SAS® Enterprise Miner™

Développez une connaissance plus précise avec un processus data mining plus productif

La transformation de données brutes en informations utiles reste une problématique pour les entreprises. Pour apporter des réponses et acquérir un réel avantage concurrentiel, il faut des solutions d'analyse performantes permettant de fouiller et de traiter des montagnes de données.

La mise en évidence de modèles et de schémas jusqu'alors inconnus permet aux décideurs d'instaurer des stratégies efficaces. Et d'aborder de manière différenciante des questions telles que : « quel client achètera quel produit, à quel moment et sur quel canal ? Quelle offre proposer pour retenir ceux qui ont des velléités de passer à la concurrence ? Sur quels critères fixer les tarifs pour optimiser les marges ? Quelle est l'incidence des calendriers de maintenance sur la durée de bon fonctionnement d'un composant ?

Ceux qui font le choix d'intégrer le data mining à leurs processus métier se donnent les moyens d'être compétitifs sur des marchés en pleine évolution.

En l'absence de méthodes et d'outils adaptés, l'analyse de grands volumes de données se révèle laborieuse.

Et si une méthode analytique peut parfaitement convenir à un ensemble de données en particulier, cette même méthode pourra se révéler inadaptée face à de nouvelles sources de données ou pour traiter de nouveaux enjeux métier.

Il est par conséquent crucial de disposer d'une solution offrant un large éventail de méthodes analytiques. Chaque méthode génère un modèle différent et c'est en comparant ces modèles que vous identifierez l'approche la plus adaptée. Si votre palette comporte un nombre restreint de méthodes (régressions ou arbres de décision seuls, par exemple), les capacités prédictives du modèle obtenu seront limitées voire inadaptées à votre problématique.

Face à la demande croissante d'informations analytiques immédiatement exploitables dans chaque secteur d'activité, les spécialistes du data mining et les analystes n'ont d'autre choix que de produire davantage de modèles de meilleure qualité dans des délais records.

Que fait SAS Enterprise Miner?

SAS Enterprise Miner industrialise le processus de data mining pour définir des modèles prédictifs et des segmentations avec une productivité inégalée. Vous pourrez ainsi intégrer absolument toutes les sources de données de votre organisation dans vos études, même les plus volumineuses, pour obtenir des modèles plus pertinents et plus précis. Les entreprises utilisent aujourd'hui SAS Enterprise Miner pour des problématiques stratégiques : améliorer les taux de réponse des campagnes marketing, minimiser l'attrition client, réduire les risques en terme de crédit, anticiper la demande, réduire le temps d'immobilisation des actifs, détecter les fraudes

Pourquoi SAS Enterprise Miner joue-t-il un rôle important ?

SAS propose la suite de méthodologies d'analyses prédictives la plus complète du marché ainsi que des fonctions interactives de visualisation. Elle permet aux utilisateurs d'explorer et d'exploiter les données efficacement et de créer une plus-value décisionnelle stratégique métier.

A qui SAS Enterprise Miner est-il destiné ?

SAS Enterprise Miner s'adresse à tous ceux qui doivent appréhender et analyser des volumes conséquents de données pour identifier et résoudre des problématiques métier ou de recherche et prendre rapidement les bonnes décisions : les data miners, les statisticiens, les analystes, les scientifiques, etc.

Intégrer l'analytique de manière sécurisée et évolutive facilite la résolution des problématiques fonctionnelles et sectorielles. Cela nécessite d'une part une collaboration étendue au sein de l'entreprise et d'autre part une solution de data mining multifonction performante, s'adaptant à la diversité des besoins.

Grâce à l'architecture optimisée de SAS Enterprise Miner, les spécialistes du data mining disposent de plus de temps pour créer des modèles prédictifs et descriptifs extrêmement précis. Les résultats peuvent être partagés dans toute l'entreprise pour diffuser les informations analytiques et intégrer ces modèles aux processus métier.

SAS Enterprise Miner repose sur le socle SAS® Business Analytics. Socle commun à l'ensemble des solutions. S'appuyer sur cette plate-forme décisionnelle assure une vision cohérente au sein de l'entreprise.

