre de segments demandés par l'utilisateur
Expressions SQL	Indique si le calcul des expressions SQL a été activé

Notifications

Variables Monotones Détectées	Indique si des variables monotones ont été trouvées dans le jeu de données, c'est-à-dire des variables dont le sens de variation est constant, dans l'ordre de lecture des données dans le jeu d'estimation.
Variables Suspectes Détectées	Ce rapport présente une liste de variables qui sont considérées comme suspectes. Ces variables suspectes ont un KI > 0.9, elles sont très fortement corrélées à la variable cible. Cela signifie que ces variables apportent probablement une information biaisée et qu'elles ne devraient pas être utilisées pour la modélisation. Une attention particulière doit être accordée à ces variables. Un rapport plus détaillé liste quelles variables particulères sont suspectes et dans quelle mesure (voir Rapports Statistiques > Compte Rendu Expert > Variables Suspectes).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Variables

Pour chaque cible nominale:

<Nom>	Nom de la variable cible
Catégorie cible	Valeur attendue de la variable cible
<catégories non-cible> - Fréquence	Pourcentage d'observations de la catégorie non-cible de la variable cible, dans le jeu de données d'estimation
<catégories cible> - Fréquence	Pourcentage d'observations de la catégorie cible de la variable cible, dans le jeu de données d'estimation

Pour chaque variable cible continue :

<Nom>	Nom de la variable cible
Min	Valeur minimale de la variable cible dans le jeu de données d'estimation
Max	Valeur maximale de la variable cible dans le jeu de données d'estimation
Moyenne	Moyenne de la variable cible pour le jeu de données d'estimation
Ecart type	Mesure de l'étendue de la dispersion des nombres autour de leur moyenne

Indicateurs de performance

Pour chaque variable cible :

Capacité prédictive (KI)	Indicateur de qualité qui correspond à la proportion d'information contenue dans la variable cible que les variables explicatives peuvent expliquer.
Confiance prédictive (KR)	Indicateur de robustesse qui précise la capacité du modèle à obtenir les mêmes performances lorsqu'il est appliqué à un nouveau jeu de données ayant les mêmes caractéristiques que le jeu de données d'apprentissage.

Nombre de segments

Pour chaque variable cible

<Nom>	nom de la variable cible
Nombre de segments demandés	Nombre de segments demandés par l'utilisateur
Nombre de segments trouvés	Nombre de segments trouvés par InfiniteInsight

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

7.3.3 Courbes de performances

Définition

Selon le type de cible, le graphique des **courbes de performances** (*model curve*) vous permet de :

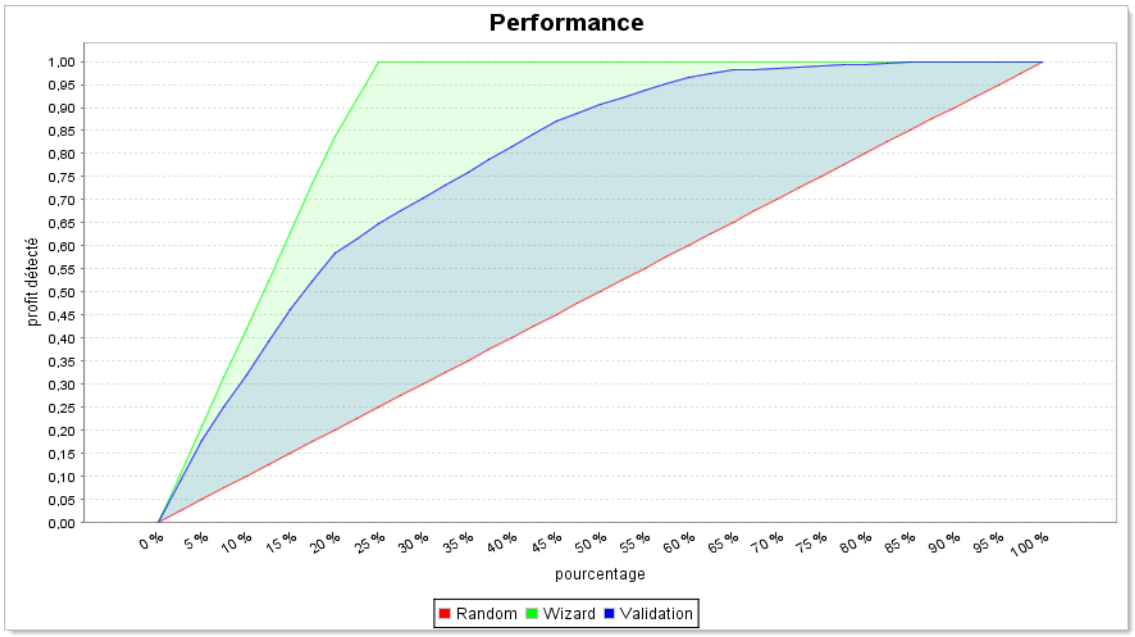
- visualiser le **profit réalisable** par rapport à votre problématique en utilisant le modèle généré lorsque la cible est nominale.
- **comparer les performances du modèle généré** à celles d'un modèle de type aléatoire et celles d'un modèle hypothétique parfait.

Sur le graphique, les courbes représentent le profit réalisable (axe des ordonnées) en fonction du taux d'observations sélectionnées sur la totalité du jeu de données initial (axe des abscisses). Les valeurs sur l'axe des abscisse sont regroupées par segment.

Afficher le graphique des courbes de performances

☑ Pour afficher le graphique des courbes de performances

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Courbes de performances*.
Les courbes de performances s'affichent.



Les paramètres par défaut affichent les courbes de performances correspondant au sous-jeu de *Validation*, à un hypothétique modèle parfait (*Wizard*) et à un modèle aléatoire (*Aléatoire*). Le type de profit utilisé est profit *Détecté*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

2 Lorsqu'il y a plus d'une variable cible, vous pouvez sélectionner la cible pour laquelle vous voulez voir les courbes de performance dans la liste *Modèles*.

> **Note**
>
> A chaque variable cible correspond un modèle. Le nom du modèle est basé sur le nom de la variable cible précédée du préfixe *kc_*.

3 Sélectionnez les options de visualisation qui vous intéressent.
Pour plus d'informations sur les options de visualisation, voir section suivante.

Options de visualisation

Pour un modèle à cible nominale

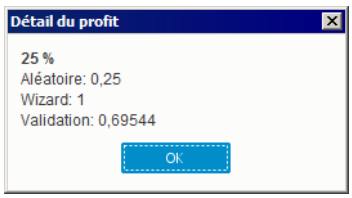
Sur le graphique des courbes de performances, différentes options vous permettent de visualiser :

- les **valeurs exactes** d'un point pour toutes les courbes représentées.
- les courbes de profit associées aux **sous-jeux d'estimation et de test**.
- les différentes courbes profit en fonction des **types de profit:**
  - *Détecté*,
  - *Lift*,
  - *Normalisé*,
  - *ROC*
  - *Lorenz 'Bon'* et *'Mauvais'*
  - *Densité 'Bon'*, *'Mauvais'* et *'Tous'*
  - *Personnalisé*.

Pour plus d'informations sur les courbes de profit (voir "Types de profit" à la page 46).

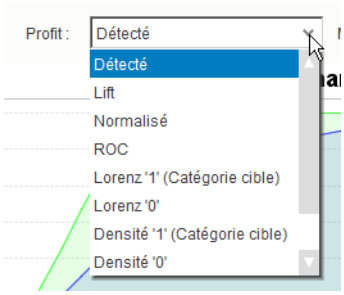
☑ Pour afficher les valeurs de profit exactes pour un point donné

Dans l' écran *Courbes de performances*, sur le graphique, cliquez sur un point de l'une des courbes représentées.
Par exemple, en cliquant sur un point de l'une des courbes ayant pour valeur en abscisse 25%, les valeurs de profit exactes apparaissent.



☑ Pour sélectionner un type de profit

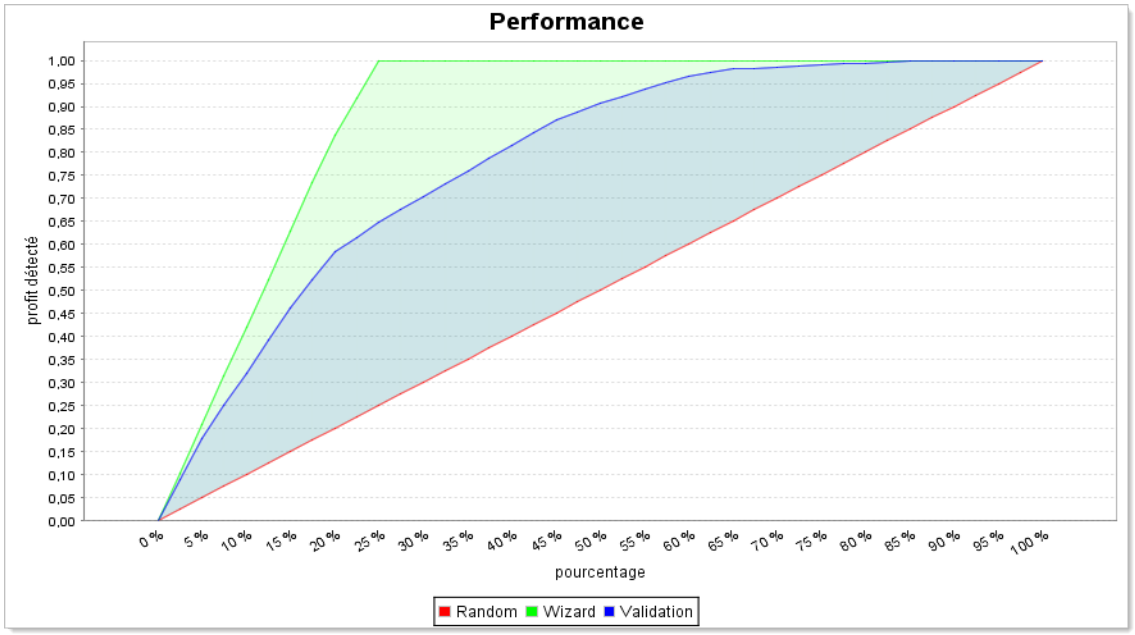
1 Dans l'écran *Courbes de performances*, au-dessus du graphique, cliquez sur la liste déroulante associée au champ *Profit*.
La liste des types de profit apparaît.



2 Sélectionnez un type de profit.
Les courbes correspondantes s'affichent.

Comprendre les courbes de profit

La figure ci-dessous représente le graphique des courbes de profit utilisant les paramètres par défaut.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Sur le graphique, les courbes représentent pour chaque type de modèle le profit réalisable (axe des ordonnées), c'est-à-dire le pourcentage d'observations appartenant à la catégorie cible de variable cible, en fonction du taux d'observations sélectionnées sur la totalité du jeu de données initial (axe des abscisses). Sur l'axe des abscisses, les observations sont ordonnées de manière décroissante en fonction de leur "score", c'est-à-dire par probabilité décroissante d'appartenir à la catégorie cible de variable cible.

Dans ce scénario d'utilisation, les courbes de profit représentent le taux de prospects susceptibles de répondre de manière positive à votre campagne marketing sur la totalité des prospects référencées dans votre base de données.

Le profit *Détecté* est le type de profit proposé par défaut. Avec ce type :

- la valeur "0" est affectée aux observations n'appartenant pas à la catégorie cible de la variable cible,
- la valeur "1/(fréquence de la variable cible dans le jeu de données)" est affectée aux observations appartenant à la catégorie cible de la variable cible.

Le tableau suivant décrit les trois courbes représentées sur le graphique utilisant les paramètres par défaut.

La courbe...	Représente...	Par exemple, en sélectionnant...
Wizard (courbe verte, la plus haute)	le profit réalisable en utilisant un hypothétique *modèle parfait*, permettant de *connaître de manière absolue* la valeur de la variable cible pour chaque observation du jeu de données	25% des observations sur la totalité de votre jeu de données à l'aide d'un modèle parfait, 100% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées. Le profit maximum est alors atteint. *Remarque* - Ces 25% correspondent au pourcentage de prospects ayant répondu de manière positive à votre campagne marketing, lors de votre phase de test. Pour ces prospects, la valeur de la variable cible, ou profit, est égale à 1.
Validation (courbe bleue, du milieu)	le profit réalisable en utilisant le *modèle généré par InfiniteInsight® Modeler / Segmentation*, permettant de prédire au mieux la valeur de la variable cible pour chaque observation du jeu de données	25% des observations de votre jeu de données initial à l'aide du modèle généré, 66,9% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées
Aléatoire (courbe rouge, la plus basse)	le profit réalisable en utilisant un *modèle aléatoire*, ne permettant de connaître en aucun cas la valeur de la variable cible pour chaque observation du jeu de données.	25% du jeu de données initial à l'aide d'un modèle aléatoire, 25% des observations appartenant à la catégorie cible de la variable cible sont sélectionnées

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Capacité prédictive, reproductibilité et courbes de profit

Sur le graphique des courbes de profit :

- du jeu de données d'estimation (graphique par défaut), la capacité prédictive correspond au rapport entre "la surface se trouvant entre la courbe du **modèle généré** et celle du **modèle aléatoire"** et **"**la surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire"**. Ainsi plus la courbe du modèle généré se rapproche de la courbe du modèle parfait, plus la capacité prédictive se rapproche de 1.
- des jeux de données d'estimation, de validation et de test (sélectionnez l'option correspondante dans la liste *Jeu de données*, située sous le graphique), la reproductibilité correspond à 1 moins le rapport entre la "surface se trouvant entre la courbe du **jeu d'estimation** et celle du **jeu de validation**" et la "surface se trouvant entre la courbe du **modèle parfait** et celle du **modèle aléatoire**".

