IBM SPSS Data Preparation 29

Remarque

Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations figurant dans la section «Remarques», à la page 25.

Notice d'édition

Cette édition s'applique à la version 29, édition 0, modification 1 d' IBM® SPSS Statistics et à toutes les éditions et modifications ultérieures, sauf indication contraire dans les nouvelles éditions.

Table des matières

Chapitre 1. Préparation des données

Les fonctions de préparation des données ci-après sont incluses dans l'édition de base.

Introduction à la préparation des données

L'augmentation de la demande d'information est proportionnelle à l'augmentation de la puissance des systèmes informatiques, provoquant la multiplication des données collectées, tout comme celle des observations, des variables et des erreurs de saisie de données. Ces erreurs représentent l'ennemi principal des modèles de prévision, ces derniers servant à entreposer les données, vous devez donc conserver des données « propres ». Cependant, la quantité de données entreposées a augmenté de telle façon qu'il n'est plus possible de vérifier manuellement les observations. Il devient alors primordial d'automatiser les processus de validation des données.

La préparation des données vous permet d'identifier les observations inhabituelles et les observations non valides, ainsi que les variables et les valeurs de données dans votre jeu de données actif, de plus ce module prépare les données pour la modélisation.

Utilisation des procédures de préparation des données

Votre utilisation des procédures de préparation des données dépend de vos besoins. Un processus standard de validation des données, une fois vos données chargées, consiste à :

- **Préparer les métadonnées :** Etudiez les variables de votre fichier de données et déterminez leur valeur valide, leur libellé et leurs niveaux de mesure. Identifiez les combinaisons des valeurs de variables impossibles qui sont couramment mal codées. Définissez les règles de validation en vous basant sur cette information. Cette tâche peut prendre beaucoup de temps, mais elle peut s'avérer vraiment utile si vous devez régulièrement valider des fichiers de données possédant des attributs similaires.
- **Valider les données :** Exécutez des vérifications et des contrôles de base des règles de validation définies afin d'identifier les observations inhabituelles, les variables et les valeurs de données. Une fois les données invalides repérées, déterminez-en la cause et corrigez le problème. Vous devrez peut-être effectuer une étape supplémentaire de préparation des métadonnées.
- **Préparer le modèle :** Utilisez une préparation automatique des données afin de transformer les champs d'origine, ce qui va améliorer la génération de modèle. Identifiez les valeurs extrêmes statistiques potentielles pouvant être à l'origine de problèmes rencontrés dans de nombreux modèles de prévision. Certaines valeurs extrêmes sont dues à des valeurs de variables invalides qui n'ont pas été identifiées. Vous devrez peut-être effectuer une étape supplémentaire de préparation des métadonnées.

Une fois que votre fichier de données est "propre", vous êtes prêt à construire des modèles à partir d'autres modules complémentaires.

Règles de validation

Une règle sert à déterminer la validité d'une observation. Il existe deux types de règles de validation :

- **Règles de variable unique :** Les règles de variable unique sont composées d'un ensemble fixe de vérification s'appliquant à une variable unique, telle que les vérifications des valeurs hors plage. Les valeurs valides peuvent être exprimées sous la forme d'une plage de valeurs ou d'une liste de valeurs possibles en ce qui concerne les règles de variable unique.
- **Règles de variable croisée :** Les règles sur variables croisées sont des règles définies par l'utilisateur qui peuvent être appliquées à une variable ou à une combinaison de variables. Les règles de variable croisée sont définies par une expression logique qui repère les valeurs non valides.

Les règles de validation sont enregistrées dans le dictionnaire de données de votre fichier de données. Vous pouvez ainsi spécifier une règle une fois et la réutiliser ensuite.

Chargement des règles de validation prédéfinies

Vous pouvez rapidement obtenir un ensemble de règles de validation prêtes à l'emploi en chargeant des règles prédéfinies à partir d'un fichier de données externe inclus dans l'installation.

Pour charger des règles de validation prédéfinies

1. À partir des menus, sélectionnez :

   Données > **Validation** > **Charger des règles prédéfinies...**

Vous pouvez également utiliser l'assistant Copier des propriétés de données pour charger les règles à partir de n'importe quel fichier de données.

Définir des règles de validation

La boîte de dialogue Définir des règles de validation vous permet de créer et d'afficher des règles de validation de variable unique et de variable croisée.

Pour créer et afficher des règles de validation

1. Á partir des menus, sélectionnez :

   Données > **Validation** > **Définir des règles...**

   La boîte de dialogue est remplie de règles de validation de variable unique et de variable croisée issues du dictionnaire de données. En l'absence de règles, une nouvelle règle de substitution que vous pouvez modifiez en fonction de vos besoins est créée automatiquement.
2. Sélectionnez des règles individuelles dans les onglets Règles de variable unique et Règles de variable croisée pour afficher et modifier leurs propriétés.

Définition des règles de variable unique

L'onglet Règles de variable unique vous permet de créer, d'afficher et de modifier les règles de validation de variable unique.

Règles : La liste affiche les règles de validation de variable unique par nom et le type de variable auquel la règle peut être appliquée. A l'ouverture de la boîte de dialogue, les règles définies dans le dictionnaire de données s'affichent ou, si aucune règle n'a été définie, une règle de substitution intitulée "Règle de variable unique 1" apparaît. Les boutons suivants apparaissent au-dessous de la liste Règles :

- **Nouveau :** Ajoute une nouvelle entrée au bas de la liste Règles. La règle est sélectionnée et le nom "SingleVarRule *n*" lui est appliqué, *n* correspondant à un nombre entier de sorte que le nom de la nouvelle règle n'ait pas de doublon parmi les règles de variable unique et de variable croisée.
- **Dupliquer :** Ajoute une copie de la règle sélectionnée au bas de la liste Règles. Le nom de la règle est ajusté de sorte qu'il n'y ait pas de doublon parmi les règles de variable unique et de variable croisée. Par exemple, si vous dupliquez "SingleVarRule 1", le nom de la première règle dupliquée sera "Copy of SingleVarRule 1", celui de la deuxième sera "Copy (2) of SingleVarRule 1", etc.
- **Supprimer.** Supprime la règle sélectionnée.

Définition de règles : Ces contrôles vous permettent d'afficher et de définir les propriétés d'une règle sélectionnée.

- **Nom** : Le nom de la règle doit être unique parmi les règles de variable unique et de variable croisée.
- **Type.** Il s'agit du type de variable auquel une règle est appliquée. Effectuez votre sélection à partir de **Numérique**, **Chaîne** et **Date**.
- **Format.** Le format vous permet de sélectionner le format de date pour les règles pouvant être appliquées à des variables de date.
- **Valeurs valides :** Vous pouvez indiquer les valeurs valides sous la forme d'une plage ou d'une liste de valeurs.

  Définition de la plage

Les contrôles de définition de la plage vous permettent de spécifier une plage de valeurs valides. Les valeurs se trouvant à l'extérieur de cette plage sont repérées et considérées comme invalides.

Entrez la valeur minimale ou la valeur maximale ou bien les deux pour spécifier une plage. Les contrôles des cases à cocher vous permettent de repérer les valeurs non libellées et non entières à l'intérieur de cette plage.

Définition de liste

Les contrôles de définition de liste vous permettent de définir une liste de valeurs valides. Les valeurs non comprises dans la liste sont repérées comme invalides.

Entrez les valeurs de la liste dans la grille. La case à cocher détermine si les observations sont importantes lorsque les valeurs de données chaîne sont comparées à la liste de valeurs possibles pour vérification.

- **Autoriser les valeurs manquantes de l'utilisateur :** Cette fonctionnalité contrôle si les valeurs manquantes de l'utilisateur sont repérées comme invalides.
- **Autoriser les valeurs système manquantes :** Cette fonctionnalité contrôle si les valeurs système manquantes sont repérées comme invalides. Elle ne s'applique pas aux types de règle chaîne.
- **Autoriser les valeurs vides :** Cette fonctionnalité contrôle si les valeurs chaîne vides (complètement vides) sont repérées comme invalides. Elle ne s'applique pas aux types de règle non-chaîne.

Définition des règles de variable croisée

L'onglet Règles de variable croisée vous permet de créer, d'afficher et de modifier les règles de validation de variable croisée.

Règles : La liste affiche les règles de validation de variable croisée par nom. A l'ouverture de la boîte de dialogue, une règle de substitution intitulée "CrossVarRule 1" s'affiche. Les boutons suivants apparaissent au-dessous de la liste Règles :

- **Nouveau :** Ajoute une nouvelle entrée au bas de la liste Règles. La règle est sélectionnée et le nom "CrossVarRule *n*" lui est appliqué, *n* correspondant à un nombre entier de sorte que le nom de la nouvelle règle n'ait pas de doublon parmi les règles de variable unique et de variable croisée.
- **Dupliquer :** Ajoute une copie de la règle sélectionnée au bas de la liste Règles. Le nom de la règle est ajusté de sorte qu'il n'y ait pas de doublon parmi les règles de variable unique et de variable croisée. Par exemple, si vous dupliquez "CrossVarRule 1", le nom de la première règle dupliquée sera "Copy of CrossVarRule 1", celui de la deuxième sera "Copy (2) of CrossVarRule 1", etc.
- **Supprimer.** Supprime la règle sélectionnée.

Définition de règles : Ces contrôles vous permettent d'afficher et de définir les propriétés d'une règle sélectionnée.

- **Nom** : Le nom de la règle doit être unique parmi les règles de variable unique et de variable croisée.
- **Expression logique :** Il s'agit de la définition de règle. Vous pouvez coder l'expression de sorte que les observations invalides aient pour résultat 1.

Création d'expressions

1. Pour construire une expression, vous pouvez soit coller les composants dans le champ Expression, soit les saisir directement depuis le clavier.

- Pour coller des fonctions ou des variables système couramment utilisées, sélectionnez un groupe dans la liste Groupe de fonctions, puis, dans la liste Fonctions et variables spéciales, double-cliquez sur la fonction ou la variable voulue (ou sélectionnez-la, puis cliquez sur **Insérer**). Entrez tous les paramètres indiqués par un point d'interrogation (cette opération ne concerne que les fonctions). Le groupe de fonctions libellé **Tous** répertorie toutes les fonctions et variables système disponibles. Une brève description de la variable ou de la fonction sélectionnée apparaît dans une zone particulière de la boîte de dialogue.
- Les constantes alphanumériques doivent être présentées entre guillemets ou apostrophes.
- Si des valeurs contiennent des chiffres décimaux, utilisez la virgule comme indicateur décimal.

Validation des données

La boîte de dialogue Valider des données vous permet d'identifier des observations suspicieuses ou invalides, des variables et des valeurs de données dans le jeu de données actif.