Principaux atouts

- **Identifier les principales relations entre variables et développer des modèles de manière rapide et intuitive.** Les utilisateurs, analystes ou experts métier, collaborent et interagissent aisément avec les informations via l'interface graphique de SAS Enterprise Miner et ce à tous les stades du cycle de modélisation. Les liaisons dynamiques entre les tables et les graphiques dans l'environnement interactif sont optimisées pour la découverte et l'exploration visuelle de la donnée. Elles permettent de cerner plus facilement les relations.
- **Créer plus efficacement des modèles de meilleure qualité grâce à la souplesse de l'environnement de travail.** SAS Enterprise Miner inclut des diagrammes de flux de processus auto-documentés qui réduisent considérablement la phase

de développement des modèles. Sa cartographie efficace du processus de data mining garantit une lecture rapide des résultats et une compréhension immédiate des techniques mises en œuvre.

- **Tirer rapidement et aisément les enseignements indispensables à vos prises de décision, de manière autonome et automatisée.** Avec SAS Rapid Predictive Modeler, fonctionnalité développée par SAS à partir de SAS Enterprise Miner, les analystes et experts métier, sans qualification statistique avancée, peuvent générer et gérer des modèles prédictifs pour les scénarii métier les plus courants. Les résultats sont présentés sous forme de graphiques et de tableaux simples à interpréter.
- **Affiner les prédictions pour prendre systématiquement les bonnes décisions et les mesures adéquates.** Des modèles plus performants utilisant des algorithmes innovants, y compris des méthodes sectorielles, génèrent des prédictions plus précises et plus stables. Ces prédictions sont facilement validées par des indicateurs et une évaluation graphique des modèles. Les résultats prédictifs et les statistiques d'évaluation de modèles basés sur des approches différentes peuvent être comparés en vis à vis. Les diagrammes obtenus peuvent être actualisés facilement et réutilisés en tant que modèle pour résoudre de nouvelles problématiques, permettant ainsi de capitaliser sur des modèles déjà construits. Etablir les profils des modèles permet également de mieux cerner le rôle des variables explicatives dans la modélisation.
- **Simplifier le déploiement de vos modèles et de vos scores pour des résultats encore plus rapides.** SAS Enterprise Miner peut automatiser le scoring de nouvelles données et fournit, à chaque stade du développement de modèles, un script de scoring complet en langages SAS, C, Java ou PMML. Ce script peut être déployé dans de multiples environnements, temps réel ou batch, dans SAS, sur le web, au sein de processus métier ou directement dans des bases de données relationnelles (traitement à l'intérieur de la base). Résultats : les temps de traitement sont considérablement réduits, vous obtenez des résultats précis, en diminuant les phases de recodage et les décisions sont plus efficaces.

Présentation du produit

SAS Enterprise Miner repose sur un système client-serveur moderne et distribué. Afin d'optimiser le processus de data mining, ce logiciel est conçu pour fonctionner avec les technologies SAS d'intégration de données, d'analyse et de reporting.

Une vue intégrée et complète des données

C'est lorsqu'il est intégré à la stratégie de diffusion de l'information que le data mining est le plus efficace. Cela inclut la collecte d'informations provenant de sources extrêmement diverses : web, centres d'appels, enquêtes, formulaires de retours clients, séries temporelles et systèmes transactionnels des points de vente etc. En utilisant conjointement SAS Text Miner à partir de la même interface, les données structurées et non structurées peuvent être également analysées. Ceci permet d'enrichir considérablement les modèles et d'étendre leurs capacités prédictives.

Articulation autour d'une interface utilisateur graphique conviviale

SAS Enterprise Miner est articulé autour d'une interface conviviale, de type glisser/déposer, conçue pour séduire aussi bien les statisticiens chevronnés que les analystes moins expérimentés.

Le processus comporte les cinq étapes clés qui constituent la méthodologie S.E.M.M.A conseillée par SAS : échantillonnage (Sample), Exploration, Modification, Modélisation et évaluation (Assess). A chacune de ces 5 étapes correspond une série d'actions - regroupant les différents algorithmes disponibles - exécutables tout au long du projet. En déployant des nœuds depuis la barre d'outils SEMMA, vous appliquez des statistiques évoluées, identifiez les variables les plus significatives, transformez des données avec des générateurs d'expressions, élaborez des modèles pour prédire les résultats, validez leur exactitude et générez une table soumise au scoring avec des valeurs prédites à déployer dans vos applications opérationnelles.