7.3.4 Détails des variables

Définition

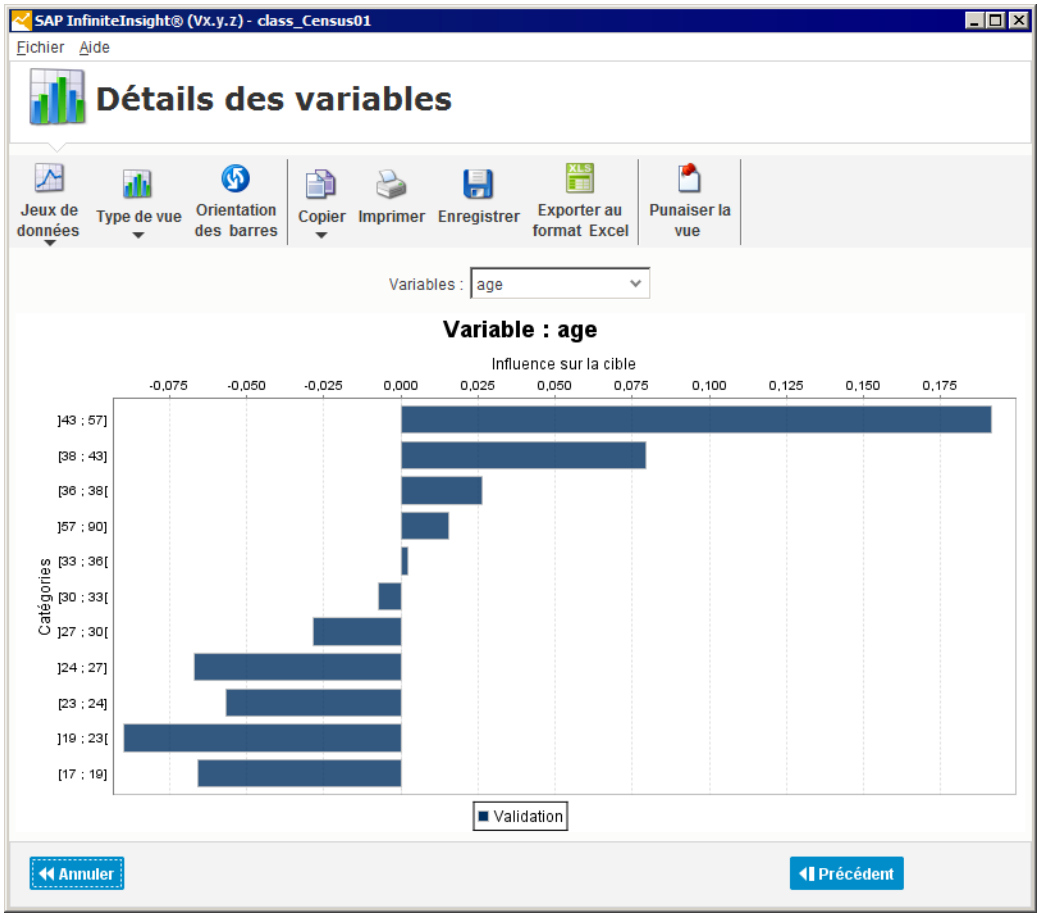
Le **graphique de détails de variable** présente l'importance des catégories d'une variable donnée par rapport à la variable cible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Afficher le graphique de détails d'une variable

☑ Pour afficher le graphique de détails d'une variable

1 Dans l'écran *Utilisation du modèle*, cliquez sur *Détails des variables*.
Le graphique de détails des variables apparaît.



2 Au-dessus du graphique, dans la liste *Variables*, sélectionnez la variable dont vous souhaitez afficher les catégories.

Si votre jeu de données contient des variables de type Date ou Datetime, des variables générées automatiquement apparaîtront dans cette liste. Pour plus d'information, reportez-vous à la section Variables de date : les variables générées automatiquement (voir "Variables de Date : les variables générées automatiquement" à la page 31).

Note

Vous pouvez afficher les détails d'une variable directement à partir du graphique *Contributions des variables*, en double-cliquant la barre représentant la variable qui vous intéresse.

Dans le cas où aucune structure utilisateur n'a été définie pour une variable continue, le graphe de détail des variables affiche les catégories créées automatiquement en utilisant le paramètre de *nombre de segments*. Le nombre de catégories affichées correspond à la valeur du paramètre de nombre de segments. Pour plus d'information au sujet de la configuration du paramètre de *nombre de segments*, reportez-vous à la section Nombre de segments pour les variables continues.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Options

En haut du panneau, une barre d'outils vous est proposée vous permettant de modifier l'affichage du graphique, de l'imprimer, copier ses données ou l'enregistrer.

Options d'affichage

☑ Pour afficher et masquer les sous-jeux d'Estimation et de Test

Cliquez sur *Jeux de données* et sélectionnez l'une des options suivantes :

-  *Tous les jeux de données*.
-  *Validation uniquement*.

☑ Pour afficher un histogramme

Cliquez sur *Type de vue* et sélectionnez  (*Histogramme*).
L'histogramme des catégories de la variable sélectionnée s'affiche.

☑ Pour afficher une courbe

Cliquez sur *Type de vue* et sélectionnez  (*Courbe de profit*).
La courbe de performances de la variable sélectionnée s'affiche.

☑ Pour ouvrir la vue courante dans une nouvelle fenêtre

Cliquez sur  (*Punaiser la vue*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Options d'utilisation

☑ Pour imprimer

1. Cliquez sur le bouton  (*Imprimer*).
   Une boîte de dialogue s'affiche vous permettant de choisir votre imprimante.
2. Sélectionnez l'imprimante et les options d'impression.
3. Cliquez sur *OK*.
   L'impression est lancée.

☑ Pour enregistrer

1. Cliquez sur le bouton  (*Enregistrer*).
   Une boîte de dialogue s'affiche vous permettant de choisir les propriétés du fichier.
2. Entrez un nom de fichier.
3. Choisissez le dossier de destination.
4. Cliquez sur *OK*.
   Le graphique est enregistré au format PNG dans le dossier sélectionné.

☑ Pour copier

1. Cliquez sur le bouton  (*Copier*) et sélectionnez l'option désirée.
   L'application copie les paramètres du graphique.
2. Collez les paramètres dans l'application de votre choix. Vous pouvez par exemple les utiliser pour générer un graphique dans un tableur (Excel, ...).

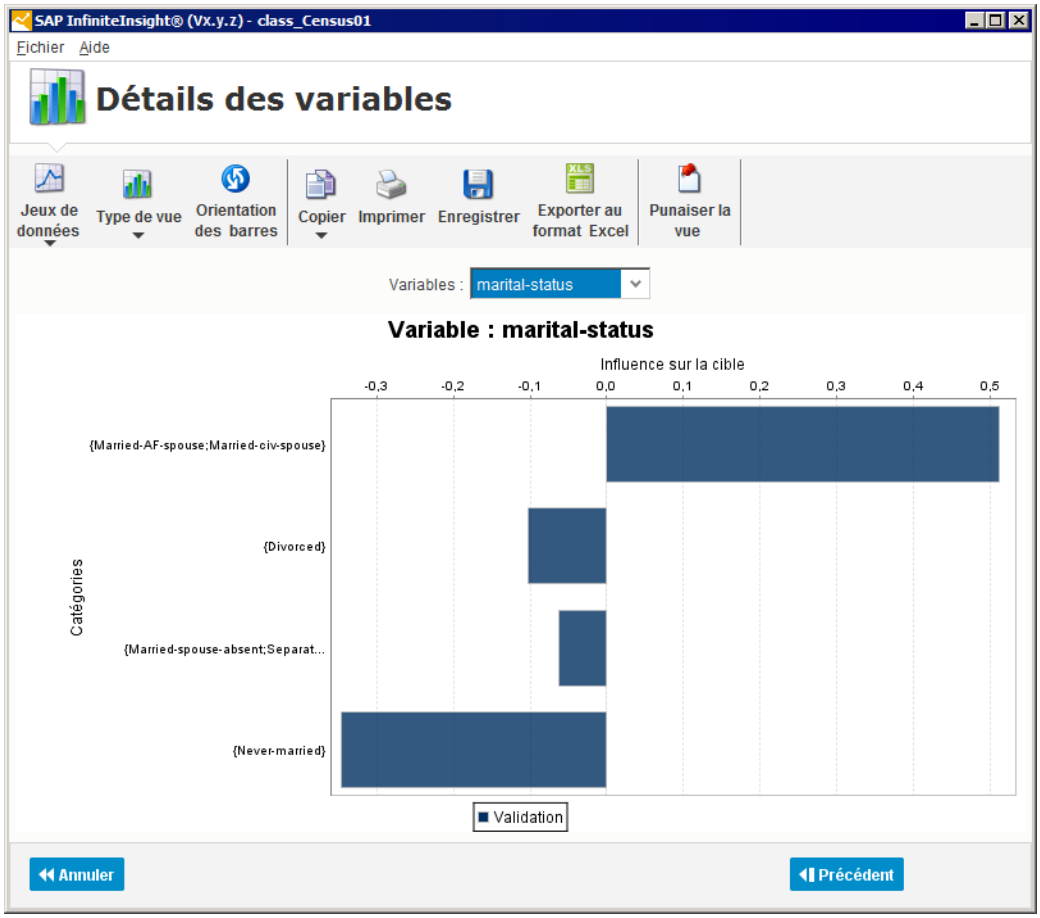
☑ Pour exporter au format Excel

Cliquez sur  (*Exporter au format Excel*).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Pour ce scénario

Sélectionnez la variable *marital-status*, qui est la variable explicative qui contribue le plus à la variable cible *Class*.



Ce graphique présente l'impact des catégories de la variable *marital-status* sur la variable cible.

7.3.5 Graphiques des segments

Il est possible d'afficher les différents types de graphiques suivants:

- Les graphiques à bulles

  Les graphiques à bulles affichent les segments en représentant la relation entre trois variables.
- Les histogrammes

  Les histogrammes permettent de visualiser en même temps les comportements de tous les segments vis à vis de la variable cible.

  Les trois graphiques suivants sont proposés :

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

- *Moyennes relatives de la cible*,
- *Fréquences*,
- *Moyennes de la cible*.

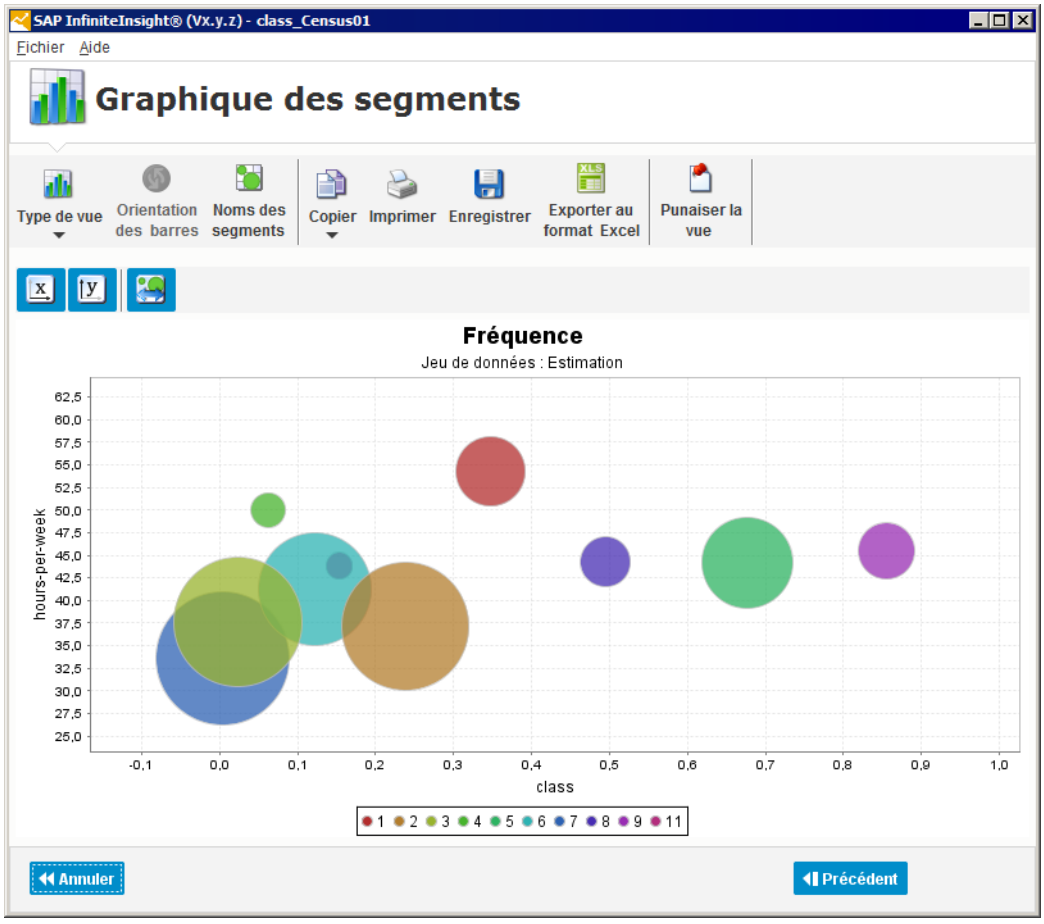
Ces trois graphiques vous permettent de visualiser :

- le pourcentage d'observations du jeu de données contenu dans chaque segment (graphique *Fréquences*),
- le pourcentage de chaque segment par rapport à la variable cible (graphiques *Moyennes de la cible* et *Moyennes relatives de la cible*).

Afficher les graphiques à bulles

☑ Pour afficher les graphiques à bulles

1 Sur l'écran *Utilisation du modèle*, cliquez *Graphique des segments*.
Le panneau *Graphique des segments* apparaît.



2 Utilisez les options pour définir les variables que vous souhaitez afficher sur le graphique à bulles.
Le tableau ci-dessous liste les options disponibles :

L'option...	vous permet...	À noter que...
	de sélectionner la variable à utiliser sur l'axe des abscisses.	Seules les variables numériques continues et nominales peuvent être utilisées.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

L'option...	vous permet...	À noter que...
	de sélectionner la variable à utiliser sur l'axe des ordonnées.	Seules les variables numériques continues et nominales peuvent être utilisées.
	de sélectionner la variable à utiliser pour la taille des bulles.	Seules la variable *Fréquence* et la variable cible peuvent être utilisées.
	d'afficher les noms des segments.	Les noms des segments peuvent être personnalisés dans *Statistiques croisées*.

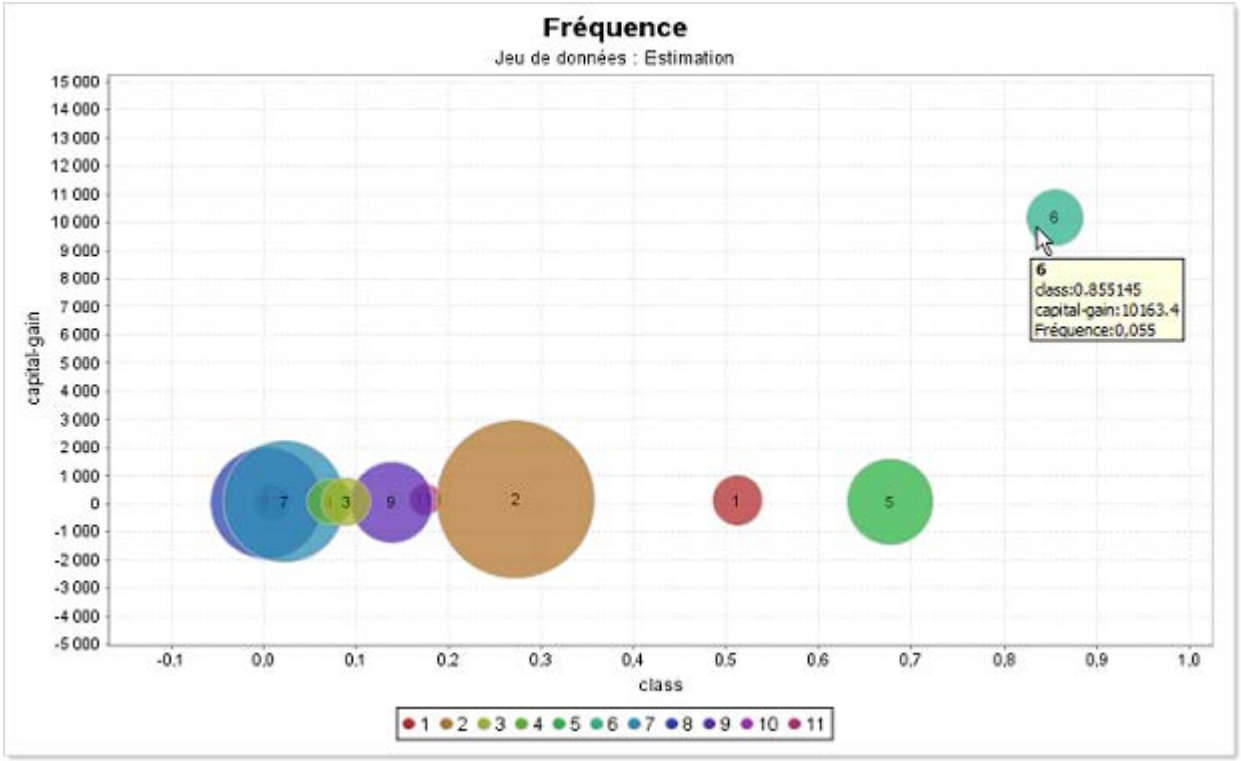
Comprendre les graphiques à bulles

Les graphiques à bulles vous permettent d'afficher les segments représentant la relation de trois variables. De ce fait, un graphique à bulles peut fournir trois types d'informations sur chaque segment.