Exemple : Un analyste de données doit fournir une enquête de satisfaction client à son client tous les mois. L'analyste doit effectuer une vérification de la qualité des données reçues chaque mois, afin de contrôler qu'il n'y a pas d'ID client incomplet, de valeurs de variables hors plage, de combinaisons de valeurs de variable régulièrement saisies par erreur. Avec la boîte de dialogue Valider des données, l'analyste peut spécifier les variables qui ne servent à identifier que les clients, définir les règles de variable unique pour les plages de variables valides et enfin définir les règles de variable croisée afin de repérer les combinaisons impossibles. La procédure renvoie un rapport sur les observations et les variables posant problèmes. De plus, les données possèdent les mêmes éléments de données chaque mois, ce qui permet à l'analyste d'appliquer les règles au nouveau fichier de données du mois suivant.

Statistiques : La procédure génère des listes de variables, d'observations et de valeurs de données qui n'ont pas passé plusieurs contrôles, des effectifs de violation des règles de variable unique et de variable croisée, ainsi que de simples récapitulatifs descriptifs des variables d'analyse.

Pondérations : La procédure ignore la spécification de la variable de pondération et la traite comme toute autre variable d'analyse.

Pour valider des données

1. A partir des menus, sélectionnez :

   Données > **Validation** > **Valider des données...**
2. Sélectionnez une ou plusieurs variables d'analyse afin de les faire valider par des vérifications de base des variables ou par des règles de validation de variable unique.

   Vous pouvez également :
3. Cliquer sur l'onglet **Règles de variable croisée** et appliquer une ou plusieurs règles de variable croisée.

(En option) Vous avez également ces possibilités :

- Sélectionner une ou plusieurs variables d'identification d'observations afin de vérifier s'ils existent des ID dupliqués ou incomplets. Les variables d'ID d'observation sont également utilisées pour libeller les sorties par observations. Si deux ou plus de deux variables d'ID d'observations sont spécifiées, la combinaison de leurs valeurs est traitée comme un identificateur d'observations.

Champs dont le niveau de mesure est inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou de plusieurs variables (champs) du jeu de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Analyser les données : Lit les données dans le jeu de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si le jeu de données est volumineux, l'opération peut prendre un certain temps.

Affecter manuellement : Ouvre une boîte de dialogue dans laquelle figurent tous les champs dont le niveau de mesure est inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans la vue de variable de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Vérifications de base de validation des données

L'onglet Vérifications de base vous permet de sélectionner les vérifications de base pour les variables d'analyse, les identificateurs d'observations ainsi que les observations complètes.

Variables d'analyse : Si vous avez sélectionné des variables d'analyse dans l'onglet Variables, vous pouvez sélectionner la ou les vérifications suivantes correspondant à leur validité. La case à cocher vous permet d'activer ou de désactiver les vérifications.

- **Pourcentage maximal de valeurs manquantes :** Répertorie les variables d'analyse dont le pourcentage de valeurs manquantes est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100.
- **Pourcentage maximal d'observations dans une catégorie unique :** Lorsque des variables d'analyse sont catégorielles, cette option répertorie alors les variables d'analyse catégorielles dont le pourcentage d'observations représentant une catégorie unique non manquante est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100. Le pourcentage est basé sur des observations n'ayant pas de valeur manquante de la variable.
- **Pourcentage maximal de catégories dont l'effectif est 1 :** Lorsque des variables d'analyse sont catégorielles, cette option répertorie alors les variables d'analyse catégorielles dont le pourcentage des catégories des variables contenant une seule observation est supérieur à la valeur indiquée. La valeur indiquée doit être un nombre positif inférieur ou égal à 100.
- **Coefficient de variation minimum :** Lorsque des variables d'analyse sont mesurées sur une échelle, cette option répertorie les variables d'analyse d'échelle dont la valeur absolue du coefficient de variation est inférieure à la valeur indiquée. Cette option ne s'applique qu'aux variables dont la moyenne n'est pas nulle. La valeur indiquée doit être un nombre non-négatif. Pour désactiver le coefficient de vérification de la variation, tapez 0.
- **Ecart type minimum :** Lorsque des variables d'analyse sont mesurées sur une échelle, cette option répertorie les variables d'analyse d'échelle dont l'écart type est inférieur à la valeur indiquée. La valeur indiquée doit être un nombre non-négatif. Pour désactiver la vérification de l'écart type, tapez 0.

Identificateurs d'observations : Si vous avez sélectionné des variables d'identificateurs d'observations dans l'onglet Variables, vous pouvez sélectionner la ou les vérifications suivantes correspondant à leur validité.

- **Repérer les ID incomplets :** Cette option répertorie les observations dont les identificateurs d'observations sont incomplets. Pour une observation donnée, un identificateur est considéré comme incomplet lorsque la valeur de toute variable ID est vide ou manquante.
- **Repérer les ID dupliqués :** Cette option répertorie les observations dont les identificateurs d'observations sont dupliqués. Les identificateurs incomplets sont exclus de l'ensemble de doublons possibles.

Repérer les observations vides : Cette option répertorie les observations dont toutes les variables sont vides ou nulles. Pour identifier des observations vides, vous pouvez utiliser toutes les variables du fichier (à l'exception des variables ID) ou seulement les variables d'analyse définies sur l'onglet Variables.

Règles de variable unique de la validation des données

L'onglet Règles de variable unique affiche les règles de validation de variable unique disponibles et vous permet de les appliquer aux variables d'analyse. Pour définir d'autres règles de variable unique, cliquez sur **Définir des règles**. Pour plus d'informations, voir «Définition des règles de variable unique», à la page 2.

Variables d'analyse : La liste affiche les variables d'analyse, récapitule leurs distributions et indique également le nombre de règles appliqué à chaque variable. Notez que les valeurs manquantes définies par l'utilisateur et par le système ne sont pas incluses dans les récapitulatifs. La liste déroulante Afficher contrôle l'affichage des variables. Vous pouvez sélectionner les affichages suivants : **Toutes les variables**, **Variables numériques**, **Variables de chaîne** et **Variables de date**.

Règles : Pour appliquer des règles à des variables d'analyse, sélectionnez une ou plusieurs variables et vérifiez toutes les règles que vous voulez appliquer dans la liste Règles. La liste Règles n'affiche que les règles appropriées aux variables d'analyse sélectionnées. Si, par exemple, vous sélectionnez des variables d'analyse numériques, seules les règles numériques s'affichent. Si vous sélectionnez une variable de chaîne, seules les règles chaîne s'affichent. Si vous n'avez sélectionné aucune variable d'analyse ou si les types de données ont été mélangés, aucune règle ne s'affiche.

Distributions de variables : Les récapitulatifs de distribution affichés dans la liste Variables d'analyse peuvent être basés sur l'ensemble des observations ou sur une analyse des premières observations *n*, comme indiqué dans la zone de texte Observations. Pour mettre à jour les récapitulatifs de distribution, cliquez sur **Réanalyser**.

Règles de variable croisée de la validation des données

L'onglet Règles de variable croisée affiche les règles de variable croisée disponibles et vous permet de les appliquer aux données. Pour définir d'autres règles de variable croisée, cliquez sur **Définir des règles**. Pour plus d'informations, voir «Définition des règles de variable croisée», à la page 3.

Sortie de la validation des données

Rapport par observation : Si vous avez appliqué des règles de validation de variable unique ou de variable croisée, vous pouvez demander un rapport répertoriant les violations des règles de validation pour les observations individuelles.

- **Nombre minimum de violations :** Cette option indique le nombre minimum de violations de règles nécessaires à l'intégration d'une observation au rapport. Spécifiez un nombre entier positif.
- **Nombre maximum d'observations :** Cette option indique le nombre maximum d'observations incluses dans le rapport d'observations. Entrez un nombre entier positif inférieur ou égal à 1000.

Règles de validation de variable unique : Si vous avez appliqué des règles de validation de variable unique, vous pouvez sélectionner le mode d'affichage et les résultats à afficher.

- **Récapituler les violations par variable d'analyse :** Pour chaque variable d'analyse, cette option affiche toutes les règles de validation de variable unique violées et le nombre de valeurs ayant violé chaque règle. Elle répertorie également le nombre total de violations de règles de variable unique pour chaque variable.
- **Récapituler les violations par règles :** Pour chaque règle de validation de variable unique, cette option affiche les variables ayant violé la règle et le nombre de valeurs non valides par variable. Elle répertorie également le nombre total de valeurs ayant violé chaque règle dans l'ensemble des variables.

Afficher les statistiques descriptives pour les variables d'analyse : Cette option vous permet de demander les statistiques descriptives pour les variables d'analyse. Une table de fréquences est généré pour chaque variable catégorielle. Un tableau de statistiques récapitulatives, comprenant la moyenne, l'écart type, les valeurs minimum et maximum, est généré pour les variables d'échelle.

Déplacer vers le haut du jeu de données actif les observations présentant des violations de règles de validation : Cette option permet de déplacer les observations contenant des violations de règles de variable unique ou de variable croisée au haut du jeu de données actif pour faciliter la lecture.

Enregistrement de la validation des données

L'onglet Enregistrer vous permet d'enregistrer les variables qui stockent les violations de règles dans le jeu de données actif.

Variables récapitulatives : Ces variables individuelles peuvent être enregistrées. Cochez une case pour enregistrer la variable. Les noms des variables par défaut sont fournis, vous pouvez les modifier.

- **Indicateur d'observations vides :** La valeur 1 est attribuée aux cas vides. Tous les autres cas sont codés 0. Les valeurs de la variable reflètent la portée spécifiée dans l'onglet Vérifications de base.
- **Dupliquer le groupe ID :** Le même numéro de groupe est attribué aux observations disposant du même identificateur d'observations (sauf les observations possédant des identificateurs incomplets) Les observations disposant d'identificateurs uniques ou incomplets sont codées 0.
- **Indicateur ID incomplet :** Les cas avec des identificateurs de cas vides ou incomplets reçoivent la valeur 1. Tous les autres cas sont codés 0.
- **Violations d'une règle de validation :** Il s'agit de l'effectif total par observation de violations des règles de validation de variable unique et de variable croisée.

Remplacer les variables récapitulatives existantes : Les variables enregistrées dans un fichier de données doivent avoir des noms identiques ou remplacer les variables de même nom.

Enregistrer les variables indicateur : Cette option vous permet d'effectuer un enregistrement complet des violations des règles de validation. Chaque variable correspond à l'application d'une règle de validation et dispose d'une valeur de 1 si l'observation viole la règle et d'une valeur de 0 dans le cas contraire.

Préparation automatisée des données

La préparation des données pour l'analyse est une des étapes les plus importantes des projets et généralement, l'une de celles qui prend le plus de temps. La préparation automatique des données (ADP) s'occupe de cette tâche à votre place, analyse vos données, identifie les corrections, supprime les champs problématiques ou inutiles, dérive de nouveaux attributs si nécessaire et améliore les performances grâce à des techniques de balayage intelligentes. Vous pouvez utiliser l'algorithme en mode complètement **automatique**, le laissant choisir et appliquer les corrections ou vous pouvez utiliser son mode **interactif** qui prévoit les modifications avant qu'elles ne soient effectuées vous laissant libre de les accepter ou de les refuser.