Génération rapide et autonome de modèles

SAS® Rapid Predictive Modeler accomplit automatiquement une série de tâches de data mining (transformation de données, sélection de variables, ajustement à divers algorithmes et évaluation de modèles) pour générer rapidement des modèles prédictifs adaptés à un grand nombre de problématiques métier. SAS Rapid Predictive Modeler s'exécute depuis SAS® Enterprise Guide® ou SAS® Add-In for Microsoft Excel et utilise les étapes de modélisation prédéfinies de SAS Enterprise Miner. Ce qui permet aux utilisateurs métier d'accéder facilement à la modélisation. En adoptant une approche collaborative, les analystes chevronnés peuvent enrichir et personnaliser les modèles développés avec SAS Rapid Predictive Modeler au travers de SAS Enterprise Miner.

Une suite inégalée de techniques et méthodes de modélisation

SAS Enterprise Miner se distingue par la profondeur de ses analyses reposant sur une suite à la fois classique et moderne d'algorithmes de modélisation prédictive et descriptive — arbres de décision, classification hiérarchique, régression linéaire et logistique, réseaux de neurones, Bagging et Boosting, data mining de séries chronologiques, machines à support vectoriel (SVM), raisonnement à base de cas (MBR), associations, analyse de séquence, analyse de chemins de navigation web, etc. Des algorithmes spécifiques à certaines problématiques métier sont aussi disponibles – scores de crédit, tarification, uplift modeling, analyse de survie – ainsi que des techniques avancées telles que la forêt décisionnelle, les splines de régression aux moindres angles et la régression des moindres carrés partiels (PLS).

Préparation, agrégation et exploration de données sophistiquées

La préparation des données représente, en règle générale, l'étape la plus longue d'un projet de data mining. Les fonctionnalités interactives de préparation de données de SAS Enterprise Miner permettent d'optimiser la gestion des valeurs manquantes, de filtrer les valeurs aberrantes et de définir des règles de segmentation. Ces fonctionnalités incluent l'importation, l'ajout, la jointure de fichiers et la suppression de variables. Les nombreuses fonctions d'agrégation et d'exploration interactives de données permettent aux néophytes d'analyser de grandes quantités de données dans des graphiques multidimensionnels à liaisons dynamiques. Il en résulte un data mining de qualité, dont les résultats sont adaptés aux spécificités des problématiques métier.

Comparaison de modèles, reporting et management orientés métier

La comparaison de modèles en termes de courbes de lift et de retour sur investissement est l'occasion pour les spécialistes du data mining de faciliter la communication de leurs résultats et de collaborer avec les experts métier.

Les modèles élaborés avec des algorithmes différents peuvent être évalués et comparés. Un nœud de seuil permet aussi d'analyser la répartition des probabilités a posteriori afin d'identifier les mesures optimales à mettre en œuvre, et de résoudre la problématique métier en question.

Une conception ouverte et évolutive, garante de souplesse

Personnalisable, l'environnement SAS Enterprise Miner permet d'ajouter des outils personnalisés et d'intégrer du code SAS personnalisé. Il est possible d'utiliser facilement dans l'interface des modèles SAS développés en dehors de SAS Enterprise Miner, et ce, tout en conservant une totale maîtrise sur la syntaxe de chaque instruction.

Le nœud Extension permet d'éditer interactivement les scripts d'apprentissage et de scoring mais aussi de les soumettre tout en consultant les journaux et de sorties. Les listes de sélection par défaut peuvent être étoffées avec des outils personnalisés intégrant du code SAS ou des fonctions XML, ouvrant ainsi l'univers SAS aux spécialistes du data mining.

Un processus de scoring automatisé pour des résultats plus rapides

Le scoring consiste à appliquer régulièrement un modèle à de nouvelles données pour une implémentation dans un environnement opérationnel réel. Ce processus peut s'avérer fastidieux, notamment s'il exige de réécrire ou de convertir du code — ce qui risque de retarder l'implémentation du modèle et d'introduire des erreurs. Le script de scoring doit refléter l'intégralité du processus aboutissant au modèle prédictif définitif, y compris chacune des étapes de prétraitement des données. SAS Enterprise Miner génère automatiquement le script de scoring dans différents langages : SAS, C, Java ou PMML. Ce script peut être déployé dans divers environnements de traitement en temps réel ou en mode batch dans SAS, sur le web ou dans des bases de données relationnelles.