De plus, les graphiques à bulles fournissent une représentation graphique de la segmentation, vous permettant de mieux visualiser les segments. Par exemple, cela peut être utile lors d'une présentation.

L'axe des abscisses, l'axe des ordonnées et la taille des bulles représentent chacun une variable. Vous pouvez choisir les variables à utiliser dans un graphique. De ce fait, vous pouvez créer un graphique à bulles qui sépare distinctement les segments l'un de l'autre, vous permettant ainsi d'identifier les segments intéressants pour votre campagne marketing.

La figure ci-dessous représente la relation entre les variables Fréquence, class et capital-gain.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Par exemple, les résultats démontrent que les clients du *segment 6* gagnent en moyenne 10 163,4 dollars par an (capital-gain: 10163,4) et représentent 5,5% (Fréquence: 0,055) de la population du jeu de données. De plus, 85,5% (class : 0,885) des clients du *segment 6* ont répondu de façon positive à la phase de test de votre campagne marketing.

En comparaison, le *segment 2* représente la plus grande population du jeu de données, soit 25,2% de la population (Fréquence 0,225), ce qui est environ cinq fois plus grand que la population du *segment 6*. Cependant, les clients du *segment 2* gagnent moins que les clients du *segment 6*, 147,542 dollars par an en moyenne (capital-gain : 147,542), soit 70% de moins que le *segment 6*. De plus, seulement 27,16% des clients du *segment 2* ont répondu de façon positive à la phase de test de votre campagne marketing.

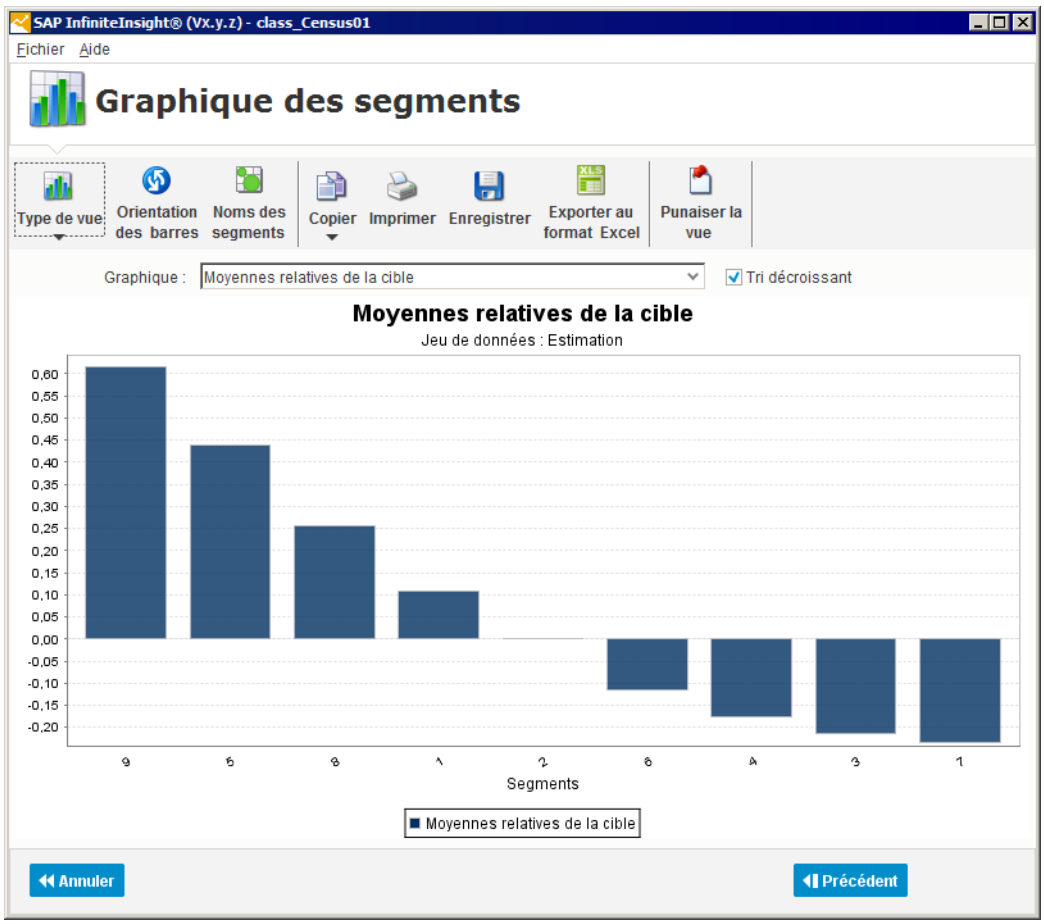
Par conséquence, comparé au *segment 2*, le *segment 6* est plus intéressant car il a montré de meilleurs résultats lors de la phase de test de votre campagne marketing.

Afficher les graphiques des segments

☑ Pour afficher les histogrammes

1 Dans l'écran *Utilisation du modèle*, cliquez sur *Graphique des segments*.
Le panneau *Graphique des segments* apparaît.

2 Cliquez sur  (*Type de vue)*, puis sélectionnez *Histogramme*.



CUSTOMER
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

3 Au-dessus du graphique, dans la liste déroulante associée au champ *Graphique*, sélectionnez le type de graphique que vous souhaitez afficher.

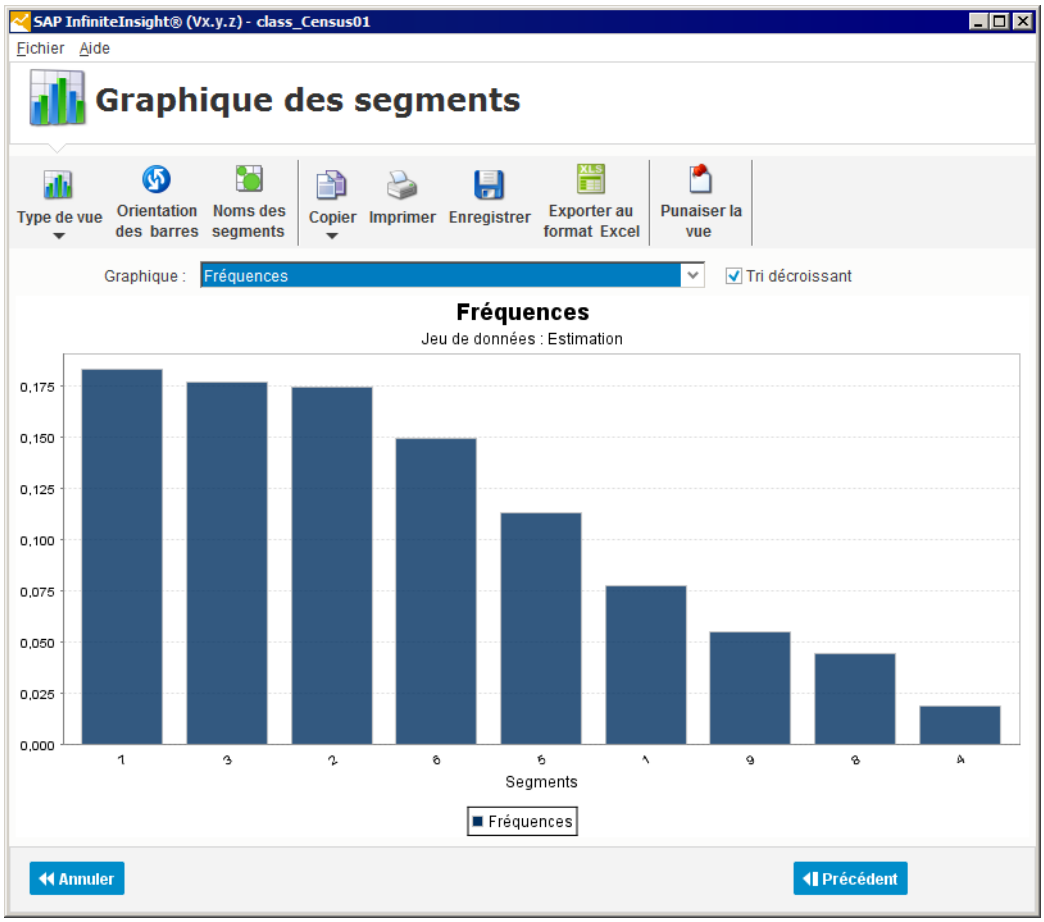
> **Remarque**
>
> Sélectionnez l'option *Tri décroissant* pour trier les barres des graphiques selon un ordre décroissant. Par exemple, sur le graphique *Moyennes relatives de la cible*, le tri décroissant permet de visualiser rapidement les segments les plus intéressants, c'est-à-dire les segments qui diffèrent le plus du comportement moyen sur l'ensemble du jeu de données.

Comprendre les graphiques des segments

Le graphique "Fréquences"

Le graphique *Fréquences* présente en pourcentage le nombre d'observations contenues dans chaque segment sur le nombre total d'observations contenues dans le jeu de données.

La figure ci-dessous présente le graphique *Fréquences* obtenu pour ce scénario. Les barres ont été triées par ordre décroissant.



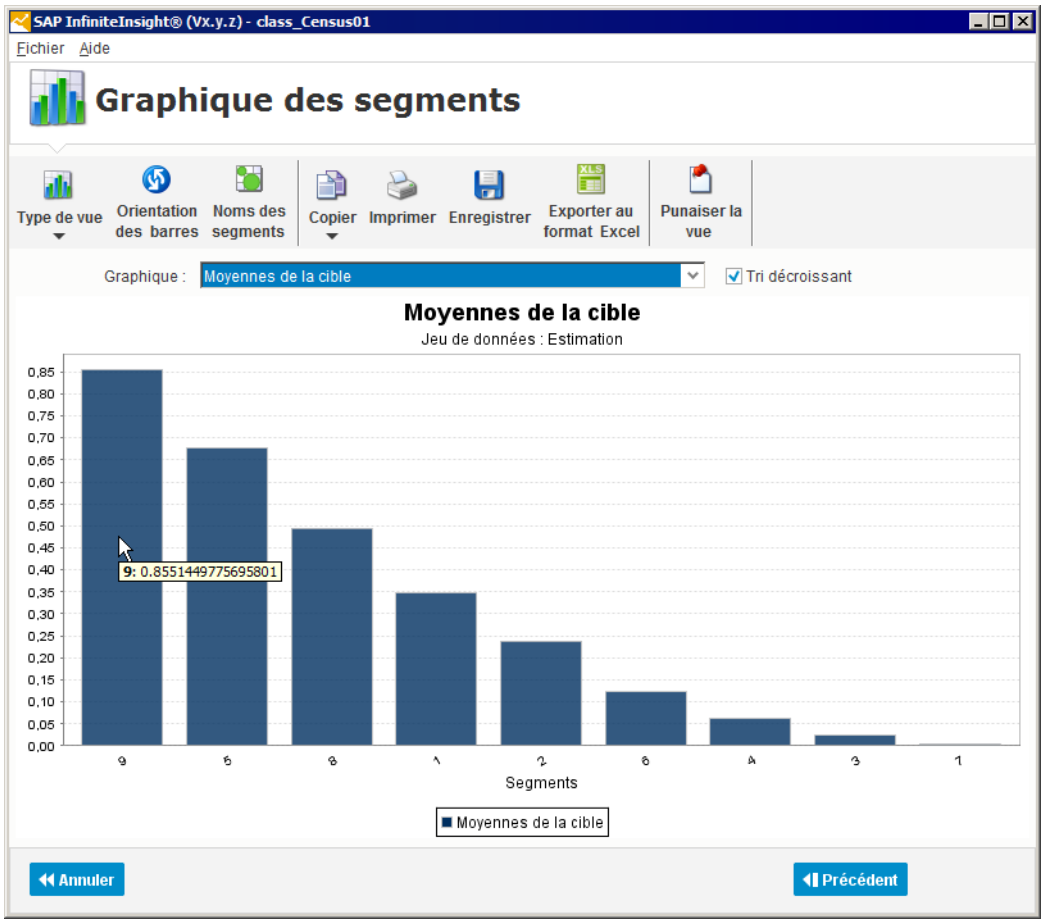
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Parmi les segments, le **segment 7** est le segment qui contient le plus grand nombre d'observations, soit 18% du nombre total de clients contenues dans le jeu de données.

Le graphique "Moyennes de la cible"

Le graphique *Moyennes de la cible* présente pour chaque segment le pourcentage d'observations appartenant à la catégorie cible de la variable cible.

La figure ci-dessous présente le graphique *Moyennes de la cible* obtenu pour ce scénario. Les barres ont été triées par ordre décroissant.



Parmi les segments, le **segment 9** est le segment qui contient le plus grand nombre d'observations appartenant à la catégorie cible. En effet, 85,5% des observations du segment 9 appartiennent à la catégorie *1* de la variable cible *Class*. Autrement dit, 85,5% des clients du segment 9 ont répondu de façon positive à la phase de test de votre campagne marketing.