L'utilisation de l'ADP vous permet de préparer facilement et rapidement vos données pour la génération de modèle, sans qu'il soit nécessaire de maîtriser les concepts de statistiques utilisés. Les modèles seront alors créés et les scores déterminés plus rapidement ; de plus, l'utilisation de l'ADP améliore la robustesse des processus de modélisation automatique.

Remarque : Lorsque la préparation automatique des données prépare un champ pour l'analyse, elle crée un nouveau champ contenant les ajustements ou les transformations, au lieu de remplacer les valeurs et les propriétés existantes de l'ancien champ. L'ancien champ n'est pas utilisé dans une analyse ultérieure ; son rôle est défini sur Aucun. Notez également que toutes les informations de valeur manquantes de l'utilisateur ne sont pas transférées vers ces champs nouvellement créés, et toutes les valeurs manquantes dans le nouveau champ sont manquantes du système.

Exemple : Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour signaler des réclamations suspectes et potentiellement frauduleuses. Avant de construire le modèle, il est nécessaire de préparer les données à l'aide de la préparation automatique des données. La compagnie souhaitant être capable de consulter et modifier les transformations avant de les appliquer, elle utilise la préparation automatique des données de manière interactive.

Un groupe automobile suit les ventes de véhicules automobiles personnels divers. Afin d'être en mesure d'identifier les modèles dont les ventes sont très satisfaisantes et ceux pour lesquels elles le sont moins, des responsables du groupe souhaitent établir une relation entre les ventes de véhicules et les caractéristiques des véhicules. Ils utilisent la préparation automatique des données pour cette analyse afin de construire des modèles à l'aide des données " avant" et " après " la préparation et de pouvoir en comparer les résultats.

Quel est votre objectif ? La préparation automatique des données recommande des étapes de préparation de données qui amélioreront la vitesse de création de modèles par les autres algorithmes et le pouvoir prédictif de ces modèles. Cela peut comprendre la transformation, la construction et la sélection de fonctions. La cible peut également être transformée. Vous pouvez spécifier les priorités de création de modèle sur lesquelles le processus de préparation des données doit se concentrer.

- **Equilibrer la vitesse et la précision.** Cette option prépare les données à accorder la même importance à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle et à la précision des prévisions.
- **Optimiser la vitesse :** Cette option prépare les données à accorder la priorité à la vitesse à laquelle les données sont traitées par les algorithmes de création de modèle. Lorsque vous travaillez avec de très grands jeux de données ou que vous recherchez une réponse rapide, sélectionnez cette option.
- **Optimiser l'exactitude.** Cette option prépare les données à accorder la priorité à la précision des prédictions produites par les algorithmes de création de modèle.

- **Analyse personnalisée :** Lorsque vous souhaitez modifier manuellement l'algorithme dans l'onglet Paramètres, sélectionnez cette option. Veuillez noter que ce paramètre est automatiquement sélectionné si vous modifiez ensuite des options dans l'onglet Paramètres qui ne sont pas compatibles avec l'un des autres objectifs.

Obtention d'une préparation automatique des données

À partir des menus, sélectionnez :

1. À partir des menus, sélectionnez :

   Transformer > **Préparer les données pour la modélisation** > **Automatique...**
2. Cliquez sur **Exécuter**.

Sinon, vous pouvez :

- Spécifier un objectif dans l'onglet Objectif.
- Spécifier les affectations de champ dans l'onglet Champs.
- Spécifier les paramètres d'expert dans l'onglet Paramètres.

Obtention d'une préparation interactive des données

1. À partir des menus, sélectionnez :

   Transformer > **Préparer les données pour la modélisation** > **Interactif...**
2. Cliquez sur **Analyser** dans la barre d'outils au-dessus de la boîte de dialogue.
3. Cliquez sur l'onglet Analyse pour consulter les étapes conseillées de préparation des données.
4. Si elles vous conviennent, cliquez sur **Exécuter**. Sinon, cliquez sur **Effacer l'analyse**, modifiez les paramètres de votre choix et cliquez sur **Analyse**.

(En option) Vous avez également ces possibilités :

- Spécifier un objectif dans l'onglet Objectif.
- Spécifier les affectations de champ dans l'onglet Champs.
- Spécifier les paramètres d'expert dans l'onglet Paramètres.
- Enregistrer les étapes de préparation des données conseillées dans un fichier XML en cliquant sur **Enregistrer XML**.

Onglet Champs

L'onglet Champs indique les champs à préparer pour une analyse ultérieure.

Utiliser des rôles prédéfinis : Cette option utilise des informations sur des champs existants. S'il n'existe qu'un champ avec le rôle Cible, il sera utilisé comme cible ; dans le cas contraire, il n'y aura pas de cible. Tous les champs avec un rôle prédéfini d'Entrée seront utilisés comme entrées. Au moins un champ d'entrée est requis. .

Utiliser des affectations de champ personnalisées : Lorsque vous remplacez des rôles de champs en les déplaçant de leur listes par défaut, la boîte de dialogue sélectionne automatiquement cette option. Lors des affectations personnalisées, spécifiez les champs suivants :

- **Cible (facultatif).** Si vous souhaitez créer des modèles nécessitant une cible, sélectionnez le champ cible. Il s'agit de la même action que lorsque l'on définit le rôle du champ sur Cible.
- **Entrées :** Sélectionnez un ou plusieurs champs d'entrée. Il s'agit de la même action que lorsque l'on définit le rôle du champ sur Entrée.

Onglet Paramètres

L'onglet Paramètres contient plusieurs groupes de paramètres différents que vous pouvez modifier pour affiner le traitement des données par l'algorithme. Si vous modifiez les paramètres par défaut et que ces

modifications sont incompatibles avec les autres objectifs, l'onglet Objectif est automatiquement mis à jour pour sélectionner l'option **Personnaliser l'analyse**.

Préparer les dates et les heures

De nombreux algorithmes de modélisation ne peuvent pas traiter directement les informations sur la date et l'heure. Ces paramètres vous permettent de calculer de nouvelles données de durée qui peuvent être utilisées comme entrées de modèle à partir des dates et des heures de vos données existantes. Les champs contenant les dates et les heures doivent être prédéfinis à l'aide des types de stockage de dates et d'heures. Il n'est pas recommandé de définir les champs de date et d'heure d'origine comme entrées de modèle après la préparation automatique des données.

Préparer les dates et les heures pour la modélisation : En désélectionnant cette option, vous désactivez tous les autres contrôles Préparer les dates et les heures, tout en conservant les sélections.

Calculer la durée écoulée jusqu'à la date de référence : Cette option génère le nombre d'années/mois/jours depuis une date de référence pour chaque variable qui contient des dates.

- **Date de référence :** Spécifiez la date à partir de laquelle la durée sera calculée en fonction des informations sur la date dans les données d'entrée. Si vous sélectionnez la **Date d'aujourd'hui**, la date du système actuelle est toujours utilisée lors de l'exécution de l'ADP. Pour utiliser une date spécifique, sélectionnez **Date fixe** et saisissez la date désirée.
- **Unités de la durée Date :** Indiquez si l'ADP doit décider automatiquement de l'unité de la durée Date ou choisissez dans les **unités fixes** des Années, Mois ou Jours.

Calculer la durée écoulée jusqu'à l'heure de référence : Cette option génère le nombre d'heures/minutes/secondes depuis une heure de référence pour chaque variable qui contient des heures.

- **Heure de référence :** Spécifier l'heure à partir de laquelle la durée sera calculée en fonction des informations sur l'heure dans les données d'entrée. Si vous sélectionnez **Heure actuelle**, cela signifie que l'heure du système actuelle est toujours utilisée lorsque l'ADP est exécuté. Pour utiliser une heure spécifique, sélectionnez **Heure fixe** et saisissez l'heure désirée.
- **Unités de la durée Heure :** Indiquez si l'ADP doit décider automatiquement de l'unité de la durée Heure ou choisissez dans les **unités fixes** des Heures, Minutes ou Secondes.

Extraire les éléments de temps cycliques : Utilisez ces paramètres pour scinder un champ de date ou d'heure en un ou plusieurs autres champs. Par exemple, si vous sélectionnez les trois cases de date, le champ de date d'entrée "1954-05-23" est divisé en trois champs : 1954, 5 et 23, chacun utilisant le suffixe défini dans le panneau **Noms de champs** et le champ de date d'origine est ignoré.

- **Extraire des dates :** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des années, des mois, des jours ou une des combinaisons possibles.
- **Extraire des heures :** Pour chaque entrée de date, spécifiez si vous souhaitez extraire des heures, des minutes ou des secondes ou une des combinaisons possibles.

Exclure des champs

Les données de mauvaise qualité peuvent affecter la précision de vos prédictions. Par conséquent, vous pouvez spécifier le niveau de qualité acceptable des fonctions d'entrée. Tous les champs constants ou avec 100% de valeurs manquantes sont automatiquement exclus.

Exclure les champs d'entrée de mauvaise qualité : En désélectionnant cette option, vous désactivez tous les autres contrôles Exclure les champs, tout en conservant les sélections.

Exclure les champs avec trop de valeurs manquantes : Les champs ayant plus que le pourcentage spécifié de valeurs manquantes sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à déselectionner cette option, et inférieure ou égale à 100, puisque les champs qui ne contiennent que des valeurs manquantes sont exclus automatiquement. La valeur par défaut est 50.

Exclure les champs nominaux avec trop de catégories uniques : Les champs nominaux ayant plus que le nombre spécifié de catégories sont supprimés de l'analyse. Spécifiez un nombre entier positif. La valeur

par défaut est 100. Cette option est utile pour supprimer automatiquement de la modélisation les champs contenant des informations d'enregistrement unique, tels que l'ID, l'adresse ou le nom.

Exclure les champs catégoriels avec trop de valeurs dans une seule catégorie : Les champs ordinaux et nominaux avec une catégorie contenant plus que le pourcentage spécifié d'enregistrements sont supprimés de l'analyse. Définissez une valeur supérieure ou égale à 0, ce qui revient à désélectionner cette option, et inférieure ou égale à 100, puisque les champs constants sont exclus automatiquement. La valeur par défaut est 95.

Réglage des mesures

Régler le niveau de mesure : En désélectionnant cette option, vous désactivez tous les autres contrôles Régler les mesures, tout en conservant les sélections.

Niveau de mesure : Indiquez si le niveau de mesure des champs continus avec "trop peu" de valeurs peut être réglé sur ordinal et si les champs ordinaux avec "trop" de valeurs peuvent être réglés sur continu.