Associés à SAS® Scoring Accelerator (un accélérateur de scoring) - disponible sur Aster, Pivotal (précédemment GreenPlum), IBM DB2, IBM Netezza, Oracle et Teradata - les modèles SAS Enterprise Miner peuvent être convertis en fonctions de scoring propres à une base de données, directement exécutables dans celle-ci. Les résultats peuvent également être directement transmis à d'autres solutions SAS (SAS® Marketing Automation, SAS® Model Manager, SAS® Real-Time Decision Manager) pour que le déploiement du data mining s'effectue dans des environnements opérationnels et en temps réel.

Une solution haute performance, nativement adaptée aux architectures de grille de calcul

L'architecture innovante client Java / serveur SAS offre une grande souplesse de configuration, permettant de passer d'un système mono-utilisateur à une solution d'entreprise de grande envergure. Dans ce type de configuration, des serveurs puissants peuvent être dédiés aux calculs et les utilisateurs peuvent accéder à tout moment aux différents projets et services de data mining, où qu'ils se trouvent (bureau, domicile ou site distant). Nombre de tâches très consommatrices — tri de données, agrégation, sélection de variables et régression — ont été redéveloppées en multithread de manière à pouvoir être exécutées en mode parallèle pour une bonne répartition et un équilibrage de charge sur une grille de serveurs, ou être programmés en traitement par lots.

Si la complexité des données et des analyses le permettent, un utilisateur pourra trouver des gains de performance conséquents sur une machine multi processeurs, et si les besoins évoluent davantage, notamment pour traiter les big data plus rapidement, SAS® High-Performance Data Mining (licence séparée) permettra de développer des modèles prédictifs adaptés à ce degré d'exigence. Pour en savoir plus : sas.com/hpdatamining.

Un système de data mining moderne et distribuable, adapté aux entreprises

SAS Enterprise Miner peut être déployé via un portail web de type client léger pour plusieurs utilisateurs, impliquant une maintenance minimale. Il peut également être entièrement configuré sur un poste de travail. Compatible avec les serveurs Windows et les plates-formes UNIX, SAS Enterprise Miner se révèle un logiciel de choix pour les entreprises gérant des projets de data mining d'envergure. En termes de restitution, des rapports couvrant l'intégralité de l'analyse peuvent être facilement créés et diffusés pour le reporting externe et la documentation interne. Les modèles peuvent être référencés de manière centralisée sur le serveur SAS pour être ensuite utilisés dans les applications SAS concernées. En fonction des besoins et des profils des utilisateurs, SAS Enterprise Miner offre donc la possibilité de développer, analyser, diffuser et gérer les modèles analytiques et les segmentations adaptés aux exigences de votre entreprise.

Principales caractéristiques

Interface intuitive

- Interface graphique intuitive pour la création de diagrammes, disponible en français :
  - élaboration de plus de modèles, plus précis, plus rapidement,
  - diffusion via le web,
  - accès à l'environnement de programmation SAS,
  - échange de diagrammes au format XML,
  - réutilisation des diagrammes comme modèle pour d'autres projets ou d'autres utilisateurs.
- Traitement batch :
  - encapsulation de toutes les fonctionnalités de l'interface,
  - à base de macros SAS,
  - incorporation des processus d'apprentissage et de scoring dans des applications personnalisées.

Evolutivité et montée en charge

- Les traitements peuvent s'exécuter sur un serveur, sur une grille, dans une base de données ou encore en mémoire
- Apprentissage asynchrone des modèles
- Possibilité d'arrêter un traitement
- Traitement parallélisé et distribué (« grid computing ») des algorithmes
- Exécution en parallèle de plusieurs diagrammes
- Tous les éléments sont stockés sur le/les serveur(s)

Accès et gestion des données

- Accès et prise en charge de données structurées et non structurées comme variables: séries chronologiques, tickets de caisse, parcours de pages web, enquêtes, etc.
- Import facilité des fichiers Excel, délimités, SAS et autres formats standards
- Prise en charge de variables contenant des caractères spéciaux et des langues différentes
- Exploration optimisée pour retrouver et visualiser rapidement des tables ou créer des graphiques interactivement
- Assistant de navigation et d'assignation des bibliothèques
- Manipulation de données : ajout ou suppression de variables, jointure ou ajout de tables
- Suppression des valeurs aberrantes
- Application de seuils sur différentes lois de distribution pour éliminer les valeurs