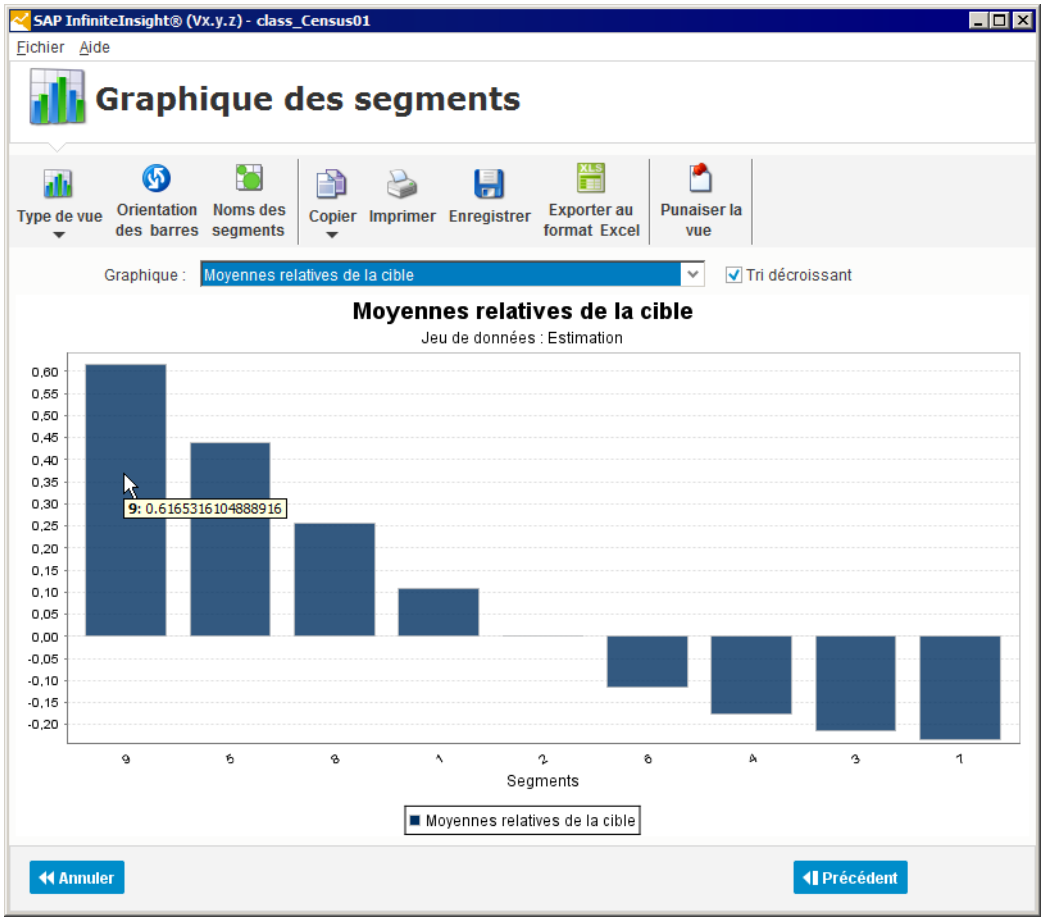
Le **segment 1** est le segment qui a la plus faible densité en observations appartenant à la catégorie cible. Moins de 1% des clients contenu dans ce segment ont répondu de manière positive à la phase de test de votre campagne marketing.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Le graphique "Moyennes relatives de la cible"

Comme le graphique *Moyennes de la cible*, le graphique *Moyennes relatives de la cible* présente pour chaque segment le pourcentage d'observations appartenant à la catégorie cible de la variable cible. Seule l'échelle adoptée pour l'axe des ordonnées diffère entre ces deux graphiques. Sur le graphique *Moyennes relatives de la cible*, le pourcentage d'observations appartenant à la catégorie cible de la variable cible sur l'ensemble du jeu de données est retranché. En d'autres mots, la valeur 0 de l'axe des ordonnées correspond au pourcentage d'observations appartenant à la catégorie cible de la variable cible sur l'ensemble du jeu de données.

La figure ci-dessous présente le graphique *Moyennes relatives de la cible* obtenu pour ce scénario. Les barres ont été triées par ordre décroissant.



Parmi les segments, le **segment 9** est le segment qui a la plus grosse proportion d'observations appartenant à la catégorie cible de la variable cible. Comparé au pourcentage d'observations appartenant à la catégorie cible sur la totalité du jeu de donnée, 61,6% des clients contenus dans le segment 9 appartiennent à la catégorie cible *1* variable cible *Class*.

Lorsqu'un segment contient près de 0% de clients appartenant à la catégorie cible, cela signifie que ce segment a quasiment la même densité en clients appartenant à la catégorie cible que le jeu de données pris dans sa totalité.

Le **segment 7** est le segment qui a la plus faible densité en observations appartenant à la catégorie cible. Comparé au pourcentage d'observations appartenant à la catégorie cible sur la totalité du jeu de donnée, -23,2% des clients contenu dans le segment appartiennent à la catégorie cible. Ce segment a donc une densité en clients appartenant à la catégorie cible plus faible que la densité du jeu de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

7.3.6 Statistiques croisées

Statistiques croisées et profils de variables

Les **statistiques croisées** permettent de visualiser pour chaque segment :

- le profil de chaque variable explicative par rapport à leur profil sur la totalité du jeu de données,
- l'expression SQL du segment si celles-ci ont été calculées.

Profil d'une variable

Le **profil d'une variable** indique la distribution des observations (appartenant à un segment ou au jeu de données global) dans les catégories de cette variable. En d'autres mots, le profil indique le pourcentage d'observations contenues dans chacune des catégories de la variable.

Exemple d'un profil de variable

La variable "sexe" d'un jeu de données peut être distribuée comme suit :

- 53% des observations appartiennent à la catégorie "homme",
- 47% des observations appartiennent à la catégorie "femme".

Cette distribution correspond au **profil** de la variable "sexe" sur le jeu de données.

Sur un segment A, issu de ce jeu de données, la même variable "sexe" peut être distribuée comme suit :

- 80% des observations appartiennent à la catégorie "homme",
- 20% des observations appartiennent à la catégorie "femme".

Cette distribution correspond au **profil** de la variable "sexe" sur le segment A.

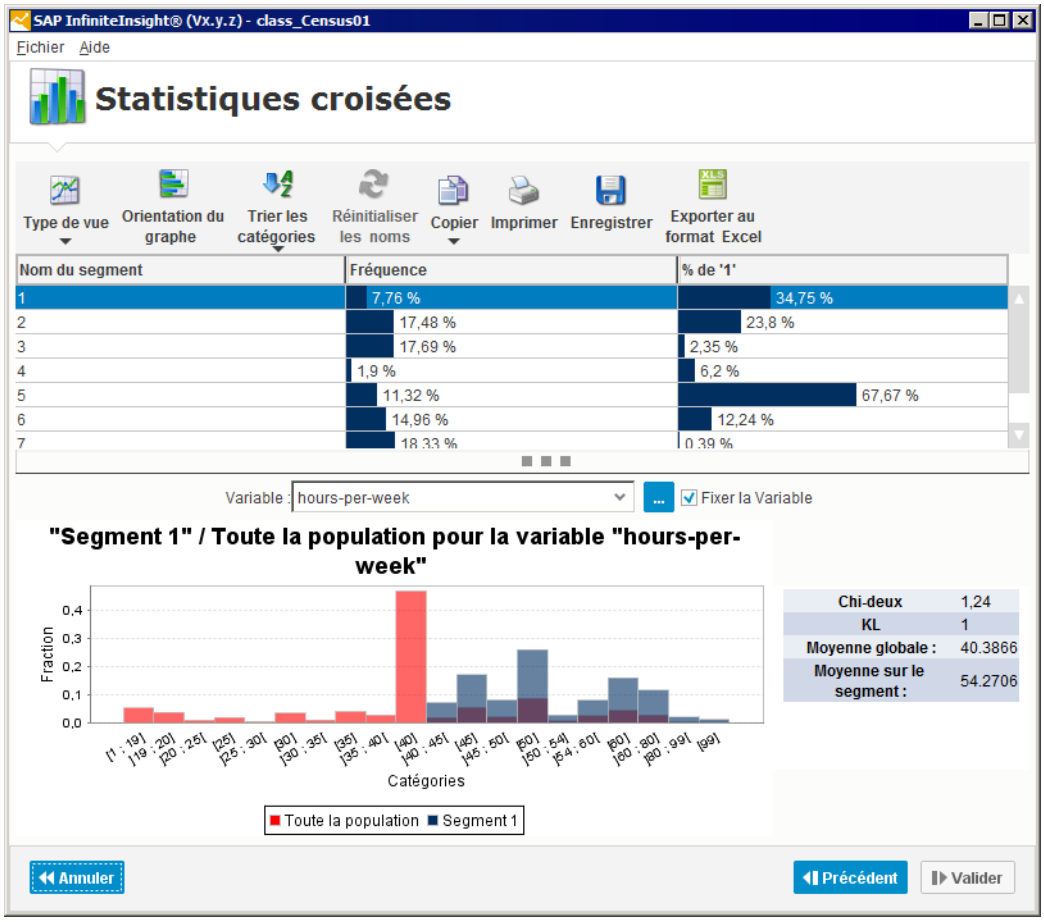
Les **statistiques croisées** permettent de visualiser et de comparer les profils de la variable "sexe" sur le jeu de données et sur les segments issus de ce jeu de données.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

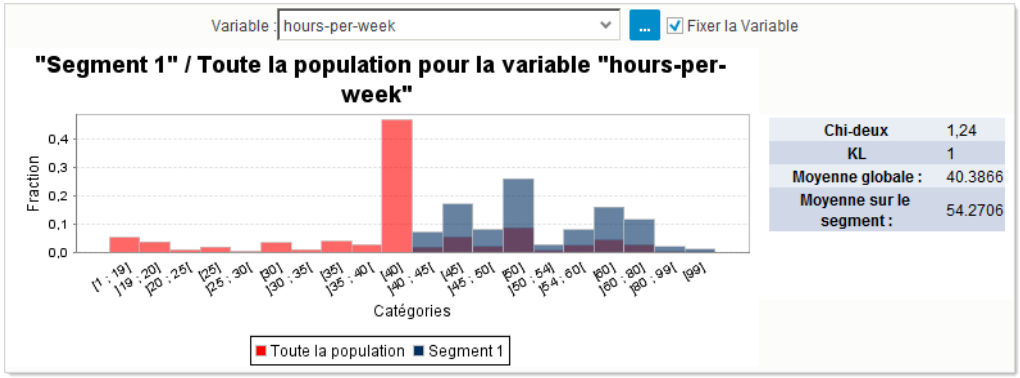
Afficher les statistiques croisées

☑ Pour afficher les statistiques croisées

1 Dans l'écran *Utilisation du modèle*, cliquez sur *Statistiques croisées* .
L'écran *Statistiques croisées* apparaît.



Par défaut, les statistiques croisées apparaissent sous forme de graphique, dans la partie inférieure de l'écran.



2 Dans le tableau, sélectionnez le segment dont vous souhaitez visualiser les statistiques croisées.

3 Dans la liste déroulante *Variable*, sélectionnez la variable dont vous souhaitez visualiser les statistiques croisées.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Comprendre les statistiques croisées

L'écran *Statistiques croisées* se décompose en trois parties :

- dans la **partie supérieure**, une liste déroulante vous permet de sélectionner la **variable** dont vous souhaitez visualiser les statistiques croisées. Les variables sont présentées par ordre décroissant en fonction de l'importance de leur contribution vis à vis de la catégorie cible de la variable cible. Quand un segment est sélectionné, les variables visibles dans la liste déroulante sont ordonnées selon la différence entre leur profil de segment et leur profile de population (on utilise la divergence de Kullback-Leibler comme mesure de cette différence). La variable apparaissant en premier dans la liste est la variable dont la différence de profils est la plus grande. Cette liste ordonnée de variables fournit l'ensemble des variables discriminantes pour décrire un segment.
- dans la **partie médiane**, un tableau présente chaque segment de manière synthétique. Il vous permet de sélectionner le **segment** dont vous souhaitez visualiser les statistiques croisées. Le tableau ci-dessous détaille le contenu du tableau synthétique :

La colonne...	Indique...	Par exemple...
Nom	le nom du segment	*Cluster 1*
Fréquence	la nombre d'observations contenues dans le segment sur le nombre total d'observations contenues dans le jeu de données	Les clients contenus dans le segment 1 représentent 7,76% du nombre total de clients contenus votre jeu de données d'apprentissage
% de '1'	la proportion d'observations contenues dans le segment appartenant à la catégorie cible de la variable cible	34,75% des clients contenues dans le segment 1 appartiennent à la catégorie cible de la variable cible *Class*. En d'autres mots, 34,75% des clients contenus dans ce segment ont répondu de manière positive à la phase de test de votre campagne marketing.

- dans la **partie inférieure**, un graphique présente soit les **statistiques croisées** correspondant au segment et à la variable sélectionnés, soit l'expression SQL définissant le segment, lorsqu'elle a été calculée.

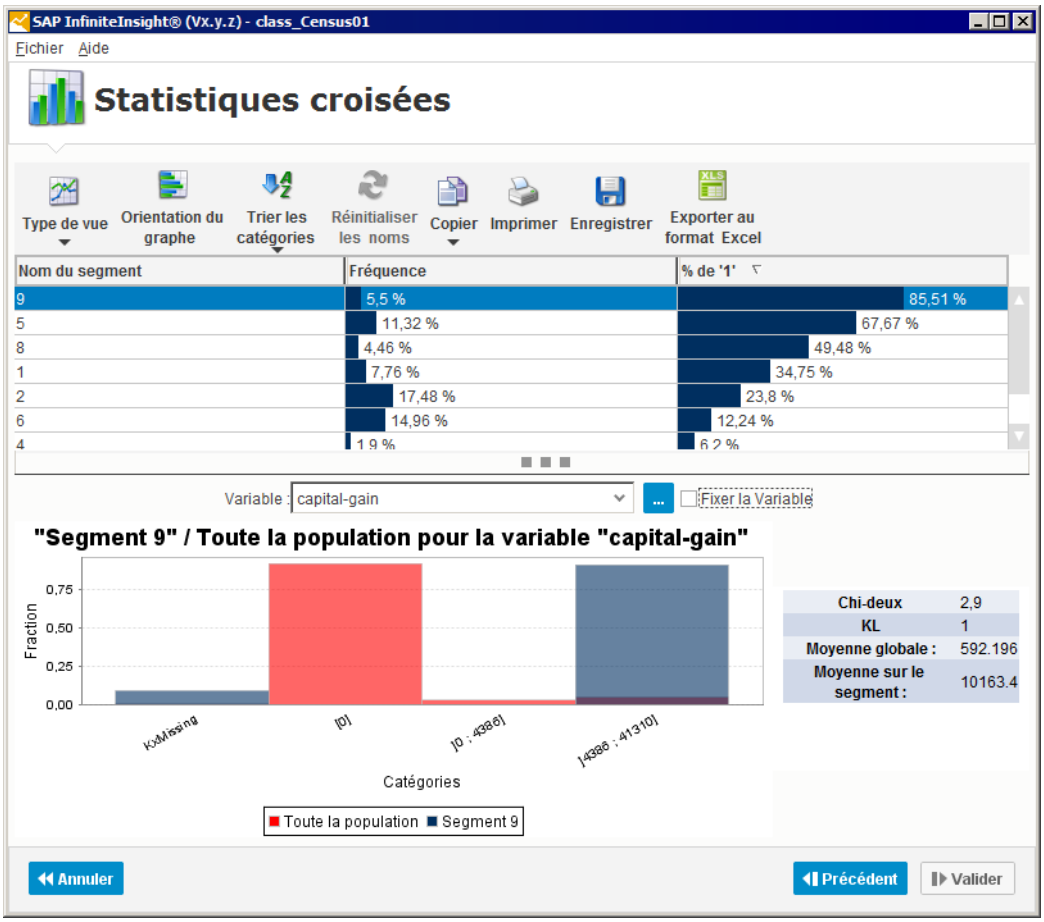
© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Les graphiques de statistiques croisées

Les **graphiques de statistiques croisées** présentent deux courbes :

- les **colonnes bleues** correspondent au profil de la variable sélectionnée sur le **segment sélectionné**,
- les **colonnes rouges** correspondent au profil de la variable sélectionnée sur la totalité du **jeu de données**.

La figure ci-dessous présente les *Statistiques croisées* obtenues pour ce scénario pour le *segment 9* et la variable *capital-gain* (gain boursier annuel).



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Dans la figure ci-dessus, le tableau permet d'identifier le *segment 9* comme le segment qui contient une des plus fortes densités d'observations appartenant à la catégorie cible de la variable cible. 85.51% des clients contenus dans ce segment appartiennent à la catégorie cible *1* de la variable cible *Class*.

Le graphique des statistiques croisées permet de visualiser et de comparer les profils de la variable capital-gain sur la totalité du **jeu de données** et sur le **segment 9**. Ces profils sont récapitulés dans le tableau ci-dessous.