- **Le nombre maximum de valeurs pour les champs ordinaux :** Les champs ordinaux ayant plus que le nombre spécifié de catégories sont reconvertis en champs continus. Spécifiez un nombre entier positif. La valeur par défaut est 10. Cette valeur doit être supérieure ou égale au nombre minimum de valeurs pour les champs continus.
- **Le nombre minimum de valeurs pour les champs continus :** Les champs continus ayant moins que le nombre spécifié de valeurs uniques sont reconvertis en champs ordinaux. Spécifiez un nombre entier positif. La valeur par défaut est 5.Cette valeur doit être inférieure ou égale au nombre maximal de valeurs pour les champs ordinaux.

Amélioration de la qualité des données

Préparer les champs pour améliorer la qualité des données : En désélectionnant cette option, vous désactivez tous les autres contrôles Améliorer la qualité des données, tout en conservant les sélections.

Traitement des valeurs extrêmes : Spécifier s'il faut remplacer les valeurs extrêmes des entrées et des cibles. Si oui, spécifier un critère de limite des valeurs extrêmes, mesuré en écarts types et une méthode de remplacement des valeurs extrêmes. Les valeurs extrêmes peuvent être remplacées soit en les tronquant (définies sur la valeur de césure) ou en les définissant comme valeurs manquantes. Les valeurs extrêmes définies comme valeurs manquantes suivent les paramètres de traitement des valeurs manquantes sélectionnées ci-dessous.

Remplacer les valeurs manquantes : Spécifier s'il faut remplacer les valeurs manquantes des champs continus, nominaux ou ordinaux.

Réorganiser les champs nominaux : Sélectionner cette option pour recoder les valeurs des champs nominaux (ensemble) de la plus petite catégorie (la moins utilisée) à la plus grande (la plus utilisée). Les valeurs des nouveaux champs démarrent à 0, 0 étant la catégorie la moins fréquente. Remarque : le nouveau champ doit être numérique même si la zone d'origine est une chaîne. Par exemple, si les valeurs d'un champ nominal sont "A", "A", "A", "B", "C", "C", la préparation automatique des données recodent "B" en 0, "C" en 1, et "A" en 2.

Rééchelonner les champs

Rééchelonner les champs : En désélectionnant cette option, vous désactivez tous les autres contrôles Rééchelonner les champs, tout en conservant les sélections.

Pondération d'analyse : Cette variable contient des pondérations (de régression ou d'échantillon) d'analyse. Les pondérations d'analyse sont utilisées pour représenter les différences de variance dans les niveaux du champ cible. Sélectionnez un champ continu.

Champs d'entrée continus : Cela normalisera les champs d'entrée continus avec une **transformation en score z** ou une **transformation min/max**. Le rééchelonnement des entrées est particulièrement utile lorsque vous sélectionnez l'option **Exécuter la construction des fonctions** dans les paramètres Sélectionner et Construire.

- **Transformation en score z :** Avec la moyenne et l'écart type observés utilisés comme estimations des paramètres de population, les champs sont standardisés puis les scores *z* sont mappés aux valeurs correspondantes d'une distribution normale avec la **moyenne finale** et l'**écart type final** spécifiés. Spécifiez un nombre pour la **moyenne finale** et un nombre positif pour l'**écart type final**. Les valeurs par défaut sont 0 et 1 respectivement, ce qui correspond au rééchelonnement standardisé.
- **Transformation min/max :** Avec la transformation minimum et maximum observée qui est utilisée comme estimations des paramètres de population, les champs sont mappés aux valeurs correspondantes d'une distribution uniforme avec la transformation **Minimum** et **Maximum** spécifiée. Spécifiez les nombres avec la transformation **Maximum** supérieure à la transformation **Minimum**.

Cible continue : Transforme une cible continue utilisant la transformation de Box-Cox en un champ ayant une distribution à peu près normale avec la **moyenne finale** et l'**écart type final** spécifiés. Spécifiez un nombre pour la **moyenne finale** et un nombre positif pour l'**écart type final**. Les valeurs par défaut sont 0 et 1 respectivement.

Remarque : Si une cible a été transformée par l'ADP, les modèles en résultant créés à l'aide de la cible transformée évaluent les unités transformées. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine. Pour plus d'informations, voir la rubrique . Pour plus d'informations, voir la rubrique «Rétablissement des scores», à la page 19.

Transformation des champs

Pour améliorer le pouvoir prédictif de vos données, vous pouvez transformer les champs d'entrée.

Transformer le champ pour la modélisation : En désélectionnant cette option, vous désactivez tous les autres contrôles Transformer les champs, tout en conservant les sélections.

Champs d'entrée qualitatifs : Les options suivantes sont disponibles :

- **Fusionner les catégories éparpillées pour optimiser l'association avec une cible.** Sélectionnez cette option pour créer un modèle plus petit en réduisant le nombre de champs à traiter en association avec la cible. Les modalités similaires sont identifiées en fonction de la relation entre l'entrée et la cible. Les catégories ne différant pas de manière significative, c'est-à-dire ayant une valeur *p* supérieure à la valeur spécifiée, sont fusionnées. Indiquez une valeur supérieure à 0 et inférieure ou égale à 1. Si toutes les catégories sont fusionnées en une seule, les versions originales et dérivées du champ sont exclues de l'analyse supplémentaire car elles n'ont aucune valeur en tant que prédicteur.
- **Lorsqu'il n'existe aucune cible, fusionner les modalités éparpillées en fonction de leur nombre.** Si le jeu de données n'a pas de cible, vous pouvez choisir de fusionner les catégories éparpillées des champs ordinaux et nominaux. La méthode d'effectifs égaux est utilisée pour fusionner les catégories ayant moins que le pourcentage minimum spécifié du nombre total d'enregistrements. Spécifiez une valeur supérieure ou égale à 0 et inférieure ou égale à 100. La valeur par défaut est 10. La fusion s'arrête lorsqu'il n'y a plus de catégorie comportant moins d'observations que le pourcentage minimal spécifié, ou lorsqu'il ne reste plus que deux catégories.

Champs d'entrée continus : Si le jeu de données comprend une cible catégorielle, vous pouvez regrouper les entrées continues ayant de fortes associations pour améliorer les performances du traitement. Les casiers sont créés en fonction des propriétés des "sous-ensembles homogènes" qui sont identifiés avec la méthode de Scheffé qui utilise la valeur de *p* comme valeur alpha de la valeur critique pour déterminer les sous-ensembles homogènes. Indiquez une valeur supérieure à 0 et inférieure ou égale à 1. La valeur par défaut est 0,05. Si l'opération de regroupement génère un regroupement unique pour un champ spécifique, les versions d'origine et regroupées du champ sont exclues car elles n'ont pas de valeur de prédicteur.

Remarque : Le regroupement dans l'ADP est différent du regroupement optimal. Le regroupement optimal utilise des informations d'entropie pour convertir un champ continu en un champ catégoriel ; il doit trier les données et les stocker dans la mémoire. L'ADP utilise des sous-ensembles homogènes pour regrouper un champ continu. Cela signifie que le regroupement ADP n'a pas besoin de trier les données et ne stocke pas toutes les données dans une mémoire. L'utilisation de la méthode des sous-ensembles homogènes pour regrouper un champ continu signifie que le nombre de catégories après le regroupement est toujours inférieur ou égal au nombre de catégories dans la cible.

Sélection et construction

Pour améliorer le pouvoir prédictif de vos données, vous pouvez construire de nouveaux champs basés sur les champs existants.

Exécuter la sélection des fonctions : Une entrée continue est supprimée de l'analyse si la valeur de *p* pour sa corrélation avec la cible est supérieure à la valeur de *p* spécifiée.

Exécuter la construction des fonctions : Sélectionnez cette option pour dériver de nouvelles fonctions d'une combinaison de plusieurs fonctions existantes. Les anciennes fonctions ne sont pas utilisées dans l'analyse ultérieure. Cette option s'applique uniquement aux fonctions d'entrée continues où la cible est continue ou lorsqu'il n'y a pas de cible.

Noms de champ

Pour identifier facilement les fonctions nouvelles et transformées, l'ADP crée et applique de nouveaux noms, préfixes ou suffixes de base. Vous pouvez modifier ces noms pour qu'ils soient plus adaptés à vos propres besoins et données.

Champs transformés et construits : Spécifiez les extensions de nom à appliquer aux champs cibles et d'entrées transformés.

En outre, spécifiez le nom du préfixe à appliquer aux fonctions construites à l'aide des paramètres Sélectionner et Construire. Le nouveau nom est créé en ajoutant un suffixe numérique à ce nom de racine du préfixe. Le format du nombre dépend du nombre de nouvelles fonctions dérivées, par exemple :

- si 1 à 9 caractéristiques sont construites, elles seront nommées : caractéristique1 à caractéristique9.
- si 10 à 99 fonctions sont construites, elles seront nommées : fonction01 à fonction 99.
- si 100 à 999 caractéristiques sont construites, elles seront nommées : caractéristique001 à caractéristique999.

Cela permet que les fonctions construites soient triées dans un ordre cohérent quel que soit leur nombre.

Durée calculée à partir des dates et des heures : Spécifier les extensions de nom à appliquer aux durées calculées à partir des dates et des heures.

Éléments cycliques extraits de dates et des heures : Spécifier les extensions de nom à appliquer aux éléments cycliques extraits des dates et des heures.

Application et enregistrement des transformations

Selon que vous utilisez la boîte de dialogue de préparation automatique ou interactive des données, les paramètres d'application et d'enregistrement des transformations des données diffèrent légèrement.

Paramètres Appliquer les transformations de la préparation automatique des données

Données transformées. Ces paramètres spécifient l'emplacement de l'enregistrement des données transformées.

- **Ajouter de nouveaux champs à le jeu de données actif.** Tous les champs créés par la préparation automatique des données sont ajoutés comme nouveaux champs à le jeu de données actif. **Mettre à jour les rôles pour les champs analysés** définira le rôle sur Aucun pour tous les champs exclus d'une analyse ultérieure par la préparation automatique des données.
- **Créer un nouveau jeu de données ou un fichier contenant les données transformées :** Les champs recommandés par la préparation automatique des données sont ajoutés à un nouveau jeu de données ou à un fichier. **Inclure les champs non analysés** ajoute les champs dans le jeu de données d'origine qui n'ont pas été spécifiés dans l'onglet Champs du nouveau jeu de données. Cette option est utile pour transférer vers le nouveau jeu de données les champs contenant des informations non utilisées dans la modélisation, telles que l'ID, l'adresse ou le nom.

Paramètre Appliquer et Enregistrer de la préparation automatique des données

Le groupe des données transformées est le même que celui de la préparation interactive des données. Les options supplémentaires suivantes sont disponibles pour la préparation automatique des données :

Appliquer les transformations. Dans les boîtes de dialogue de la Préparation automatique des données, déselectionner cette option revient à désactiver tous les autres contrôles Appliquer et Enregistrer, tout en conservant les sélections.

Enregistrer les transformations comme syntaxe. Cette option enregistre les transformations recommandées comme syntaxe de commande dans un fichier externe. La boîte de dialogue Préparation interactive des données ne contient pas ce contrôle car elle collera les transformations comme syntaxe de commande dans la fenêtre de syntaxe si vous cliquez sur **Coller**.