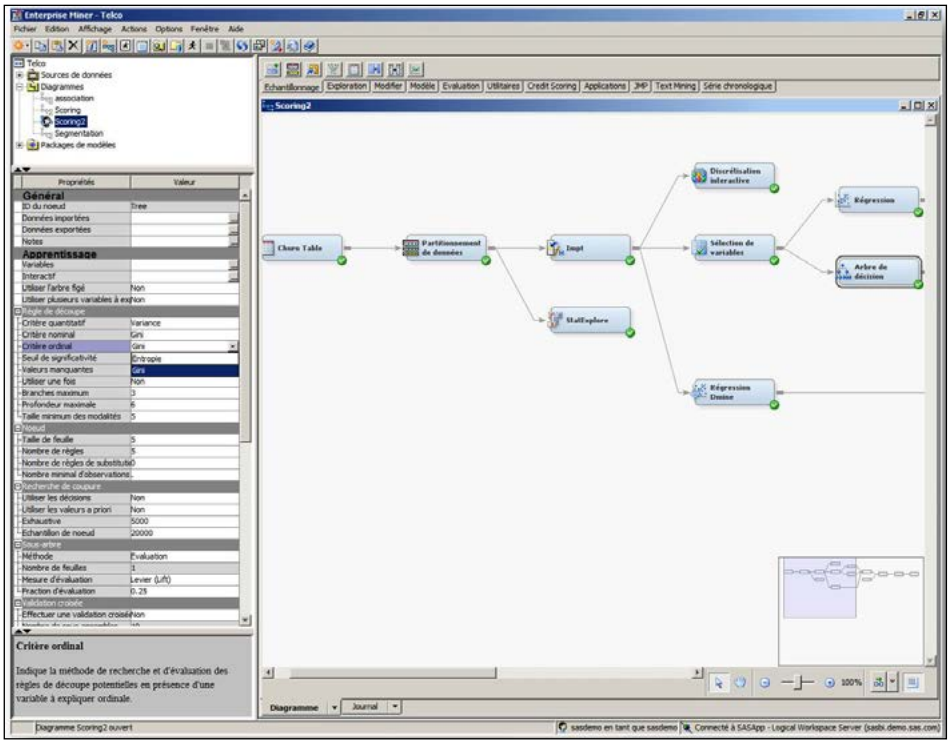

Figure1 : Dans l'interface utilisateur de SAS Enterprise Miner, le diagramme est un schéma auto-documenté, facile à actualiser. Ce diagramme peut être appliqué à de nouveaux jeux de données ou partagé avec d'autres analystes