Catégories de la variable "capital-gain"	Profil sur le jeu de données	Profil sur le segment 6
KxMissing	1%	9%
[0]	92%	0%
]0 ; 4386]	3%	0%
]4386, 41310]	5%	91%

La distribution des données sur la catégorie *]4386 ; 41310]* met clairement en évidence que la majorité des clients contenus dans le *segment 9* réalisent des gains boursiers annuels importants par rapport à l'ensemble des clients contenus dans le jeu de données. De plus, la distribution des données sur la catégorie *[0]* indique que la majorité des clients contenus dans le jeu de données, soit 92%, ne réalisent aucun gain boursier annuel, tandis qu'aucun des clients contenus dans le *segment 9* ne réalisent un gain boursier annuel nul.

En cochant la case *Fixer la variable*, vous pouvez comparer les profils de la variable capital-gain pour les différents segments.

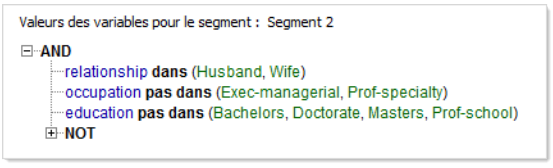
Afficher les expressions SQL

L'écran *Statistiques croisées* vous permet également d'afficher les expressions SQL correspondant à chaque segment.

> **Remarque**
>
> Les expressions SQL ne sont visibles que si vous avez coché la case *Calculer les expressions SQL* dans les paramètres spécifiques du modèle avant de le générer.

☑ Pour afficher l'expression SQL d'un segment

1. Sélectionnez le segment dans le tableau en haut de l'écran.
   Le graphique correspondant au segment s'affiche.
2. Cliquez sur  (*Type de vue*), puis sélectionnez  (*SQL*).
   L'expression SQL du segment s'affiche à la place du graphique.



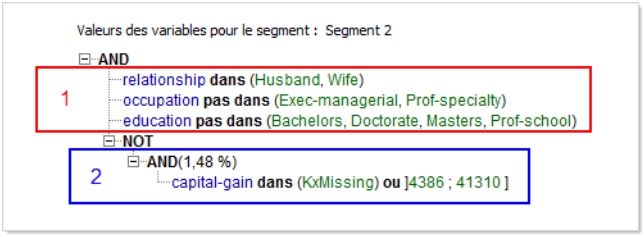
3. Cliquez sur + pour explorer la structure de l'expression SQL.
4. Cliquez sur  (*Type de vue*), puis sélectionnez  (*Mode comparaison*) pour retourner au graphique.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Comprendre les expressions SQL

- L'écran *Expressions SQL* se décompose en deux parties :
- dans la **partie supérieure**, un tableau présente chaque segment de manière synthétique. Il vous permet de sélectionner le **segment** dont vous souhaitez visualiser l'expression SQL.
- dans la **partie inférieure**, un arbre présente l'**expression SQL** correspondant au segment sélectionné.

La figure ci-dessous présente l'*expression SQL* du segment 2



L'expression SQL est structurée de la façon suivante :

- la première partie (notée **1** dans la figure ci-dessus) définit un ensemble d'observations dont les variables correspondent aux valeurs indiquées,
- la seconde partie (notée **2** dans la figure ci-dessus) définit des ensembles d'observations qui sont exclus de l'ensemble obtenu par la première partie de l'expression. Les pourcentages indiquent la proportion de chaque ensemble exclu par rapport à l'ensemble obtenu par la première partie de l'expression.
  Dans l'exemple ci-dessus on peut voir que le premier ensemble exclu correspond aux observations pour lesquelles la variable *capital-gain* est soit manquante (*KXMissing*) soit comprise entre 4386 exclu et 41310 (*]4386 ; 41310]*), ce qui représente 1,48% des observations obtenues par la première partie de l'expression.

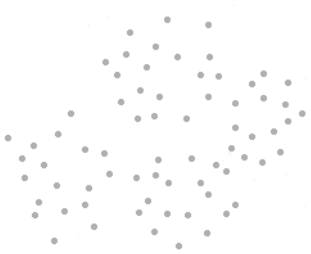
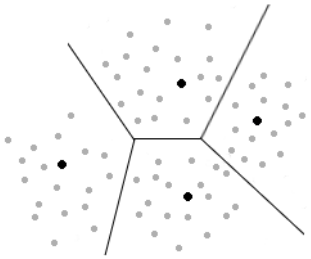
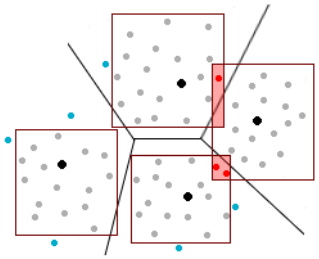
> **Note**
>
> Les segments sont créés en appliquant les expressions SQL dans un ordre précis défini par le moteur SAP InfiniteInsight®. Si vous appliquez les règles dans un ordre aléatoire, vous pouvez ne pas obtenir tout à fait les même résultats.

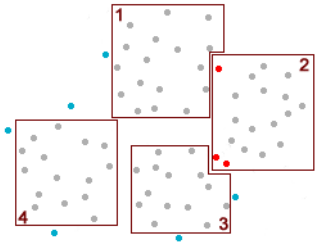
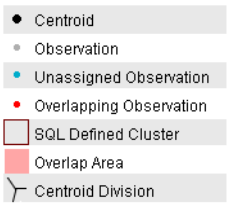
© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Différence entre statistiques croisées classiques et expressions SQL

La segmentation créée avec les expressions SQL est différente de celle créée sans. La raison d'être des expressions SQL est de proposer des segments faciles à comprendre et à appliquer. Les expressions SQL doivent se rapprocher au plus près des segments de base (c'est-à-dire ceux que vous obtenez sans calculer les expressions SQL). SQL peut être utilisé à la fois pour mieux comprendre les segments et pour les déployer sur la totalité de la base de données ou sur de nouvelles données (ce qui n'est généralement pas évident avec d'autres techniques).

L'utilisation de schémas permet de mieux comprendre la différence entre les segments créés à partir de centroïdes et ceux créés à partir d'expressions SQL.

Schéma	Explication
	Ce schéma représente un ensemble d'observations issues d'un jeu de données.
	Pour créer un segment, le moteur de InfiniteInsight® Modeler / Segmentation utilise l'approche centroïde. Les centroïdes sont le résultat d'un algorithme de segmentation, cela signifie qu'ils sont le barycentre de l'ensemble des points les plus proches. Quand on applique InfiniteInsight® Modeler / Segmentation sur ce jeu de données, les observations sont regroupées en fonction de leur distance par rapport à chaque centroïde. Ce schéma représente le jeu de données regroupé en quatre segments. C'est ce qu'on appelle le diagramme de Voronoï.
	Pour créer les expressions SQL qui définissent les segments, le moteur InfiniteInsight® Modeler / Segmentation utilise ce qu'on appelle la longueur minimum de description (Minimum Description Length ou MDL). Cela signifie que les segments initiaux créés à partir de l'approche centroïde sont retravaillés pour correspondre à l'expression la plus simple possible essayant ainsi de trouver le meilleur compromis entre la taille de l'expression et la perte d'information. Ce schéma représente les expressions SQL des segments (en rouge) comparées aux centroïdes. Sur ce schéma vous pouvez voir que : ▪ certaines observations qui se trouvaient dans un segment avec l'approche centroïde se retrouvent dans un autre quand on utilise les expressions SQL. ▪ d'autres observations ne peuvent être décrites par les expressions SQL et sont donc laissées hors des segments. C'est ce qu'on appelle les observations non . ▪ certaines observations peuvent être décrites par deux expressions SQL distinctes et donc apparaître dans deux segments différents.

	Ce schéma représente le résultat final obtenu avec les expressions SQL. Une observation ne peut pas apparaître dans deux segments différents, donc lorsque deux segments se recoupent, l'observation concernée est attribuée au premier segment créé. Le second segment auquel l'observation appartenait également est redéfini pour l'exclure. Vous pouvez voir que les observations qui apparaissaient dans deux segments sont conservées dans un seul. Le choix du segment dans lequel les observations seront conservées dépend de l'ordre dans lequel les règles SQL sont appliquées. Dans ce cas, la règle définissant le segment 2 a été appliquée avant celles définissant les segments 1 et 3.
	*Légende des schémas*

Comment choisir le type de segmentation le plus adapté ?

Grâce à la segmentation supervisée, InfiniteInsight® Modeler / Segmentation vous propose des indicateurs de performance (capacité prédictive et reproductibilité). Ils peuvent être utilisés pour comparer les deux types de segmentations (car le nombre de segments est identique). Si la capacité prédictive ne change pas de façon significative, la segmentation avec expressions SQL peut être préférable car plus facile à comprendre. En revanche, si la capacité prédictive baisse, il vaut mieux en rester à la segmentation de base.

La capacité prédictive n'est pas forcément ce que vous voulez optimiser pour une segmentation. Vous pouvez visualiser le profil cible de chaque segment dans l'interface graphique. Parmi les quatre segments, un ou deux peuvent être particulièrement intéressants. Dans ce cas, il vaut mieux se concentrer sur ces segments et étudier leur évolution lors de la génération des expressions SQL.

7.3.7 Rapport de modélisation

SAP InfiniteInsight® vous propose un ensemble de rapports vous permettant une analyse fine de votre modèle. Ces tables sont regroupées en plusieurs niveaux :

- les *statistiques descriptives*, qui fournissent des statistiques sur les variables, leurs catégories et les jeux de données ainsi que les statistiques croisées des variables par rapport aux variables cibles.

> **Note**
>
> Si votre jeu de données contient des variables de type Date ou Datetime, des variables générées automatiquement apparaîtront dans ces rapports. Pour plus d'information, reportez-vous à la section Variables de date : les variables générées automatiquement (voir "Variables de Date : les variables générées automatiquement" à la page 31).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

- les *performances du modèle*, dans lesquelles vous trouverez les indicateurs de performance du modèle, les individus non assignés, ainsi que les statistiques détaillées du score.
- le *détail des segments*, qui détaille pour chaque segment son KL, les fréquences et la moyenne de la cible, son expression SQL et ses statistiques.
- la *vérification des déviations*, qui vous permet de vérifier la présence de déviation pour chaque variable et catégorie de variable entre les jeux de données de validation et de test.
- les *rapports avancés*, dans lesquels vous trouverez d'autres indicateurs de performance, l'encodage des variables, ...

Options des rapports de modélisation

Une barre d'outils vous est proposée vous permettant de modifier l'affichage du rapport courant, de le copier, l'imprimer, le sauvegarder ou l'exporter sous format Excel.

Options d'affichage

Vue		Cette option permet d'afficher la vue courante du rapport dans un tableau graphique qui peut être triés par colonne.
		Cette option permet d'afficher la vue courante du rapport sous forme de tableau HTML.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme d'histogramme. Cet histogramme peut être trié par ordre ascendant ou descendant des valeurs ainsi que par ordre alphabétique ascendant ou descendant. Vous pouvez également choisir quelles données afficher.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme de secteurs.
		Pour certains rapports, vous pouvez choisir d'afficher la vue courante sous forme de courbe.
Trier		Quand le rapport en cours est affiché sous la forme d'un histogramme cette option vous permet de modifier son orientation (d'horizontal à vertical et inversement).
		Cette option vous permet d'afficher le rapport courant sans triage.
		Cette option vous permet de trier les valeurs du rapport courant par ordre ascendant.
		Cette option vous permet de trier les valeurs du rapport courant par ordre descendant.
		Cette option vous permet de trier les noms du rapport courant par ordre ascendant.
		Cette option vous permet de trier les noms du rapport courant par ordre descendant.
Séries	Cette option permet de sélectionner quelles informations afficher dans le rapport courant.	

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Options d'utilisation

Copier		Cette option permet de copier les données de la vue courante du rapport affiché. Les informations ainsi copiées peuvent être collées dans un éditeur de texte, un tableur, un document de traitement de texte.
		Si le rapport courant contient plusieurs vues (pour différentes variables, différents jeux de données, etc.) Cette option permet de copier l'ensemble des vues pour ce rapport.
		Si le rapport en cours est affiché sous forme de graphique, cette option vous permet de le copier au format image et de le coller dans un éditeur de texte ou dans un logiciel graphique.
Imprimer	Cette option permet d'imprimer la vue courante du rapport sélectionné selon le mode d'affichage choisi (rapport HTML, graphique, ...).	
Exporter		Cette option permet d'enregistrer sous différents formats (texte, html, pdf, rtf) les données de la vue courante du rapport affiché.
		Cette option permet d'enregistrer sous différents formats (texte, html, pdf, rtf) les données de l'ensemble des vues du rapport affiché.
		Cette option, qui est disponible pour toutes les formes d'affichage, permet d'exporter la vue courante vers Excel (compatible avec Excel 2002, 2003, XP et 2007).
		Cette option vous permet de sauvegarder tous les rapports.
		Cette option vous permet de sauvegarder la personnalisation des rapports.

7.4 Etape 4 - Utiliser le modèle

Une fois généré, un modèle de segmentation peut être **enregistré** pour utilisation ultérieure.

Un modèle de segmentation peut être **appliqué sur de nouveaux jeux de données**. Le modèle vous permet alors de déterminer à quel segment appartiennent les observations décrites dans ces jeux de données.

Cette partie présente l'option *Appliquer le modèle sur un nouveau jeu de données* de InfiniteInsight® Modeler / Segmentation. Les autres options de déploiement des modèles de segmentation sont similaires à celles proposées pour les modèles générés avec la fonctionnalité InfiniteInsight® Modeler / Régression ou Classement. Pour plus d'informations sur ces options, voir :

- Enregistrer un modèle
- Ouvrir un modèle
- Générer le code source d'un modèle

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

7.4.1 Appliquer un modèle sur un nouveau jeu de données

Le modèle en cours d'utilisation peut être appliqué sur de nouveaux jeux de données. Le modèle permet alors de déterminer à quel segment appartiennent les observations décrites dans ces jeux de données.

Contrainte d'utilisation d'un modèle

Pour qu'un modèle puisse être appliqué sur un jeu de données, le format du jeu de données d'application doit être identique à celui du jeu de données d'apprentissage utilisé pour générer le modèle. La même variable cible doit notamment être contenue dans les deux jeux de données, même si ses valeurs ne sont pas renseignées dans le jeu de données d'application.

Types de résultats proposés

L'application d'un modèle sur un jeu de données permet d'obtenir trois types de résultats :

- le **numéro du segment** auquel appartient chaque observation.
- le **codage disjonctif des numéros de segments**, ce qui signifie que pour chaque segment, une variable booléenne est créée indiquant si l'observation en cours appartient à ce segment ou non. Pour une observation donnée, la valeur "1" est assignée à la variable correspondant au segment contenant l'observation, et la valeur "0" est assignée aux variables correspondant aux autres segments. Les noms des variables sont générés selon la syntaxe suivante:
  `kx_<Nom de la Cible>_<Index du segment>`

  Prenons comme exemple un modèle à cinq segments. Lorsque vous appliquez ce modèle, SAP InfiniteInsight® crée cinq variables correspondant aux cinq segments générés. Pour une observation appartenant au segment 3, le résultat est le suivant :

KxIndex	class	kc_class	kc_class_1	kc_class_2	kc_class_3	kc_class_4
15	1	3	0	0	1	0

- la **moyenne de la cible pour chaque segment**, c'est-à-dire le pourcentage d'observations appartenant à la catégorie cible de la variable cible que contient chaque segment.