Enregistrer les transformations comme XML. Cette option enregistre les transformations recommandées au format XML dans un fichier externe, qui peut être fusionné avec le modèle PMML à l'aide de la commande `TMS MERGE` ou appliqué à un autre jeu de données à l'aide de la commande `TMS IMPORT`. La boîte de dialogue Préparation interactive des données ne contient pas ce contrôle car elle enregistrera les transformations au format XML si vous cliquez sur **Enregistrer XML** dans la barre d'outils au-dessus de la boîte de dialogue.

Onglet Analyse

Remarque : L'onglet Analyse est utilisé dans la boîte de dialogue Préparation interactive des données pour vous permettre de passer en revue les transformations. La boîte de dialogue de préparation automatique des données ne comprend pas cette étape.

1. Lorsque les paramètres d'ADP vous conviennent, y compris les modifications effectuées dans les onglets Objectif, Champs et Paramètres, cliquez sur **Analyser les données**. L'algorithme applique les paramètres aux entrées de données et affiche les résultats dans l'onglet Analyse.

L'onglet Analyse contient à la fois des sorties en tableaux et des sorties graphiques qui résument le traitement de vos données et affichent les recommandations sur la façon de modifier ou d'améliorer les données pour l'évaluation. Vous pouvez ensuite revoir puis accepter ou refuser ces recommandations.

L'onglet Analyse est composé de deux panneaux, la vue principale à gauche et la vue liée, ou auxiliaire, à droite. Il existe trois vues principales :

- Récapitulatif de traitement des champs (par défaut). Pour plus d'informations, voir «Récapitulatif de traitement des champs », à la page 13.
- Champs. Pour plus d'informations, voir «Champs », à la page 14.
- Récapitulatif des actions. Pour plus d'informations, voir «Récapitulatif des actions », à la page 15.

Il existe quatre vues liées/auxiliaires :

- Pouvoir prédictif (par défaut). Pour plus d'informations, voir «Pouvoir prédictif », à la page 15.
- Tableau des champs. Pour plus d'informations, voir «Tableau des champs », à la page 15.
- Détails des champs. Pour plus d'informations, voir «Détails des champs », à la page 16.
- Détails des actions. Pour plus d'informations, voir «Détails des actions », à la page 17.

Liens entre les vues

Dans la vue principale, le texte souligné dans les tableaux contrôle ce qui apparaît dans la vue liée. Si vous cliquez sur ces parties de texte, vous obtenez des détails sur un champ, un ensemble de champs ou une étape de traitement spécifique. Le lien que vous avez sélectionné en dernier apparaît en une couleur plus foncée qui permet d'identifier la connexion entre les contenus des deux panneaux de la vue.

Réinitialisation des vues

Pour afficher de nouveau les recommandations d'analyse d'origine et abandonner les modifications effectuées sur les vues Analyse, cliquez sur **Réinitialiser** au bas du panneau de la vue principale.

Récapitulatif de traitement des champs

La table récapitulative de traitement des champs fournit un instantané de l'impact du traitement général projeté, y compris les modifications de l'état des fonctions et le nombre de fonctions construites.

Veuillez noter que le modèle est bien construit, et que par conséquent il n'y a pas de mesure ou de graphique de la modification du pouvoir prédictif général avant et après la préparation des données. Par contre, vous pouvez afficher les graphiques du pouvoir prédictif des prédicteurs individuels recommandés.

Le tableau affiche les informations suivantes :

- le nombre de champs cible.
- Le nombre de prédicteurs (d'entrée) d'origine.
- Les prédicteurs recommandés pour l'analyse et la modélisation. Cela comprend le nombre total de champs recommandés ; le nombre de champs d'origine non transformés recommandés ; le nombre de champs transformés recommandés (sans les versions intermédiaires des champs, champs dérivés des prédicteurs de date/heure et prédicteurs construits) ; le nombre de champs dérivés recommandés des champs date/heure ; et le nombre de prédicteurs construits.
- Le nombre de prédicteurs d'entrée non recommandés quelle que soit leur forme, que ce soit sous leur forme d'origine, comme champ dérivé, ou comme entrée d'un prédicteur construit.

Lorsque des informations sur les **champs** sont soulignées, cliquez pour afficher plus de détails dans une vue liée. Les détails de la **Cible**, des **Fonctions d'entrée**, et des **Fonctions d'entrée non utilisées** apparaissent dans la vue liée Tableau des champs. Pour plus d'informations, voir la rubrique «Tableau des champs », à la page 15. **Les fonctions recommandées pour l'analyse** s'affichent dans la vue liée au pouvoir prédictif. Pour plus d'informations, voir «Pouvoir prédictif », à la page 15.

Champs

La vue principale Champs affiche les champs traités et si l'ADP recommande de les utiliser dans les modèles en aval. Vous pouvez ignorer les recommandations pour n'importe quel champ ; par exemple, exclure les fonctions construites ou inclure les fonctions que l'ADP recommande d'exclure. Si un champ a été transformé, vous pouvez décider d'accepter ou non la transformation suggérée ou d'utiliser ou non la version d'origine.

La vue Champs est composée de deux tableaux, un pour la cible et un pour les prédicteurs qui ont été traités ou créés.

Tableau cible

Le tableau **Cible** n'apparaît que si une cible est définie dans les données.

Elle contient deux colonnes :

- **Nom** : Nom ou libellé du champ cible ; le nom d'origine est toujours utilisé, même si le champ a été transformé.
- **Niveau de mesure.** Affiche l'icône représentant le niveau de mesure. Placez la souris sur l'icône pour afficher un libellé (continu, ordinal, nominal, etc.) qui décrit les données.

  Si la cible a été transformée, la colonne **Niveau de mesure** reflète la version transformée finale. *Remarque* : vous ne pouvez pas désactiver les transformations pour la cible.

Tableau des prédicteurs

Le tableau **Prédicteurs** est affiché en permanence. Chaque ligne du tableau représente un champ. Les lignes sont triées par défaut dans l'ordre décroissant du pouvoir prédictif.

Pour les fonctions ordinaires, le nom d'origine est toujours utilisé comme nom de ligne. Les versions d'origine et dérivée des champs date/heure apparaissent dans le tableau (dans des lignes séparées) ; le tableau contient également les prédicteurs construits.

Veuillez noter que les versions transformées des champs apparaissant dans le tableau représentent toujours les versions finales.

Par défaut, seuls les champs recommandés sont affichés dans le tableau des prédicteurs. Pour afficher les champs restants, sélectionnez la boîte de dialogue **Inclure les champs non recommandés dans le tableau** au-dessus du tableau ; ces champs sont ensuite affichés au bas du tableau.

La table contient les colonnes suivantes :

- **Version à utiliser.** Affiche une liste déroulante qui contrôle l'utilisation d'un champ en aval et s'il faut utiliser les transformations recommandées. Par défaut, la liste déroulante reflète les recommandations.

  Pour les prédicteurs ordinaires qui ont été transformés, la liste déroulante contient trois choix : **Transformée**, **Originale** et **Originale**.

  Pour les prédicteurs non transformés ordinaires, les choix sont : **Originale** et **Ne pas utiliser**.

  Pour les champs dérivés date/heure et les prédicteurs construits, les choix sont : **Transformée** et **Ne pas utiliser**.

  Pour les champs de date d'origine, la liste déroulante est désactivée et définie sur **Ne pas utiliser**.

  Remarque : Pour les prédicteurs contenant à la fois les versions d'origine et transformées, passer des versions **d'origine** aux versions **transformées** met automatiquement à jour les paramètres **Niveau de mesure** et **Pouvoir prédictif** pour ces fonctions.
- **Nom** : Chaque nom de champ est un lien. Cliquez sur un nom pour afficher plus d'informations sur le champ dans la vue liée. Pour plus d'informations, voir «Détails des champs », à la page 16.
- **Niveau de mesure.** Affiche l'icône représentant le type de données ; passez la souris sur l'icône pour afficher un libellé (continu, ordinal, nominal, etc.) qui décrit les données.
- **Pouvoir prédictif :** Le pouvoir prédictif est affiché uniquement pour les champs recommandés par l'ADP. Cette colonne n'apparaît pas si aucune cible n'est définie. Le pouvoir prédictif est compris entre 0 et 1, les valeurs les plus élevées, indiquant des prédicteurs de "meilleur" qualité. En général, le pouvoir prédictif est utile pour comparer les prédicteurs dans une analyse ADP, mais les valeurs du pouvoir prédictif ne peuvent être comparées entre des analyses différentes.

Récapitulatif des actions

Pour chaque action effectuée par la préparation automatique des données, les prédicteurs d'entrée sont transformées et/ou supprimées ; les champs qui survivent à une action sont utilisés à la suivante. Les champs qui survivent jusqu'à la dernière étape sont ensuite recommandés pour la modélisation, alors que les entrées des prédicteurs transformés et construits sont supprimés.

Le récapitulatif des actions est un simple tableau qui répertorie les actions effectuées par l'ADP. Lorsqu'une **Action** est soulignée, vous pouvez cliquer dessus pour afficher plus de détails sur les actions effectuées dans une vue liée. Pour plus d'informations, voir la rubrique «Détails des actions », à la page 17.

Remarque : Seules les versions d'origine et transformées finales de chaque champ sont affichées, et pas les versions intermédiaires utilisées pendant l'analyse.

Pouvoir prédictif

Affiché par défaut au début de l'analyse ou lorsque vous sélectionnez **Prédicteurs recommandés pour l'analyse** dans la vue principale Récapitulatif du traitement des champs, le graphique affiche le pouvoir prédictif des prédicteurs recommandés. Les champs sont triés par pouvoir prédictif, avec le champ ayant la plus haute valeur apparaissant en premier.

Pour les versions transformées des prédicteurs ordinaires, le nom des champs reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

Les icônes de niveau de mesure sont affichées après les noms de champ individuels.

Le pouvoir prédictif de chaque prédicteur recommandé est calculé à partir d'une régression linéaire ou d'un modèle bayésien naïf selon que la cible est continue ou catégorielle.

Tableau des champs

La vue Tableau des champs est un simple tableau qui répertorie les fonctions importantes et qui apparaît lorsque vous cliquez sur **Cible**, **Prédicteurs**, ou **Prédicteurs non utilisés** dans la vue principale Récapitulatif du traitement des champs.

Elle contient deux colonnes :

- **Nom** : Nom du prédicteur.

  Pour les cibles, le libellé ou le nom d'origine du champ est utilisé, même si la cible a été transformée.

  Pour les versions transformées des prédicteurs ordinaires, le nom reflète votre choix de suffixe dans le panneau Noms de champ de l'onglet Paramètres ; par exemple : *_transformed*.

  Pour les champs dérivés des dates et des heures, le nom de la version transformée finale est utilisé ; par exemple : *bdate_years*.