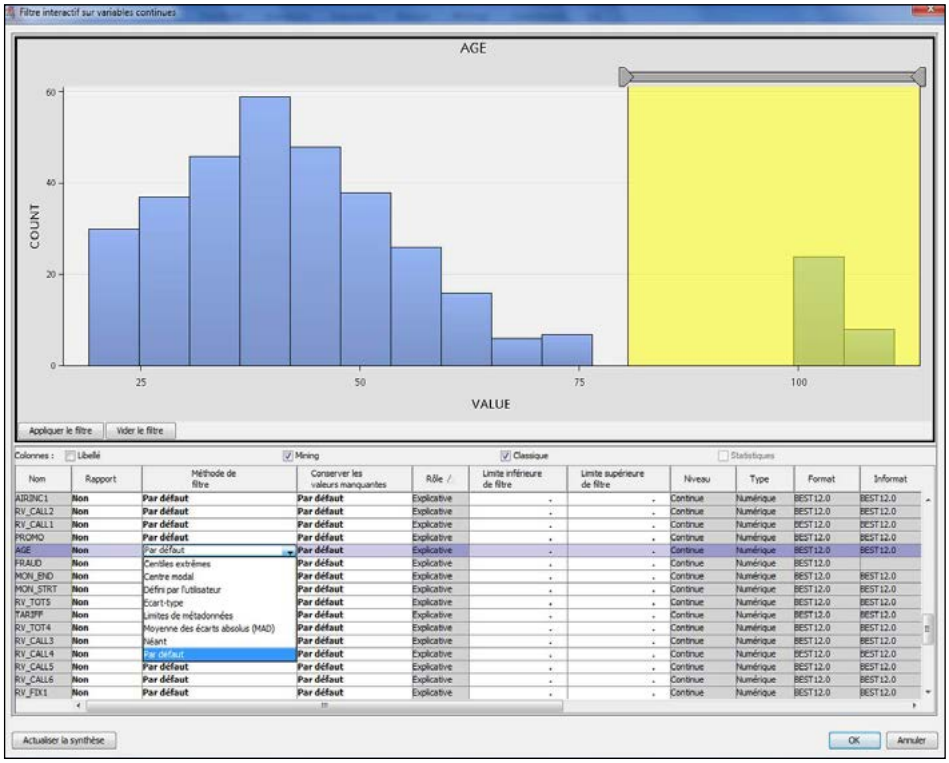

Figure 2 : Eliminez les valeurs aberrantes de manière interactive avec le nœud Filtrer.

extrêmes
- Regroupement des valeurs dont la fréquence est inférieure à N occurrences
- Filtrage interactif des valeurs
- Intégration avec R pour étoffer les types de modèles et leur comparaison
- Gestion des métadonnées permettant de modifier les attributs des variables tels que le rôle, le type, l'ordre
- Intégration native avec les autres composants de SAS
- Déploiement du script de scoring

Echantillonnage

- Aléatoire simple
- Stratifié
- Pondéré
- Par grappe
- Systématique
- Sélection des N premières observations
- Préférentiel
- Classifié et stratifié (Teradata 13

Partitionnement des données

- Création de tables d'apprentissage, de validation, de test et de sorties
- Evaluation et validation de la performance des modèles (méthode Hold-out)
- Stratification par défaut de la variable qualitative à expliquer
- Partitionnement équilibré sur toute variable qualitative

Transformation des variables

- Fonctions simples: log, log base 10, racine carrée, inverse, quadratique, exponentielle, normalisée
- Regroupement par classes, par quantiles ou discrétisation en fonction de la variable à expliquer
- Puissance optimale : normalité maximale, corrélation maximale et égalisation de l'étendue des niveaux avec la variable à expliquer
- Editeur d'interactions : définition des effets des interactions polynomiales et de Nième degré
- Définition interactive des transformations :
  - définitions de transformations personnalisées via le générateur d'expressions ou l'éditeur SAS,
  - comparaison de la distribution de la nouvelle variable avec la variable d'origine.