En fonction du niveau d'information souhaité, vous pouvez choisir de générer :

- uniquement le numéro de segment auquel appartient chaque observation (option *valeur prévue*).
- le numéro de segment et le codage disjonctif des numéros de segments (option *Codage disjonctif des numéros de segments*). Vous pouvez également décider d'inclure dans le fichier de résultats obtenu les variables contenues dans le jeu de données d'application (option *Codage disjonctif et recopie des var. explicatives*).
- le numéro de segment et la moyenne de la cible pour chaque segment (option *Moyenne de la cible pour les segments*).

Pour ce scénario

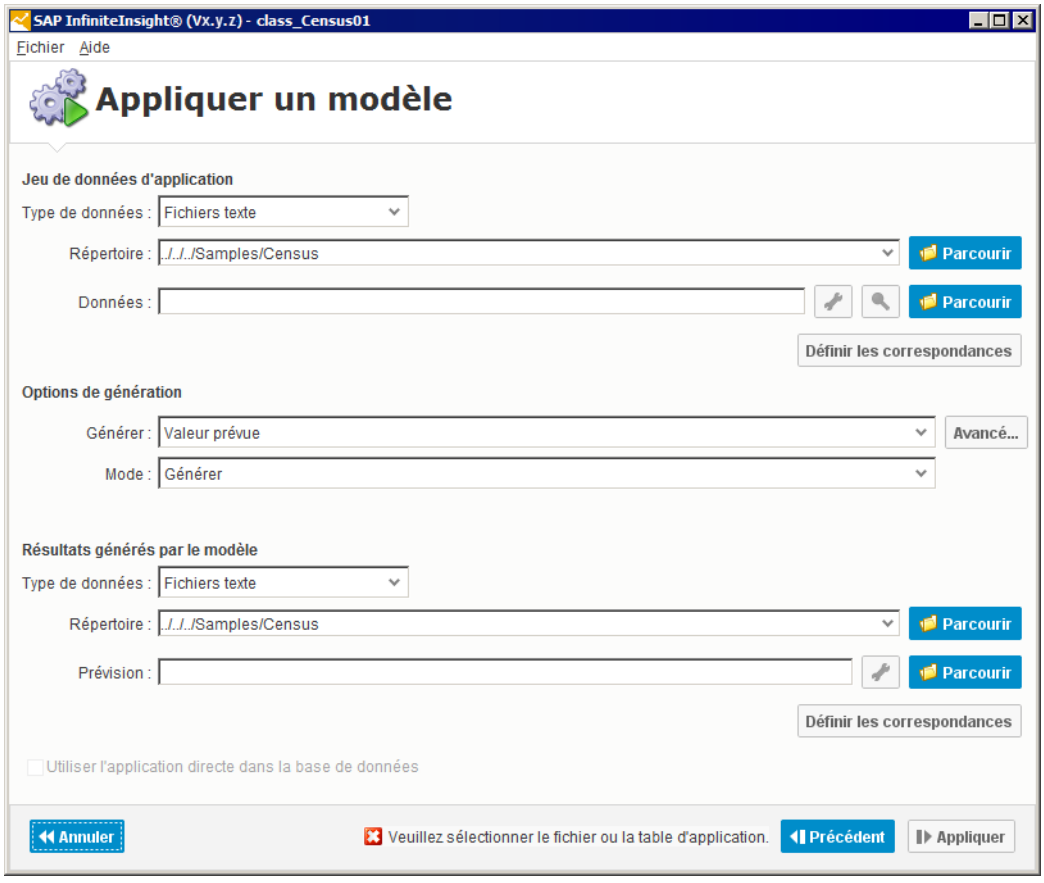
Vous allez appliquer le modèle sur le fichier *Census01.csv*, que vous avez utilisé pour générer le modèle.

Dans la procédure *Pour appliquer le modèle sur un nouveau jeu de données* :

- sélectionnez le format *Fichiers texte*,
- dans le champ *Générer*, sélectionnez l'option *Moyenne de la cible pour les segments*,
- sélectionnez un répertoire de votre choix pour enregistrer le fichier de résultats (*Résultats générés par le modèle*).

Pour appliquer le modèle sur un nouveau jeu de données

1 Dans l'écran *Utilisation du modèle*, cliquez sur l'option *Application du modèle*.
L'écran *Appliquer un modèle* apparaît.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

2 Dans la partie *Jeu de données d'application*, sélectionnez le format de la source de données dans la liste *Type de donnée*.

3 Cliquez sur les boutons *Parcourir* pour indiquer respectivement :
- dans le champ *Répertoire*, le répertoire dans lequel est stocké votre jeu de données,
- dans le champ *Données*, le nom du fichier correspondant à votre jeu de données.

4 Dans le cadre *Options de génération*, sélectionnez dans la liste *Générer* le type de valeurs de sortie que vous souhaitez obtenir pour la variable cible.

5 Sélectionnez dans la liste *Mode*, le type de résultats voulu.

6 Dans le cadre *Résultats générés par le modèle*, sélectionnez le format du ficher de sortie

7 Cliquez sur le bouton *Appliquer*.
L'écran *Application du modèle* apparaît.
Une fois l'application du modèle terminée, le fichier de résultats de l'application est automatiquement enregistré à l'emplacement que vous avez défini sur l'écran *Appliquer le modèle*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Utiliser l'application directe dans la base de données

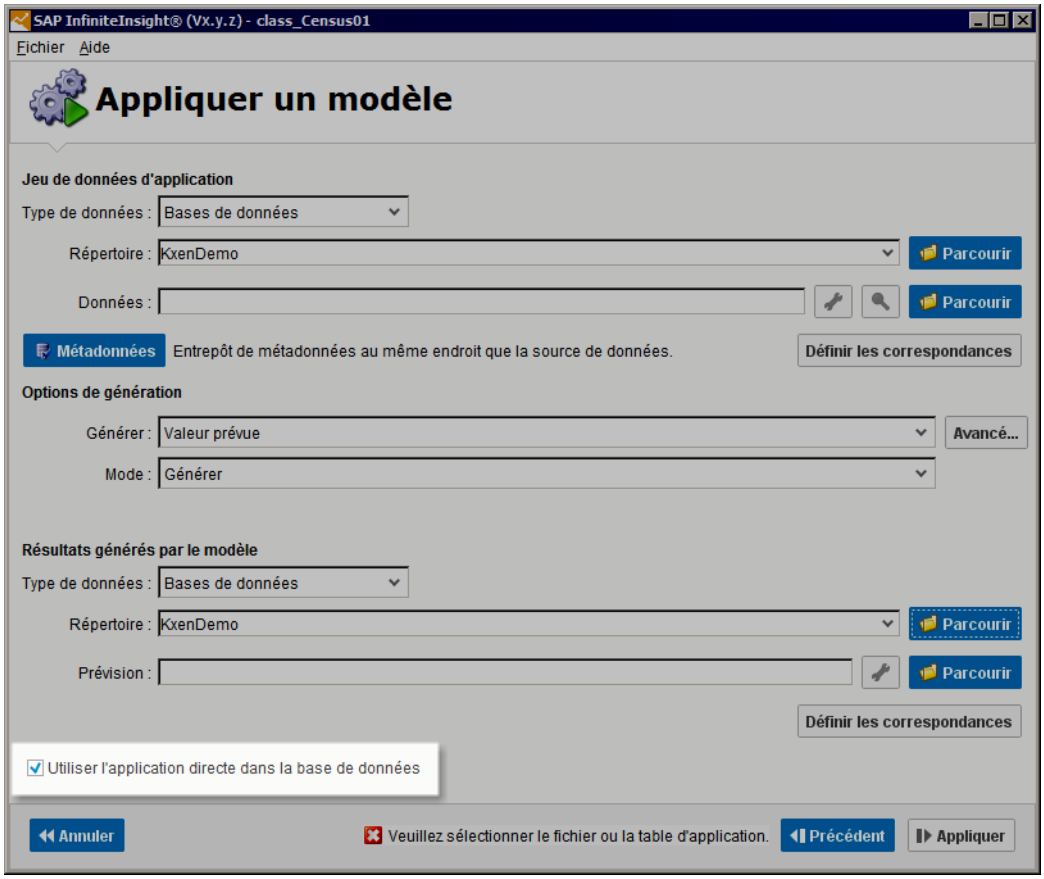
Pré-requis pour l'utilisation du mode d'application direct dans la base de données

Ce mode optimisé du score peut être utilisé si toutes les conditions suivantes sont remplies:

- le jeu de données d'application (table, vue, requête, manipulation de données) et les résultats du jeu de données sont des tables provenant de la même base de données,
- le modèle calculé contient au moins une variable avec une clé physique pré-définie dans SAP InfiniteInsight®,
- une licence InfiniteInsight® Scorer valide,
- aucune erreur apparue,
- un mode d'application dans la base de données activé,
- un accès de lecture et d'écriture (créer une table).

☑ Pour utiliser le mode d'application directe dans la base de données

Cochez l'option *Utiliser l'application directe dans la base de données*.



© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Paramètres avancés

Copier la variable de poids

Cette option vous permet d'ajouter au fichier de sortie la variable de poids si elle a été définie lors de la sélection des variables du modèle.

Copier les variables

Cette option vous permet d'ajouter au fichier de sortie une ou plusieurs variables du jeu de données.

☑ Pour ajouter toutes les variables du jeu de données

Cochez l'option *Toutes*.

☑ Pour sélectionner uniquement les variables qui vous intéressent

1. Sélectionnez l'option *Sélection*.
2. Cliquez sur le bouton >> pour afficher le tableau de sélection des variables.
3. Sélectionnez dans la liste *Eléments disponibles* les variables que vous voulez ajouter (utilisez la touche *Ctrl* pour sélectionner plusieurs variables à la fois).
4. Cliquez sur le bouton > pour ajouter les variables sélectionnées à la liste *Eléments sélectionnés*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Constantes définies par l'utilisateur

Cette option vous permet d'ajouter au fichier de sortie des constantes comme par exemple la date de l'application du modèle, le nom du jeu de données utilisé, ou toute autre information utile pour l'exploitation du fichier de sortie.

Une constante est définie par les informations suivantes:

Paramètre	Description	Valeur
Générer	indique si la constante sera générée dans le jeu de données de sortie.	*coché*: la constante sera générée *décoché*: la constante ne sera pas générée
Nom	nom de la constante	1 Le nom ne peut être identique à celui d'une variable du jeu de données de référence. 2 Si le nom est identique à celui d'une constante existante, celle-ci sera remplacée par la nouvelle constante.
Format	type de la constante	*number*: nombre *string*: chaîne de caractères *integer*: entier *date*: date *datetime*: date et heure
Valeur	valeur de la constante	format des dates: `YYYY-MM-DD` format des dates avec horaire: `YYYY-MM-DD HH:MM:SS`
Clé	spécifie si la constante est une variable clé ou un identifiant de l'enregistrement. Il est possible de déclarer des clés multiples qui seront construites selon l'ordre indiqué (1-2-3-...).	*0*: la constante n'est pas un identifiant *1*: identifiant primaire *2*: identifiant secondaire ...

☑ Pour définir une constante

1. Cliquez sur le bouton *Ajouter*. Une fenêtre s'ouvre vous permettant de saisir les paramètres de la constante.
2. Dans le champ *Nom*, saisissez le nom de la constante.
3. Dans la liste *Format de sortie*, sélectionnez son type.
4. Dans le champ *Valeur de sortie*, saisissez la valeur que vous souhaitez donner à la constante.
5. Cliquez sur le bouton *OK* pour valider la création de la constante. La nouvelle constante apparaît dans la liste. Vous pouvez choisir de générer ou non les constantes définies en cochant la case *Générer* correspondante.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Sorties par rang de segment

Segments par ordre de proximité

Cette option vous permet d'ajouter au fichier de sortie les numéros des segments dont le centroïde est le plus proche de l'observation en cours. Le segment dont le centroïde est le plus proche est celui auquel appartient l'observation, son numéro apparaît dans le fichier de sortie dans la colonne *kc_<Variable cible>*. Le segment suivant apparaît dans la colonne *kc_<Variable cible>_2*, et ainsi de suite en terminant par le segment dont le centroïde est le plus éloigné. Vous pouvez choisir d'ajouter tous les segments, ou seulement les plus proches.

☑ Pour ajouter tous les segments

Cochez l'option *Tous*.

☑ Pour ajouter les segments les plus proches

1. Cochez l'option *Les plus proches*.
2. Saisissez dans le champ texte le nombre de segments à ajouter (c'est-à-dire les deux, trois ou quatre premiers par exemple).

Noms des segments par ordre de proximité

Cette option vous permet d'ajouter au fichier de sortie les noms des segments dont les centroïdes sont les plus proches de l'observation en cours. Le segment dont le centroïde est le plus proche est celui auquel appartient l'observation, son nom apparaît dans le fichier de sortie dans la colonne *kc_name_<Variable cible>*. Le segment suivant apparaît dans la colonne *kc_name_<Variable cible>_2*, et ainsi de suite en terminant par le segment dont le centroïde est le plus éloigné. Vous pouvez choisir d'ajouter tous les segments, ou seulement les plus proches.

☑ Pour ajouter tous les segments

Cochez l'option *Tous*.

☑ Pour ajouter les segments les plus proches

1. Cochez l'option *Les plus proches*.
2. Saisissez dans le champ texte le nombre de segments à ajouter (c'est-à-dire les deux, trois ou quatre premiers par exemple).

> **Note**
>
> Le nom par défaut d'un segment est son numéro. Vous pouvez modifier les noms des segments dans la colonne *Nom* du panneau *Statistiques croisées* accessible par le menu.

Distances par ordre croissant

Cette option vous permet d'ajouter au fichier de sortie les distances de chaque observation aux centroïdes des segments. La distance au centroïde le plus proche apparaît dans la colonne *kc_best_dist_<Variable cible>*, la distance du second centroïde le plus proche apparaît dans la colonne *kc_best_dist_<Variable cible>_2*, et ainsi de suite jusqu'au centroïde le plus éloigné de l'observation en cours. Vous pouvez ajouter les distances par rapport à tous les centroïdes ou seulement les plus courtes.

☑ Pour ajouter toutes les distances

Cochez l'option *Toutes*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

☑ Pour ajouter les distances les plus courtes

1. Cochez l'option *Les plus proches*.
2. Saisissez dans le champ texte le nombre de distances à ajouter (c'est-à-dire les deux, trois ou quatre premières par exemple).

> **Remarque**
>
> Lorsque le mode SQL est activé, la notion de segment le plus proche n'est pas pertinente. Si un enregistrement appartient à un segment, la distance vaut 0. Si un enregistrement n'appartient pas à un segment, la distance vaut 1.

Probabilité

Cette option vous permet d'ajouter au fichier de sortie les probabilités que l'observation en cours appartiennent à chacun des segments. La probabilité que l'observation appartienne au segment dont le centroïde est le plus proche apparaît dans la colonne *kc_best_proba_<Variable cible>*, cette probabilité est généralement la plus haute. La probabilité que l'observation appartienne au second segment le plus proche apparaît dans la colonne *kc_best_proba_<Variable cible>_2*, et ainsi de suite jusqu'au segment dont le centroïde est le plus éloigné. Vous pouvez ajouter toutes les probabilités ou seulement celles correspondant aux segments dont les centroïdes sont les plus proches.

☑ Pour ajouter toutes les probabilités

Cochez l'option *Toutes*.