  Pour les prédicteurs construits, le nom du prédicteur construit est utilisé ; par exemple : *Predictor1*.
- **Niveau de mesure.** Affiche l'icône représentant le type de données.

  Pour la cible, le **Niveau de mesure** reflète toujours la version transformée (si la cible a été transformée), par exemple, changée d'ordinale (ensemble ordonné) à continue (plage, échelle) et vice versa.

Détails des champs

La vue Détails des champs contient les graphiques de distribution, des valeurs manquantes et du pouvoir prédictif (le cas échéant) pour le champ sélectionné et s'affiche lorsque vous cliquez sur un **Nom** de la vue principale Champs. De plus, l'historique du traitement pour le champ et le nom du champ transformé apparaissent également (le cas échéant).

Pour chaque ensemble de graphiques, deux versions apparaissent côte à côte pour comparer le champ avec et sans transformations appliquées ; si aucune version transformée du champ n'existe, un graphique apparaît pour la version d'origine uniquement. Pour les champs de date ou d'heure dérivés et les prédicteurs construits, les graphiques n'apparaissent que pour le nouveau prédicteur.

Remarque : Si un champ est exclu parce qu'il contient trop de catégories, seul l'historique de traitement apparaît.

Graphique de distribution

La distribution des champs continus apparaît dans un histogramme, avec une courbe normale superposée et une ligne de référence verticale pour la valeur moyenne ; les champs catégoriels apparaissent sous forme de graphique à barres.

Les histogrammes sont libellés pour montrer l'écart-type et l'asymétrie, toutefois l'asymétrie n'apparaît pas si le nombre des valeurs est inférieur ou égal à 2 ou si la variance du champ d'origine est inférieure à 10-20.

Passez la souris sur le graphique pour afficher la moyenne des histogrammes ou le nombre et le pourcentage du nombre total d'enregistrements des catégories dans les graphiques à barres.

Graphique des valeurs manquantes

Les graphiques circulaires comparent le pourcentage des valeurs manquantes avec et sans transformations appliquées ; les libellés de graphique indiquent le pourcentage.

Si l'ADP traite les valeurs manquantes, le graphique circulaire après la transformation comprend la valeur de remplacement comme libellé, c'est-à-dire la valeur utilisée à la place des valeurs manquantes.

Passez la souris sur le graphique pour afficher le nombre des valeurs manquantes et le pourcentage du nombre total d'enregistrements.

Graphique de pouvoir prédictif

Pour les champs recommandés, les graphiques à barres affichent le pouvoir prédictif avant et après la transformation. Si la cible a été transformée, le pouvoir prédictif calculé tient compte de la cible transformée.

Remarque : Les graphiques de pouvoir prédictif ne sont pas affichés si aucune cible n'est définie, ou si la cible est atteinte depuis le panneau de la vue principale.

Passez la souris sur le graphique pour afficher la valeur du pouvoir prédictif.

Tableau des historiques du traitement

Ce tableau indique la façon dont la version transformée d'un champ a été dérivée. Les actions entreprises par l'ADP sont répertoriées dans l'ordre dans lequel elles ont été exécutées ; mais, pour certaines étapes, plusieurs actions ont pu être exécutées pour un champ particulier.

Remarque : Ce tableau n'apparaît pas pour les champs qui n'ont pas été transformés.

Les informations du tableau sont divisées en deux ou trois colonnes :

- **Action :** Le nom de l'action. Par exemple, Prédicteurs continus. Pour plus d'informations, reportez-vous à la rubrique «Détails des actions », à la page 17.
- **Détails :** La liste des traitements effectués. Par exemple, Transformer en unités standard.
- **Fonction :** Apparaît uniquement pour les prédicteurs construits et affiche la combinaison linéaire de champs d'entrée, par exemple, 0,06*âge + 1,21*hauteur.

Détails des actions

La vue liée Détails des actions apparaît lorsque vous cliquez sur **Action** dans la vue principale Récapitulatif des actions.La vue liée Détails des actions affiche des informations relatives aux actions et des informations communes pour chaque étape de traitement effectuée. Les détails relatifs à chaque action spécifique apparaissent d'abord.

La description de chaque action est utilisée comme titre en haut de la vue liée. Les détails relatifs à chaque action sont affichés sous le titre, et peuvent contenir des détails sur le nombre de prédicteurs dérivés, de champs reconvertis, de transformations de cible, de catégories fusionnées ou réorganisées et de prédicteurs construits ou exclus.

Au cours du traitement des actions, le nombre de prédicteurs utilisés pour le traitement peut varier, par exemple lorsque des prédicteurs sont exclus ou fusionnés.

Remarque : Si une action est désactivée ou qu'aucune cible n'est spécifiée, un message d'erreur apparaît à la place des détails de l'action lorsque vous cliquez sur l'action dans la vue principale Récapitulatif des actions.

Il existe neuf actions possibles, toutefois, toutes ne sont pas nécessairement actives pour chaque analyse.

tableau Champs de texte

Ce tableau affiche le nombre :

- Prédicteurs exclus de l'analyse.

Tableau Prédicteurs de date et d'heure

Ce tableau affiche le nombre :

- Durées dérivées des prédicteurs de date et d'heure.
- d'éléments Date et heure.
- Prédicteurs de date et d'heure dérivées, au total.

La date ou heure de référence est affichée comme note de bas de page si des durées de date ont été calculées.

Tableau Balayage des prédicteurs

Ce tableau affiche le nombre des prédicteurs suivants exclus du traitement :

- Constantes.
- Prédicteurs avec trop de valeurs manquantes.
- Prédicteurs avec trop d'observations dans une seule catégorie.
- Champs nominaux (ensembles) avec trop de catégories.
- Prédicteurs supprimés, au total.

Tableau Vérifier le niveau de mesure

Ce tableau affiche le nombre de champs reconvertis, répartis selon les catégories suivantes :

- Champs ordinaux (ensembles ordonnés) reconvertis en champs continus.
- Champs continus reconvertis en champs ordinaux.
- Nombre total des champs reconvertis.

Si aucun champ d'entrée (cible ou de prédicteurs) n'est un ensemble continu ou ordinal, cela apparaît en note de bas de page.

Tableau Valeurs extrêmes

Ce tableau affiche le nombre de valeurs extrêmes traitées.

- soit le nombre de champs continus pour lesquels des valeurs extrêmes ont été recherchées et tronquées, ou le nombre de champs continus pour lesquels les valeurs extrêmes ont été recherchées et définies sur manquantes, en fonction de vos paramètres dans le panneau Préparer les entrées & la cible dans l'onglet Paramètres.
- Le nombre de champs continus exclus parce qu'ils étaient constants après le traitement des valeurs extrêmes.

Une note de bas de page indique la valeur de césure des valeurs extrêmes et une autre note de bas de page apparaît si aucun champ d'entrée (cible ou de prédicteurs) n'est continu.

Tableau Valeurs manquantes

Ce tableau affiche le nombre de champs qui contenaient des valeurs manquantes remplacées, selon les catégories suivantes :

- Cible. Cette ligne n'apparaît pas si aucune cible n'est spécifiée.
- Prédicteurs. Elles sont divisées en nombre de champs nominaux (ensemble), ordinaux (ensemble ordonné) et continus.
- Le nombre total de valeurs manquantes remplacées.

Table cible

Ce tableau indique si la cible a été transformée :

- Transformation de Box-Cox en normalité. Cette catégorie est elle-même divisée en colonnes qui indiquent le critère spécifié (moyenne et écart type) et le Lambda.
- Catégories cibles réorganisées pour améliorer la stabilité.

Tableau prédicteurs catégoriels

Ce tableau affiche le nombre de prédicteurs catégoriels :

- dont les catégories ont été réorganisées de la plus faible la plus élevée pour améliorer la stabilité.
- dont les catégories ont été fusionnées pour optimiser l'association avec la cible.
- dont les catégories ont été fusionnées pour traiter les catégories éparpillées.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après la fusion.

Une note de bas de page apparaît si aucun prédicteur catégoriel n'existe.

Tableau Prédicteurs continus

Il existe deux tableaux. Le premier affiche une des transformations suivantes :

- Les valeurs des prédicteurs transformées en unités standard. De plus, il indique le nombre de prédicteurs transformés, la moyenne spécifiée et l'écart type.
- Les valeurs des prédicteurs mappées sur une plage commune. De plus, il indique le nombre de prédicteurs transformés utilisant une transformation min-max, ainsi que les valeurs minimum et maximum spécifiées.

- Les valeurs des prédicteurs et le nombre de prédicteurs regroupés.

Le deuxième tableau affiche les détails de construction de l'espace des prédicteurs, sous la forme du nombre de prédicteurs :

- construites.
- exclues en raison d'une faible association avec la cible.
- exclues parce qu'elles étaient constantes après le regroupement.
- exclues parce qu'elles étaient constantes après la construction.

Une note de bas de page apparaît si aucun prédicteur continu n'a été saisi.

Rétablissement des scores

Si une cible a été transformée par l'ADP, les modèles en résultant créés à l'aide de la cible transformée évaluent les unités transformées. Afin d'interpréter et d'utiliser les résultats, vous devez reconvertir la valeur observée dans son échelle d'origine.

1. Pour rétablir les scores, dans les menus, choisissez :

   Transformation > **Préparation des données pour la modélisation** > **Rétablissement des scores...**
2. Sélectionnez un champ à rétablir. Ce champ doit contenir des valeurs prévues par le modèle de la cible transformée.
3. Spécifiez un suffixe pour le nouveau champ. Ce nouveau champ contiendra des valeurs prévues par le modèle à l'échelle d'origine de la cible non transformée.
4. Spécifiez l'emplacement du fichier XML contenant les transformations de l'ADP. Ce doit être un fichier enregistré à partir des boîtes de dialogue Préparation automatique ou interactive des données. Pour plus d'informations, voir «Application et enregistrement des transformations », à la page 12.

Identification des observations inhabituelles

La procédure de détection des anomalies vise à repérer les observations inhabituelles en se basant sur les écarts par rapport aux normes de leurs groupes de clusters. La procédure est destinée à détecter rapidement les observations inhabituelles afin de vérifier les données à l'étape d'analyse exploratoire des données, avant d'effectuer toute sorte d'analyse inférentielle de ces mêmes données. Cet algorithme sert à détecter des anomalies générales. Il est vrai que la définition d'une observation anormale ne s'applique pas à tous les secteurs. Par exemple, la définition d'une anomalie peut être clairement définie lorsqu'il s'agit de détecter des moyens de paiements inhabituels dans l'industrie pharmaceutique ou du blanchissement d'argent dans l'industrie bancaire.

Exemple : Un analyste de données employé pour construire des modèles capables de prédire les résultats obtenus suite au traitement d'attaques cardiaques cherche des données de qualité, car de tels modèles sont sensibles aux observations inhabituelles. Certaines de ces observations éloignées sont des observations tout à fait uniques et s'avèrent donc inexploitables en matière de prédiction, alors que d'autres sont dues à des erreurs de saisie de données dans lesquelles les valeurs sont techniquement « correctes » sans pouvoir toutefois être prises en compte par les procédures de validation de données. La procédure d'identification des observations inhabituelles sert à identifier ces valeurs extrêmes et à en dresser la liste afin que l'analyste puisse décider de la manière de les traiter.

Statistiques : La procédure génère des groupes d'homologues, des normes de groupes d'homologues pour des variables continues et catégorielles, des indices d'anomalies basés sur les écarts par rapport aux normes de groupes d'homologues, ainsi que des valeurs d'impact de variables pour les variables contribuant le plus à une observation considérée comme inhabituelle.

Analyse des données

Données : Cette procédure fonctionne avec des variables continues et catégorielles. Chaque ligne représente une observation distincte tandis que chaque colonne représente une variable différente sur laquelle les groupes d'homologues sont basés. Une variable d'identification d'observations est disponible

dans le fichier de données pour marquer les sorties, mais elle ne sera pas utilisée dans l'analyse. Les valeurs manquantes sont autorisées. La variable de pondération est ignorée, si indiquée auparavant.

Le modèle de détection peut être appliqué à un nouveau fichier de données de test. Les éléments des données du test doivent être identiques aux éléments contenus dans les données d'apprentissage. Et, en fonction des paramètres d'algorithme, le traitement de la valeur manquante utilisé pour créer le modèle doit être appliqué au fichier de données de test avant d'effectuer une évaluation.

Tri par observation : Notez que la solution peut dépendre de l'ordre des observations. Pour réduire les effets de tri, classez les observations de manière aléatoire. Pour vérifier la stabilité d'une solution donnée, vous pouvez obtenir différentes solutions dans lesquelles les observations sont triées de différentes manières aléatoires. Si les fichiers sont très volumineux, vous pouvez effectuer plusieurs fois l'opération sur un échantillon des observations triées de différentes manières aléatoires.

Hypothèses : L'algorithme suppose que toutes les variables sont non constantes et indépendantes, et qu'aucune observation ne possède de valeur manquante pour les variables d'entrée. Chaque variable continue est considérée comme ayant une distribution normale (gaussienne) et chaque variable catégorielle comme ayant une distribution multinomiale. Des tests internes empiriques indiquent que la procédure est assez résistante aux violations de l'hypothèse d'indépendance et des hypothèses de distribution, mais vous devez savoir comment ces hypothèses sont vérifiées.

Pour identifier les observations inhabituelles

1. A partir des menus, sélectionnez :

   Données > **Identification des observations inhabituelles...**
2. Sélectionnez au moins une variable d'analyse.
3. Vous pouvez également sélectionner une variable d'identificateur d'observation à utiliser pour le libellé de la sortie.

Champs dont le niveau de mesure est inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou de plusieurs variables (champs) du jeu de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Analyser les données : Lit les données dans le jeu de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si le jeu de données est volumineux, l'opération peut prendre un certain temps.

Affecter manuellement : Ouvre une boîte de dialogue dans laquelle figurent tous les champs dont le niveau de mesure est inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans la vue de variable de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Identification de la sortie d'observations inhabituelles

Liste des observations inhabituelles et des raisons pour lesquelles elles sont considérées comme inhabituelles : Cette option propose trois tableaux :

- La liste d'index des observations présentant une anomalie affiche les observations identifiées comme étant inhabituelles, ainsi que leur valeur d'index d'anomalie correspondante.
- La liste d'ID des pairs d'observation présentant une anomalie affiche les observations inhabituelles ainsi que les informations relatives à leur groupe d'homologues correspondant.
- La liste des raisons expliquant les anomalies affiche le numéro de l'observation, la variable de raison, la valeur d'impact de la variable, la valeur de la variable et la norme de la variable pour chaque raison.

Tous les tableaux sont triés par index d'anomalie en ordre décroissant. De plus, les ID des observations ne sont affichés que si la variable d'identificateur de l'observation est indiquée dans l'onglet Variables.

Principales statistiques : Les contrôles de ce groupe génèrent des récapitulatifs de distribution.

- **Normes de groupes d'homologues :** Cette option affiche le tableau des normes de variables continues (en cas d'utilisation de variables continues dans l'analyse) et le tableau des normes de variables catégorielles (en cas d'utilisation de variables catégorielles dans l'analyse). Le tableau des normes de variables continues affiche la moyenne et l'écart type de chaque variable continue pour chaque groupe d'homologues. Le tableau des normes de variables catégorielles affiche le mode (catégorie la plus utilisée), sa fréquence et le pourcentage de fréquence de chaque variable catégorielle pour chaque groupe d'homologues. La moyenne d'une variable continue et le mode d'une variable catégorielle sont utilisés comme les valeurs standard dans l'analyse.
- **Indices d'anomalies :** Le récapitulatif de l'index d'anomalie affiche les statistiques descriptives pour l'index d'anomalie des observations identifiées comme étant les plus inhabituelles.
- **Occurrence de raisons par variable d'analyse :** Pour chaque raison, le tableau affiche la fréquence et le pourcentage de fréquence de chaque occurrence de variable exprimé sous la forme d'une raison. Le tableau indique également les statistiques descriptives de l'observation de chaque variable. Si le nombre de raisons maximal est défini sur 0 dans l'onglet Options, cela signifie que cette option n'est pas disponible.
- **Observations traitées :** Le récapitulatif du traitement des observations affiche les effectifs et les pourcentages d'effectif pour toutes les observations dans un jeu de données actif, les observations incluses et exclues de l'analyse, et les observations de chaque groupe d'homologues.

Identification des enregistrements d'observations inhabituelles

Enregistrer les variables : Les contrôles de ce groupe vous permettent d'enregistrer des variables de modèle dans le jeu de données actif. Vous pouvez également choisir de remplacer les variables existantes dont le nom est en conflit avec les variables à enregistrer.

- **Indices d'anomalies :** Enregistre la valeur de l'index d'anomalie pour chaque observation dans une variable portant le nom indiqué.
- **Groupes d'homologues :** Enregistre l'ID du groupe d'homologues, le nombre d'observations et la taille en tant que pourcentage pour chaque observation dans les variables portant le nom de racine spécifié. Par exemple, si le nom de racine *Peer* est spécifié, les variables *Peerid*, *PeerSize* et *PeerPctSize* sont générées. *Peerid* est l'ID du groupe d'homologues de l'observation, *PeerSize* la taille du groupe et *PeerPctSize* la taille du groupe exprimée en pourcentage.
- **Raisons :** Enregistre les ensembles de variables de raison portant le nom de racine spécifié. Un ensemble de variables de raison est composé du nom de la variable en tant que raison, de sa mesure de l'impact de la variable, de sa propre valeur et de la valeur standard. Le nombre d'ensembles dépend du nombre de raisons demandées dans l'onglet Options. Par exemple, si le nom de racine *Reason* est spécifié, les variables *ReasonVar_k* , *ReasonMeasure_k*, *ReasonValue_k* et *ReasonBorm_k* sont alors générées, *k* correspondant à la raison *k*. Cette option n'est pas disponible si le nombre des raisons est défini sur 0.

Exporter un fichier de modèle : Cette option vous permet d'enregistrer le modèle au format XML.

Identification des valeurs manquantes des observations inhabituelles

L'onglet Valeurs manquantes sert à contrôler le traitement des valeurs manquantes spécifiées par l'utilisateur et les valeurs système manquantes.

- **Exclure les valeurs manquantes de l'analyse :** Les observations contenant des valeurs manquantes sont exclues de l'analyse.
- **Inclure les valeurs manquantes dans l'analyse :** Les valeurs manquantes des variables continues sont remplacées par leur moyenne générale correspondante, et les catégories manquantes des variables catégorielles sont groupées et traitées en tant que catégorie valide. Les variables traitées sont ensuite utilisées dans l'analyse. Vous pouvez également demander la création d'une variable supplémentaire représentant la proportion de variables manquantes dans chaque observation et utiliser cette variable dans l'analyse.

Options d'identification des observations inhabituelles

Critères d'identification des observations inhabituelles : Ces sélections déterminent le nombre d'observations à inclure dans la liste d'anomalies.

- **Pourcentage d'observations ayant les valeurs d'index d'anomalie les plus élevées :** Indiquez un nombre positif inférieur ou égal à 100.
- **Nombre fixe d'observations ayant les valeurs d'index d'anomalie les plus élevées :** Indiquez un entier positif inférieur ou égal au nombre total d'observations contenues dans le jeu de données actif et utilisées dans l'analyse.
- **Identifiez les observations dont la valeur d'index d'anomalie atteint ou dépasse une valeur minimum uniquement :** Spécifiez un nombre non négatif. Une observation est considérée comme anormale si la valeur d'index d'anomalie est supérieure ou égale à la limite d'inclusion spécifiée. Cette option est employée avec les options **Pourcentage d'observations** et **Nombre fixe d'observations**. Par exemple, si vous spécifiez un nombre fixe de 50 observations et une valeur de césure de 2, la liste d'anomalie sera composée de 50 observations au moins, chaque observation aura une valeur d'index d'anomalie supérieure ou égale à 2.

Nombre de groupes d'homologues : La procédure cherche le meilleur nombre de groupes d'homologues compris entre la valeur minimum et la valeur maximum spécifiées. Les valeurs doivent être des entiers positifs dont la valeur minimum ne doit pas dépasser la valeur maximum. Lorsque les valeurs spécifiées sont égales, la procédure part du principe que le nombre de groupes d'homologues est fixe.

Remarque : En fonction de la variance de vos données, il peut arriver que le nombre de groupes d'homologues pris en charge par les données soit inférieur au nombre spécifié comme valeur minimum. Dans une telle situation; la procédure risque d'engendrer un nombre de groupes d'homologues plus petit.

Nombre maximum de raisons : Une raison est constituée de la mesure de l'impact d'une variable, du nom de la variable pour cette raison, de la valeur de la variable et de la valeur du groupe d'homologues correspondant. Spécifiez un nombre entier non-négatif. Si cette valeur égale ou dépasse le nombre de variables traitées qui sont ensuite utilisées dans l'analyse, les variables sont alors affichées.

Fonctions supplémentaires de la commande DETECTANOMALY

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Omettre de l'analyse quelques variables du jeu de données actif sans indiquer de façon explicite toutes les variables d'analyse (à l'aide de la sous-commande EXCEPT).
- Spécifier un ajustement pour équilibrer l'influence des variables continues et catégorielles (à l'aide du mot-clé `MLWEIGHT` de la sous-commande `CRITERIA`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Regroupement optimal

La procédure Regroupement optimal discrétise une ou plusieurs variables d'échelle (désormais appelées **variables d'entrée de regroupement**) en distribuant les valeurs de chaque variable dans des casiers. La formation de casiers est optimale par rapport à une variable guide catégorielle qui « supervise » le regroupement par casiers. Les casiers peuvent ensuite être utilisés à la place des valeurs de données d'origine pour de plus amples analyses.