☑ Pour ajouter les probabilités des segments les plus proches

1. Cochez l'option *Les meilleurs*.
2. Saisissez dans le champ texte le nombre de probabilités à ajouter (c'est-à-dire les deux, trois ou quatre meilleures par exemple).

> **Remarque**
>
> Lorsque le mode SQL est activé, la notion de segment le plus proche n'est pas pertinente. Si un enregistrement appartient à un segment, la probabilité vaut 1. Si un enregistrement n'appartient pas à un segment, la probabilité vaut 0.

Sorties par identifiant de segment

Distance aux segments

Cette option vous permet d'ajouter au fichier de sortie la distance de chaque observation par rapport aux différents segments. Les distances sont générées dans les colonnes *kc_dist_cluster_<Variable cible>_<Identifiant segment>*. Par exemple si la variable cible est **Age**, la distance au segment **1** apparaîtra dans la colonne *kc_dist_cluster_Age_1*.

☑ Pour ajouter les distances à tous les segments

Cochez l'option *Tous*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

☑ Pour sélectionner les distances à ajouter

1. Cochez l'option *Sélection*.
2. Cliquez sur le bouton >>. La liste des segments s'affiche.
3. Cochez les segments pour lesquels vous souhaitez avoir les distances.

> **Remarque**
>
> Lorsque le mode SQL est activé, la notion de segment le plus proche n'est pas pertinente. Si un enregistrement appartient à un segment, la distance vaut 0. Si un enregistrement n'appartient pas à un segment, la distance vaut 1.

Probabilité du segment

Cette option vous permet d'ajouter au fichier de sortie la probabilité de chaque observation d'appartenir aux différents segments. Les probabilités sont générées dans les colonnes *kc_proba_cluster_<Variable cible>_<Identifiant segment>*. Par exemple si la variable cible est **Age**, la probabilité que l'observation appartienne au segment **1** apparaîtra dans la colonne *kc_dist_cluster_Age_1*.

☑ Pour ajouter les probabilités pour tous les segments

Cochez l'option *Tous*.

☑ Pour sélectionner les probabilités à ajouter

1. Cochez l'option *Sélection*.
2. Cliquez sur le bouton >>. La liste des segments s'affiche.
3. Cochez les segments pour lesquels vous souhaitez avoir les distances.

> **Remarque**
>
> Lorsque le mode SQL est activé, la notion de segment le plus proche n'est pas pertinente. Si un enregistrement appartient à un segment, la probabilité vaut 1. Si un enregistrement n'appartient pas à un segment, la probabilité vaut 0.

Autres

Codage disjonctif de la valeur prévue

Une colonne est créée pour chaque segment et contient 0 ou 1 selon que l'observation appartient au segment correspondant. Les colonnes créées sont nommées kc_disj_<variable cible>_<id segment>. Par exemple, si votre modèle comporte cinq segments et que la variable cible s'appelle *Age*, les cinq colonnes suivantes seront créées : *kc_disj_age_1*, *kc_disj_age_2*, *kc_disj_age_3*, *kc_disj_age_4*, *kc_disj_age_5*.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

Valeur moyenne de la cible / Probabilité de la catégorie cible

Cette option vous permet d'ajouter au fichier de sortie :

- pour les variables cibles continues :
  - la valeur moyenne de la cible pour le segment contenant l'observation (affichée dans la colonne *kc_<VariableCible>_Mean*),
  - la différence entre la moyenne de la cible pour le segment et la valeur réelle de la variable cible pour l'observation courante si elle est disponible (affichée dans la colonne *kc_<VariableCible>_Error*). A noter que lorsque la valeur de la cible actuelle n'est pas disponible, celle-ci équivaut à 0 par défaut. Par conséquent, la différence entre la valeur de la cible actuelle et la valeur moyenne de la cible calculée équivaut à la même valeur, ce qui signifie que les colonnes *kc_<VariableCible>_Mean* et *kc_<VariableCible>_Error* affichent la même valeur.
- pour les variables cibles nominales :
  - la proportion de la catégorie cible de la variable cible dans le segment contenant l'observation (affichée dans la colonne *kc_<Variable cible>_Mean*).

Analyser les résultats de l'application

Pour ce scénario

Dans Microsoft Excel, ouvrez le fichier de résultats au format texte que vous avez obtenu suite à l'application du modèle sur le fichier *Census01.csv*.

Pour ouvrir le fichier de résultats de l'application d'un modèle

1. En fonction du format du fichier de résultats généré, utilisez *Microsoft Excel* ou toute autre application pour ouvrir ce fichier.
   La figure ci-dessous présente les premières et les colonnes du fichier de résultats obtenu pour le scénario.

	A	B	C	D
1	KxIndex	class	kc_clusterId	kc_TargetMeanClustId
2	1	0	3	0,017524
3	2	0	5	0,476401001
4	3	0	3	0,017524
5	4	0	2	0,237075999
6	5	0	5	0,476401001
7	6	0	4	0,308898985
8	7	0	3	0,017524
9	8	1	5	0,476401001
10	9	1	1	0,942696989
11	10	1	1	0,942696989
12	11	1	5	0,476401001
13	12	1	5	0,476401001
14	13	0	3	0,017524
15	14	0	4	0,308898985
16	15	1	2	0,237075999
17	16	0	2	0,237075999

2. Vous pouvez maintenant analyser les résultats obtenus et utiliser les résultats de vos analyses pour prendre les bonnes décisions.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-Créer un modèle de segmentation ou de regroupement avec InfiniteInsig

Description du fichier de résultats

En fonction des options que vous avez sélectionnées, le fichier de résultats contient une partie ou la totalité des informations suivantes, dans l'ordre dans lequel elles sont présentées ci-dessous :

- la **variable clé** définie lors de la description des données à l'étape de définition des paramètres de modélisation. Si votre jeu de données ne contenait pas de variable clé, alors la variable clé *KxIndex* a été automatiquement générée par SAP InfiniteInsight®.
- éventuellement la **variable cible** renseignée par des valeurs connues si celles-ci figuraient dans le jeu de données d'application, comme c'est le cas pour ce scénario.
- la variable *kc_clusterId*, qui indique le numéro du segment auquel appartient chaque observation.
- la variable *kc_TargetMeanClusterId*, qui indique le pourcentage d'observations appartenant à la catégorie cible de la variable cible que contient chaque segment.
- les variables correspondant à chaque segment, et indiquant le **codage disjonctif des numéros de segments**. Le nom de ces variables correspondent aux numéros des segments, préfixés par *kc_cluster_*, par exemple *kc_cluster_1* pour le segment 1.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

8 Glossaire

A

agrégation de données

Le processus de consolider des valeurs de données dans un plus petit nombre de valeurs. Par exemple, des données de ventes peuvent être relevées quotidiennement et puis additionnées pour une semaine.

analyse de réseaux sociaux

L'analyse de réseaux sociaux est utilisée pour identifier des communautés ainsi que pour connaître la propagation dans des graphes (adoption d'un produit, épidémiologie), l'évolution d'un graphe ou l'influence d'un individu dans une communauté.

antécédent

X est appelé l'antécédent de la règle. Il peut être constitué d'un Item ou d'un Itemset.

application directe en base de données (in-database application)

Le fait d'envoyer une requête d'application du modèle à une base de données. Cette requête SQL est alors traitée dans la base elle-même.

apprentissage

Un autre terme pour l'estimation des paramètres d'un modèle basée sur le jeu de données disponible.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

attribut

En calcul informatisé, un attribut est une spécification qui définit une propriété d'un objet, d'un élément ou d'un fichier.

AUC

La statistique AUC mesure la performance ou la capacité prédictive d'un modèle. Il s'agit de la surface sous la courbe ROC.

auto-sélection

L'auto-sélection de SAP InfiniteInsight® est une sélection automatisée d'attributs.

B

barre d'erreur

voir intervalle de prédiction

base de données

Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter l'exploitation (ajout, mise à jour, recherche de données).

bibliothèque de variables

La bibliothèque de variables permet de stocker les descriptions des variables que vous avez déjà utilisées afin de pouvoir les réutiliser automatiquement lors d'une description par analyse.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

borne inférieure

La borne inférieure est définie comme un élément de P qui est inférieur ou égal à tous les éléments de S.

borne supérieure

Une borne supérieure d'un sous-ensemble S d'un ensemble partiellement ordonné (P, ≤) est un élément de P qui est supérieur ou égal à tous les éléments de S.

C

carte de score

Cet écran montre les coefficients associés aux catégories de toutes les variables du modèle (uniquement dans le cas d'un modèle régressive (Segmentation)).

catégorie

Une catégorie est une des valeurs possibles d'une variable discrète. Une variable discrète est une variable nominale ou ordinale. Il s'agit de l'élément de base utilisé pour coder la variable et pour rassembler des statistiques descriptives.

catégorie cible

La catégorie cible est la valeur attendue de la cible.

centroïde

Point fictif à l'intérieur d'un polygone dont les coordonnées correspondent au centre de celui-ci.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

chunk (by chunk)

Nombre de lignes d'un tableau qui sont traitées comme paquet.

coefficient de détermination (R2)

rapport entre la variabilité des prédictions (somme des carrés expliqués) et la variabilité des données (somme des carrés totaux).

confiance

La confiance d'une règle est une mesure qui indique le pourcentage de sessions qui vérifient le conséquent parmi celles qui vérifient l'antécédent. Par exemple le nombre de sessions qui contiennent l'Item D parmi celles qui contiennent l'Itemset {A,B,C}.

conséquent

Y est appelé le conséquent d'une règle. Il est constitué d'un seul Item, par exemple Y peut être l'Item {D}.

contribution

L'importance relative de chaque variable dans un modèle créé

contributions intelligentes des variables

La contribution des variables dans un modèle en prenant en compte la corrélation de variable.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

corrélation

Il s'agit d'une mesure qui quantifie la fait que deux variables partagent la même information. Ceci peut être mesuré en prenant la variation relative de deux variables pour différentes entités. La statistique classique définit la corrélation linéaire pour calculer la mesure sur des variables continues. SAP InfiniteInsight® peut calculer les corrélation entre variables de type différent en regardant la corrélation des codes des deux variables par rapport à une cible.

D

délai d'expiration

Une période de temps définie après laquelle un événement spécifique a lieu, sauf si un autre événement spécifique a lieu avant.

détail des variables

La mesure de l'impact d'une catégorie sur la cible.

déviation

La déviation est la différence entre la valeur observée et la moyenne d'un intervalle ou d'un rapport de variable.

domaine

Voir enregistrement analytique. Le domaine comportemental est généralement obtenu par des agrégats d'entité ou par des tables de transaction.

E

écart-type

L'écart type mesure la dispersion d'une série de valeurs autour de leur moyenne.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

écart-type de l'erreur

dispersion des erreurs autour du résultat réel

échantillonnage

L'échantillonnage est la sélection d'une partie dans un tout : lorsqu'on ne peut pas saisir un événement dans son ensemble, il faut effectuer des mesures en nombre fini, afin de représenter l'événement.

éditeur de formule

Un panneau qui permet de créer des champs comme expressions complexes dans l'éditeur de jeux de données analytiques.

encodage

L'encodage consiste à mettre une séquence de caractères (lettres, chiffres, signes de ponctuation et certains symboles) dans un format spécialisé pour une transmission ou un stockage efficace.

enregistrement

Il s'agit de la structure de données de base pour appliquer l'analyse de données. On l'appelle aussi une ligne de tableau. Un enregistrement typique serait la structure qui contient toutes les informations pertinentes sur un client ou compte en particulier.

enregistrement analytique

Un enregistrement analytique est une vue logique de tous les attributs qui correspondent à une entité. Un enregistrement analytique peut être divisé en plusieurs domaines qui regroupent des attributs liés.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

entité

Une entité est un objet d'intérêt d'une tâche analytique : il peut s'agir d'un client, d'un produit ou d'un store.

Note

Dû à une contrainte technique, les entités doivent avoir un identifiant unique.

erreur absolue moyenne (L1)

moyenne arithmétique des valeurs absolues des écarts (distance Manhattan ou City block)

erreur maximale (LInf)

écart maximum (distance de Chebyshev)

erreur moyenne

moyenne arithmétique des écarts

erreur quadratique moyenne (L2)

racine carré de la moyenne arithmétique des carrés des écarts (l'importance des grosses erreurs est majorée) (distance Euclidienne)

F

faux positif

signaux incorrectement identifiés comme positifs

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

filtre numérique

En électronique, un filtre numérique est un élément qui effectue un filtrage à l'aide d'une succession d'opérations mathématiques sur un signal discret.

fluctuation

Une évolution du signal qui n'est ni stable ni cyclique (InfiniteInsight® Modeler / Séries temporelles).

G

graphe à bulles / graphe en bulles

Un graphe à bulles est une représentation spécifique dans InfiniteInsight® Modeler / Segmentation qui affiche les segments en bulles. Les coordonnées d'une bulle donnée sont les valeurs du centroïde du segment correspondant de deux variables continues au choix. La taille de la bulle est donnée par la fréquence du cluster correspondant.

I

index de GINI

L'index GINI est une mesure de la capacité prédictive d'un modèle qui repose sur la courbe de Lorenz. Il est proportionnel à la superficie entre la courbe aléatoire et la courbe du modèle.

indicateur de performance clé

Les indicateurs clé de performance (ICP), ou KPI (selon l'acronyme anglais), sont des indicateurs mesurables d'aide décisionnelle dont le but est de représenter un aperçu d'évolution des facteurs clés de succès des processus de l'entreprise afin d'évaluer sa performance globale en fonction des objectifs à atteindre.

indicateur de qualité : capacité prédictive

La capacité prédictive (KI) est l'indicateur de qualité des modèles générés par SAP InfiniteInsight®. Cet indicateur correspond au taux d'information contenu dans la variable cible que les variables explicatives permettent d'expliquer.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

indicateur de robustesse : reproductibilité

La reproductibilité (KR) est l'indicateur de robustesse des modèles générés par SAP InfiniteInsight® . Elle indique la capacité d'un modèle à conserver les mêmes performances dans le cas où il est appliqué à un nouveau jeu de données présentant les mêmes attributs que le jeu de données d'apprentissage.

installation avec plusieurs instances

Il s'agit d'un mode d'installation SAP InfiniteInsight® qui consiste à lancer plusieurs instances sur un seul serveur afin de répartir la charge.

intervalle de prédiction

Les valeurs extrêmes de l'intervalle de prédiction se calculent de la façon suivante : `{TargetMean - (sqrt(TargetVariance)); TargetMean + (sqrt(TargetVariance))}`

Item

Un composant d'une règle d'association.

itemset

Un ensemble d'Items est appelé un Itemset.

itération

Une itération est un seul passage d'un cycle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

J

jeu de données

Un jeu de données est une collection de donnée, habituellement représentée sous forme de tableau. Chaque colonne représente une variable et chaque ligne attribue une valeur pour chacune des variables.

jeu de données d'application

Un jeu de données d'application est un jeu de données sur lequel on applique un modèle et qui contient une variable cible dont on veut connaître la valeur.

jeu de données d'apprentissage

Un jeu de données d'apprentissage est un jeu de données utilisé pour la génération d'un modèle. En analysant le jeu de données d'apprentissage, les composants SAP InfiniteInsight® génèrent un modèle qui permet d'expliquer la variable cible, grâce aux variables explicatives.

jeu de données d'événements

Un jeu de données d'événement devrait comporter au moins :

- une date d'événement comme une date de naissance ou le début de l'essai dans le format `AAAA/MM/JJ`.
- un identifiant de référence dans deux colonnes (par exemple un identifiant de client) qui sera utilisé pour créer des nœuds et des liens et éventuellement pour joindre un jeu de données de décoration (jeu de données qui contient des informations complémentaires telles que des informations géo-démographiques).