Exemples : La réduction du nombre de valeurs distinctes que prend une variable a un certain nombre d'utilisations :

- Les données requises d'autres procédures. Les variables discrétisées peuvent être traitées comme catégorielles lors d'une utilisation dans des procédures faisant appel à ce type de variable. Par exemple, la procédure Tableaux croisés nécessite que toutes les variables soient catégorielles.
- Confidentialité des données. Signaler des valeurs regroupées par casiers au lieu des valeurs réelles aide à protéger la confidentialité de vos sources de données. La procédure Regroupement optimal peut guider le choix des casiers.

- Performances en matière de vitesse. Certaines procédures sont plus efficaces lorsque vous travaillez avec un nombre réduit de valeurs distinctes. Par exemple, la vitesse de la régression logistique multinomiale peut être améliorée grâce à l'utilisation de variables discrétisées.
- Révélation de la séparation complète ou quasi complète des données.

Recodage supervisé optimal et regroupement visuel : Les boîtes de dialogue Regroupement visuel proposent plusieurs méthodes automatiques de création de casiers sans utiliser de variable guide. Ces règles "non supervisées" sont utiles pour générer des statistiques descriptives, telles que des tables de fréquences, mais le recodage supervisé optimal donne de meilleurs résultats si votre objectif final est de générer un modèle de prévision.

Sortie : La procédure génère des tableaux de divisions pour les casiers et les statistiques descriptives de chaque variable d'entrée de regroupement. En outre, vous pouvez enregistrer de nouvelles variables dans le jeu de données actif contenant les valeurs regroupées par casiers des variables d'entrée de regroupement et enregistrer les règles de regroupement comme syntaxe de commande pour les utiliser dans la discrétisation de nouvelles données.

Remarques sur les données de recodage supervisé optimal

Données : Cette procédure exige que les variables d'entrée de regroupement soient des variables d'échelle numériques. La variable guide doit être catégorielle et peut être chaîne ou numérique.

Obtention du regroupement optimal

1. A partir des menus, sélectionnez :

   Transformer > **Regroupement optimal...**
2. Sélectionnez une ou plusieurs variables d'entrée de regroupement.
3. Sélectionnez une variable guide.

Les variables contenant les valeurs des données regroupées par casiers ne sont pas générées par défaut. Utilisez l'onglet Enregistrer pour enregistrer ces variables.

Sortie du recodage supervisé optimal

L'onglet Sortie contrôle l'affichage des résultats.

- **Extrema pour les casiers :** Affiche l'ensemble des points finaux pour chaque variable d'entrée de regroupement.
- **Statistiques descriptives pour les variables regroupées :** Pour chaque variable d'entrée de regroupement, cette option affiche le nombre d'observations dotées de valeurs valides, le nombre d'observations dotées de valeurs manquantes, le nombre de valeurs valides distinctes, et les valeurs minimale et maximale. Pour la variable guide, cette option affiche la distribution de classe pour chaque variable d'entrée de regroupement liée.
- **Entropie de modèle pour les variables regroupées :** Pour chaque variable d'entrée de regroupement, cette option affiche une mesure de l'exactitude des prévisions de la variable par rapport à la variable guide.

Enregistrement du recodage supervisé optimal

Enregistrer les variables dans le jeu de données actif : Les variables contenant les valeurs de données regroupées par casiers peuvent se substituer aux variables d'origine pour une analyse ultérieure.

Enregistrer les règles de regroupement en tant que syntaxe : Génère une syntaxe de commande qui peut être utilisée pour regrouper d'autres jeux de données par casiers. Les règles de recodage sont basées sur les divisions déterminées par l'algorithme de regroupement par casiers.

Valeurs manquantes de recodage supervisé optimal

L'onglet Valeurs manquantes spécifie si les valeurs manquantes sont gérées par la suppression des observations incomplètes ou des composantes non valides seulement. Les valeurs manquantes de

l'utilisateur sont toujours traitées comme non valides. Lors du recodage des valeurs de variable d'origine en nouvelle variable, les valeurs manquantes de l'utilisateur sont converties en valeurs système manquantes.

- **Seulement composantes non valides :** Cette option concerne chaque paire de variables guide et d'entrée de regroupement. La procédure utilise toutes les observations ayant des valeurs non manquantes sur la variable guide et d'entrée de regroupement.
- **Toute observation incomplète :** Cette option concerne toutes les variables spécifiées dans l'onglet Variables. Si une variable est manquante pour une observation, l'observation est intégralement exclue.

Options Regroupement optimal

Prétraitement : Les variables d'entrée de "pré-regroupement" dotées de nombreuses valeurs distinctes améliorent le temps de traitement sans altérer la qualité des casiers finaux. Le nombre maximal de casiers fournit la limite supérieure du nombre de casiers créés. Ainsi, si vous spécifiez 1 000 comme maximum, mais qu'une variable d'entrée de regroupement possède moins de 1 000 valeurs distinctes, le nombre de casiers prétraités créés pour la variable d'entrée de regroupement sera égal au nombre de valeurs distinctes dans cette variable.

Casiers faiblement remplis : La procédure génère parfois des casiers avec très peu d'observations. La stratégie suivante supprime ces pseudo-divisions :

Pour une variable donnée, supposons que l'algorithme ait trouvé n divisions $_{\text{finales}}$ et donc n+1 casiers $_{\text{finaux}}$. Pour les casiers i = 2, ..., n $_{\text{final}}$ (compris entre le deuxième casier avec la plus faible valeur et le deuxième casier avec la valeur la plus élevée), calculez

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

où *taillede(b)* est le nombre d'observations dans le casier.

Lorsque cette valeur est inférieure au seuil de fusion spécifié, b_i est considéré comme faiblement peuplé et fusionné avec b_{i-1} ou b_{i+1}, en fonction de la valeur possédant la plus faible entropie d'informations de classe.

La procédure effectue un seul passage à travers les casiers.

Points finaux des casiers : Cette option indique comment la limite inférieure d'un intervalle est définie. Puisque la procédure détermine automatiquement les valeurs des divisions, il s'agit surtout d'une question de préférence.

Premier casier (le moins élevé)/Dernier casier (le plus élevé) : Ces options spécifient comment les divisions minimal et maximal de chaque variable d'entrée de regroupement sont définis. Généralement, la procédure suppose que les variables d'entrée de regroupement peuvent prendre n'importe quelle valeur sur la ligne des nombres réels, mais si vous avez une raison théorique ou pratique de limiter la plage, vous pouvez le faire sur la base des valeurs les plus faibles/les plus élevées.

Fonctions supplémentaires de la commande OPTIMAL BINNING

Le langage de syntaxe de commande vous permet également d'effectuer les actions suivantes :

- Effectuer un recodage non supervisé à l'aide de la méthode d'effectifs égaux (via la sous-commande `CRITERIA`).

Reportez-vous au manuel *Command Syntax Reference* pour plus d'informations sur la syntaxe.

Remarques

Le présent document a été développé pour des produits et des services proposés aux Etats-Unis. et peut être mis à disposition par IBM dans d'autres langues. Vous pouvez toutefois devoir détenir une copie du produit ou une version du produit dans cette langue pour pouvoir y accéder.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Contactez votre interlocuteur IBM pour plus d'informations sur les produits et services disponibles dans votre région. Toute référence à un produit, logiciel ou service IBM n'implique pas que seul ce produit, logiciel ou service IBM puisse être utilisé. Tout autre élément fonctionnellement équivalent peut être utilisé, s'il n'enfreint aucun droit d'IBM. Il est de la responsabilité de l'utilisateur d'évaluer et de vérifier lui-même les installations et applications réalisées avec des produits, logiciels ou services non expressément référencés par IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits décrits dans le présent document. La remise de ce document ne vous accorde aucun droit de licence sur ces brevets ou demandes de brevet. Si vous désirez recevoir des informations concernant l'acquisition de licences, veuillez en faire la demande par écrit à l'adresse suivante :

IBM Director of Licensing

IBM Corporation

North Castle Drive, MD-NC119Armonk, NY 10504-1785U.S.A.

Les informations sur les licences concernant les produits IBM utilisant un jeu de caractères double octet peuvent être obtenues par écrit à l'adresse suivante :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-kuTokyo 103-8510, Japon

LE PRESENT DOCUMENT EST LIVRE "EN L'ETAT" SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties implicites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Le présent document peut contenir des inexactitudes ou des coquilles. Il est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, apporter des améliorations et des modifications aux produits et aux logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci) et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing

IBM Corporation

North Castle Drive, MD-NC119Armonk, NY 10504-1785U.S.A.

Ces informations peuvent être soumises à des conditions particulières, prévoyant notamment le paiement d'une redevance.

Le logiciel sous licence décrit dans ce document et tous les éléments sous licence disponibles pour ce dernier sont fournis par IBM d'après les termes du contrat client IBM, des conditions internationales d'utilisation de logiciels IBM ou de tout contrat équivalent entre nous.

Les données de performances et les exemples de clients ne sont présentés qu'à des fins d'illustration. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins d'illustration ou d'explication uniquement. Toute ressemblance avec des noms de personnes et de sociétés serait purement fortuite.

LICENCE DE COPYRIGHT :

Le présent logiciel contient des exemples de programmes d'application en langage source destinés à illustrer les techniques de programmation sur différentes plateformes d'exploitation. Vous avez le droit de copier, de modifier et de distribuer ces exemples de programmes sous quelque forme que ce soit et sans paiement d'aucune redevance à IBM, à des fins de développement, d'utilisation, de vente ou de distribution de programmes d'application conformes aux interfaces de programmation des plateformes pour lesquelles ils ont été écrits ou aux interfaces de programmation IBM. Ce exemples n'ont pas été testés en détails, ni dans toutes les conditions. C'est pourquoi IBM ne peut pas garantir ou assurer la fiabilité, la serviçabilité ou le fonctionnement de ces programmes. Les modèles de programmes sont fournis "en l'état", sans garantie d'aucune sorte. IBM ne sera en aucun cas responsable des dommages liés à l'utilisation de ces programmes.

Toute copie totale ou partielle de ces programmes exemples et des oeuvres qui en sont dérivées doit comprendre une notice de copyright, libellée comme suit :

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corp., dans de nombreux pays. Les autres noms de services et de produits peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques d'Adobe Systems Incorporated aux Etats-Unis et/ou dans d'autres pays.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques d'Intel Corporation ou de ses filiales aux Etats-Unis et dans certains autres pays.

Linux est une marque de Linus Torvalds aux Etats-Unis et/ou dans certains autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques de Microsoft Corporation aux Etats-Unis et/ou dans certains autres pays.

UNIX est une marque enregistrée de The Open Group aux Etats-Unis et/ou dans certains autres pays.

Java ainsi que tous les logos et toutes les marques incluant Java sont des marques d'Oracle et/ou de ses sociétés affiliées.

Index

S

T

V