K

KL (Kullback-Leibler)

La divergence Kullback-Leibler est utilisée pour mesurer la différence entre le profil de cluster et le profil de population des variables.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

L

Lift

Le Lift d'une règle est une mesure qui indique les chances de trouver le conséquent en utilisant l'antécédent comparé aux chances de trouver le conséquent au hasard. Une valeur supérieure à 1 indique que l'utilisation de l'antécédent augmente vos chances de trouver le conséquent.

M

MAPE globale sur l'horizon

Cet indicateur de performance pour le modèle de prévision est la moyenne des valeurs MAPE observées dans tout l'horizon d'apprentissage. Une valeur de zéro indique un modèle parfait tandis qu'une valeur supérieure à 1 indique un modèle de mauvaise qualité. Une MAPE globale sur l'horizon de 0.09 veut dire que le modèle prend en compte 91 % du signal, l'erreur de prévision est alors de 9 %.

matrice confusion

La matrice de confusion permet de visualiser les valeurs de la cible prédites par le modèle par rapport aux valeurs réelles et de fixer le score à partir duquel les observations seront considérées comme positives, c'est-à-dire pour lesquelles la valeur de la cible est celle recherchée.

métadonnées

les informations sur les données elles-mêmes

méta-opérateur

Des opérateurs qui sont utilisés sur d'autres opérateurs.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

modèle descriptif

Modèle, qui permet de décrire des jeux de données

modèle explicatif

Modèle, qui permet de prédire et d'expliquer des phénomènes

monotonicité

Le sens de variation de fonctions monotones ne change pas.

moteur

La partie d'une application qui est indépendant de l'interface et qui fait la manipulation de données et d'autres opérations de base indépendamment de la façon que ceci est présenté à l'utilisateur.

moyenne

La valeur moyenne arithmétique d'une collection de données numériques.

moyenne des valeurs absolues de l'erreur relative (MAPE)

Indicateur qui donne la qualité de l'adéquation entre le signal et le modèle sur la l'ensemble des prévisions. Sa valeur est un pourcentage indiquant le niveau d'erreur (0% indique un modèle parfait).

N

normaliser

Transformer des valeurs numériques et les ramener dans un intervalle prédéfini.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

O

observation non assignée

Lorsque Segmentation crée les expressions SQL des segments, certaines observations ne peuvent être décrites par les expressions SQL et sont donc laissées hors des segments. C'est ce qu'on appelle les observations non assignées.

P

pivot

Un pivot est (tableau croisé dynamique) une fonctionnalité de certains tableurs qui permet de générer une synthèse d'une table de données brutes. Cela permet de créer des mises en forme de tableaux en choisissant les différents champs voulus, en abscisses ou en ordonnées.

polynôme

Un polynôme peut être de degré 1, 2, 3 ou plus. En définissant l'ordre du polynôme, vous définissez le degré de complexité du modèle.

population

Une population est une liste d'identifiants d'entités. Une population peut être définie comme liste de valeurs. Cette liste peut être extraite d'une colonne de table : on dit alors qu'elle est définie en extension, ou à travers une expression filtrante d'une autre population : on dit alors qu'elle est définie en intension.

population estampillée

Une population estampillée est une liste de paires <identifiant ; estampille temporelle> : le signifiance sémantique d'une telle construction peut être associée à des instantanés des entités à un moment donné : une entité donnée peut être représentée à différents estampilles temporelles dans une seule population estampillée.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

profit détecté

Le profit détecté est le type de profit proposé par défaut. Il permet de visualiser le pourcentage d'observations appartenant à la catégorie cible de la variable cible, c'est-à-dire la catégorie la moins fréquente, en fonction du taux d'observations sélectionné sur la totalité du jeu de données.

profit Lift

Le profit Lift permet de visualiser la différence entre un modèle parfait (Wizard) et un modèle aléatoire et entre le modèle généré et un modèle aléatoire. Le modèle aléatoire sert de référence et est toujours égal à 1.

profit normalisé

Le profit normalisé permet de visualiser l'apport du modèle généré par les composants SAP InfiniteInsight® par rapport à un modèle de type aléatoire, c'est-à-dire un modèle qui vous permettrait de sélectionner uniquement au hasard des observations dans votre base de données. Ce profit est utilisé pour les graphiques de détail des variables, qui présentent l'importance de chacune des catégories d'une variable donnée par rapport à la variable cible.

profit personnalisé

Le profit personnalisé vous permet de définir vos propres valeurs de profit, c'est-à-dire d'associer à chaque valeur de la variable cible un coût et un bénéfice.

R

racine

Morphème terminologique qui peut être utilisé soit tel quel comme mot - mot-racine -, soit comme élément de base dans un mot dérivé.

rapport de modélisation

Il s'agit d'un ensemble de rapports vous permettant une analyse fine de votre modèle.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

règle d'association

Une règle d'association est une relation d'implication de la forme X => Y. Cette règle signifie : si l'attribut X est présent dans une session, l'attribut Y l'est également. Deux mesures permettent d'indiquer la qualité d'une règle : le support et la confiance.

représentation en graphe bipartite / représentation en graphe non bipartite

L'affichage en graphe bipartite montre deux populations distinctes de nœuds (ou jeux de nœuds) avec les liens entre les deux jeux de nœuds. Par exemple, le premier jeux de nœuds peut représenter des clients et le second, des produits. De cette vue globale, une représentation en graphe non bipartite peut être dérivée pour se concentrer sur les liens entre les nœuds d'un jeu de données.

robustesse

Le degré de robustesse correspond à la capacité prédictive d'un modèle appliqué à un jeu de données d'application.

ROC

La courbe ROC (Receiver Operating Characteristic) est dérivée de la théorie de détection du signal. Elle permet d'étudier les variations de la spécificité et de la sensitivité d'un test pour différentes valeurs du seuil de discrimination.

rôle

Une variable peut avoir trois rôles différents, à savoir :

- variables cibles
- variables explicatives
- variables de poids.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

S

saisonal

Variations liées à des événements calendaires

score

La note d'évaluation numérique par rapport à une problématique donnée

segment

Un segment est un ensemble de valeurs défini par ses bornes (borne supérieure et borne inférieure). Les segments sont le résultat d'une manipulation qu'on appelle la segmentation.

sensibilité

La sensibilité, qui apparaît sur l'axe des ordonnées, est la proportion de signaux trouvés qui ont été correctement identifiés (également appelés vrais positifs).

série temporelle

Une série temporelle est une suite de données ponctuelles, mesurées en général successivement et dans des intervalles de temps (souvent uniformes).

serveur authentifié

Les utilisateurs peuvent seulement communiquer avec le serveur authentifié SAP InfiniteInsight® s'ils ont le mot de passe. Le serveur authentifié SAP InfiniteInsight® redirige l'authentification vers des modules enfichables d'authentification (PAM).

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

session

Une session est identifiée par une clé unique et elle est composée d'une ou plusieurs transactions.

simulation

Application d'un modèle à un seul individu

source de données

La source de données comprend la source elle-même, comme une base de données relationnelle, une base de données simple ou un fichier texte, et l'information de connexion nécessaire pour accéder les données.

statistiques croisées

Une méthode de prédire l'exactitude d'un modèle de classification ou de régression. Le jeu de données est divisé en plusieurs parties qui sont utilisées l'une après l'autre pour tester un modèle qui correspond aux autres parties.

stockage

SAP InfiniteInsight® utilise plusieurs types de stockage pour décrire des données :

- date,
- date et horaire,
- nombre,
- entier,
- chaine de caractères.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

stratégie de découpage

Une stratégie de découpage est une technique qui permet de décomposer un jeu de données d'apprentissage en trois sous-jeux distincts :

- un sous-jeu d'estimation,
- un sous-jeu de validation,
- un sous-jeu de test.

Ce découpage permet une validation croisée des modèles générés. Il existe neuf types de stratégies de découpage.

stratégie de découpage aléatoire

Cette stratégie distribue les données du jeu de données initial de manière aléatoire dans les trois sous-jeux d'estimation, de validation et de test.

stratégie de découpage périodique

Cette stratégie suit le cycle de distribution suivant :

1. Trois lignes du jeu de données initial sont distribuées dans le sous-jeu d'estimation.
2. Une ligne est distribuée dans le sous-jeu de validation.
3. Une ligne est distribuée dans le sous-jeu de test.
4. La distribution reprend à l'étape *1*.

stratégie de découpage personnalisée

La stratégie de découpage personnalisée vous permet de définir vos propres sous-jeux de données. Pour l'utiliser, vous devez préparer au préalable (avant de lancer les fonctionnalités SAP InfiniteInsight®) trois sous-jeux correspondant aux sous-jeux d'estimation, de validation et de test.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

stratégie de découpage séquentielle

Cette stratégie découpe le jeu de données initial en trois blocs, correspondant aux proportions de découpage habituelles :

- les lignes correspondant aux premiers 3/5 du jeu de données initial sont distribuées d'un bloc dans le jeu de données d'estimation,
- les lignes correspondant aux 1/5 suivant du jeu de données initial sont distribuées d'un bloc dans le jeu de données de validation,
- les lignes correspondant aux derniers 1/5 du jeu de données initial sont distribuées d'un bloc dans le jeu de données de test.

support

Le support d'une règle est la mesure qui indique le nombre de sessions qui vérifient la règle. Par exemple le numéro de session qui contient l'Itemset {A,B,C} et l'Item D.

T

table calendaire

Une table calendaire est utilisée pour faciliter le développement de solutions pour tous les modèles qui impliquent des dates. C'est une pratique courante d'avoir une table calendaire qui contient déjà une partie ou toute l'information nécessaire pour accomplir la plupart des tâches complexes comprenant des dates avec des requêtes simples dans la base de données.

table d'analyse

Une table d'analyse est un tableau avec des lignes et des colonnes. Chaque ligne représente une "observation". Des rôle peuvent être assignés aux colonnes, tels que “Input”, “skip”, “target” or “weight”.

table de données

Une table de données est un ensemble de données présentées sous la forme d'un tableau à deux dimensions.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

taux de classement

rapport entre le nombre d'enregistrements classifiés correctement et le nombre total d'enregistrements

test de K-S

Le K-S est le critère de Kolmogorov-Smirnov appliqué comme mesure de la déviation par rapport aux taux de réponse uniformes pour les catégories d'une variable. K-S est un test d'ajustement non paramétrique qui repose sur la déviation maximale entre les fonctions de distribution cumulative et empirique.

transaction

Une transaction est définie par :

- une clé unique
- la clé de la session liée
- un attribut, appelé Item.

type de profit

Un type de profit permet de calculer le profit réalisable grâce à l'utilisation d'un modèle. De manière générale, un bénéfice est associé aux valeurs souhaitées (ou attendues) de la variable cible et un coût est associé à ses valeurs non souhaitées (ou non attendues).

type de variable

Il existe plusieurs types de variables :

- variables continues,
- variables ordinales,
- variables nominales,
- variables textuelles.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

V

valeur aberrante

Les individus qui présentent, sur une ou plusieurs variables, des caractéristiques très éloignées de la majorité. On parle aussi de valeurs extrêmes.

valeur manquante

Des valeurs peuvent manquer parce qu'elles n'ont pas été relevées, renseignées ou parce qu'elles ont été perdues.

variable

Une variable correspond à un attribut qui décrit les observations stockées dans votre base de données.

Dans les composants SAP InfiniteInsight®, une variable est définie soit par un type, soit par un format de stockage, ou bien un rôle.

variable cible

Une variable cible est une variable que vous cherchez à expliquer ou dont vous souhaitez prédire les valeurs dans un jeu de données d'application. Elle correspond à votre problématique métier.

variable continue

Les variables continues sont des variables dont les valeurs sont numériques continues et ordonnées. Des opérations arithmétiques peuvent être effectuées sur ces valeurs, telles que la somme ou la moyenne.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

variable de poids

Une variable de poids permet d'attribuer un poids relatif à chacune des observations qu'elle décrit, et d'orienter le processus d'apprentissage en conséquence. Déclarer une variable comme variable de poids revient à faire un nombre de copies pour chacune des observations du jeu de données qui soit proportionnel à la valeur qu'elles possèdent pour cette variable.

variable exclue

la véritable cible

variable explicative

Une variable explicative est une variable qui décrit vos données et qui sert à expliquer une variable cible.

variable extra prédictible

Variable dont on connaît la valeur pour la période qu'on veut prédire.

variable nominale

Les variables nominales sont des variables dont les valeurs sont discrètes, c'est-à-dire appartenant à des catégories, et non ordonnées. Les variables nominales peuvent être :

- numériques, c'est-à-dire avoir pour valeurs des nombres (number).
- textuelles, c'est-à-dire avoir pour valeurs des chaînes de caractères (string).
- Important
- Les variables binaires sont considérées comme des variables nominales.

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

variable ordinale

Les variables ordinales sont des variables dont les valeurs sont discrètes, c'est-à-dire appartenant à des catégories, et ordonnées. Les variables ordinales peuvent être :

- numériques, c'est-à-dire avoir pour valeurs des nombres (number). Elles sont alors ordonnées selon l'ordre numérique naturel (0, 1, 2, etc.).
- textuelles, c'est-à-dire avoir pour valeurs des chaîne de caractères (string). Elles sont alors ordonnées de manière alphabétique. Important
- Les variables binaires sont considérées comme des variables nominales.

vrai négatif

signaux correctement identifiés comme faisant partie de la classe des non-signaux

vrai positif

signaux correctement identifiés

© 2014 SAP SE or an SAP affiliate company. All rights reserved-

www.sap.com/contactsap

© 2014 SAP SE ou société affiliée SAP. Tous droits réservés.

Toute reproduction ou communication de la présente publication, même partielle, par quelque procédé et à quelque fin que ce soit, est interdite sans l'autorisation expresse et préalable de SAP SE ou d'une société affiliée SAP. Les informations du présent document sont susceptibles d'être modifiées sans préavis.

Certains logiciels commercialisés par SAP SE et ses distributeurs contiennent des composants logiciels qui sont la propriété d'éditeurs tiers. Les spécifications des produits peuvent varier d'un pays à l'autre.

Les informations du présent document sont fournies par SAP SE ou par une société affiliée SAP uniquement à titre informatif, sans engagement ni garantie d'aucune sorte. SAP SE ou ses sociétés affiliées ne pourront en aucun cas être tenues responsables des erreurs ou omissions relatives à ces informations. Les seules garanties fournies pour les produits et les services de SAP SE ou d'une société affiliée SAP sont celles énoncées expressément à titre de garantie accompagnant, le cas échéant, lesdits produits et services. Aucune des informations contenues dans le présent document ne saurait constituer une garantie supplémentaire.

SAP et tous les autres produits et services SAP mentionnés dans ce document, ainsi que leurs logos respectifs, sont des marques commerciales ou des marques déposées de SAP SE (ou d'une société affiliée SAP) en Allemagne ainsi que dans d'autres pays. Tous les autres noms de produit et service mentionnés sont des marques commerciales de leurs sociétés respectives.

Pour plus d'informations sur les marques déposées, voir http://www.sap.com/corporate-en/legal/copyright/index.epx.