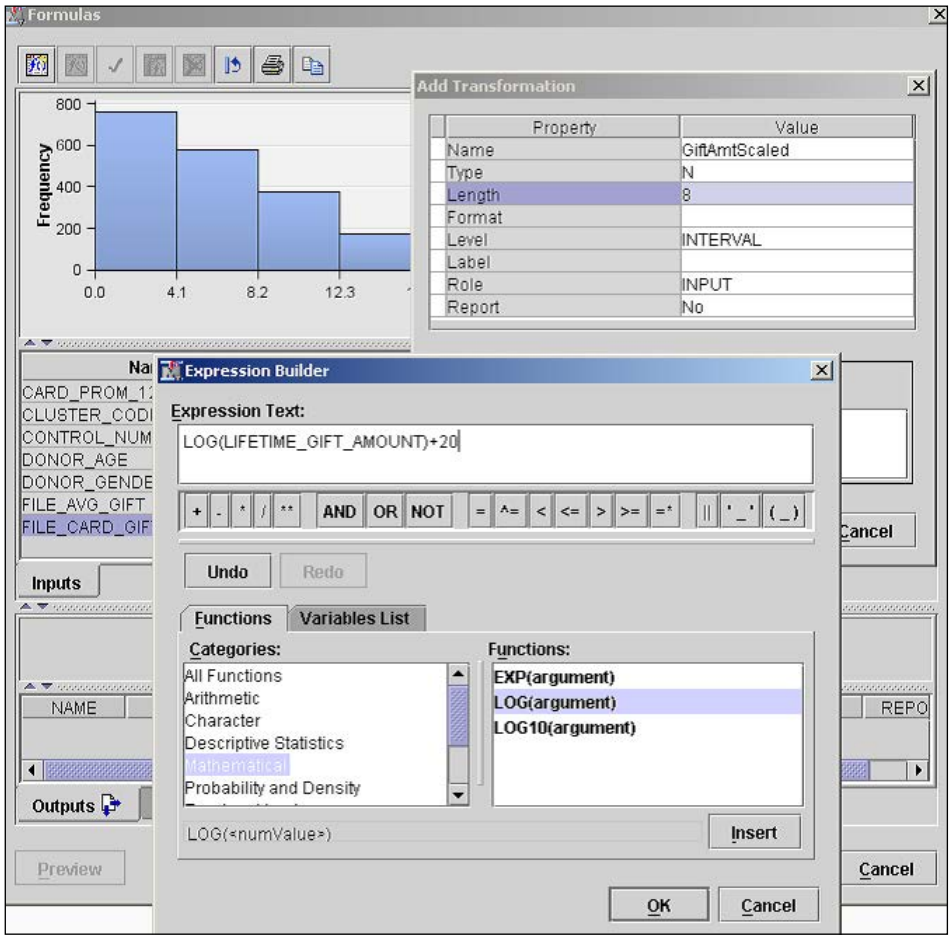

Figure3 : Créez des variables personnalisées via le générateur d'expressions.

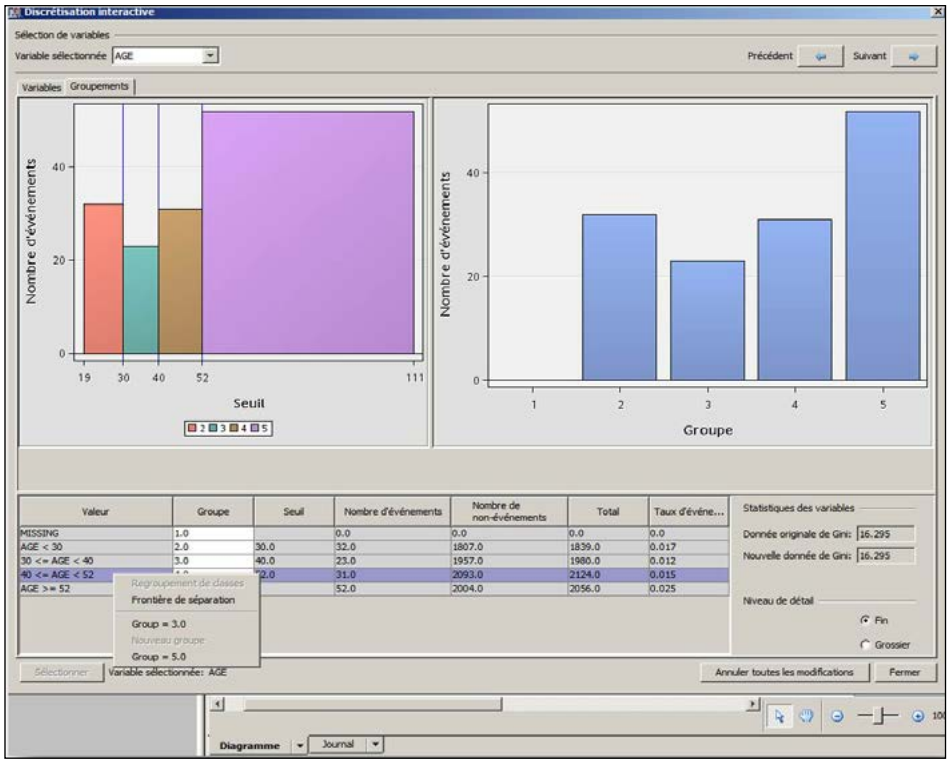

Figure 4 : Regroupez interactivement les variables en classes, pour refléter les règles métier : vous pourrez aussi découper, regrouper et sauvegarder ces variables modifiées afin de pouvoir les réutiliser par ailleurs.

Discrétisation interactive des variables

- En quantiles ou en nombre de classes
- Sélection de variables et mesure des effets via la statistique de Gini
- Regroupement des valeurs manquantes dans un groupe distinct
- Visualisation globale ou détaillée des groupes
- Profil de classes en fonction de la variable à expliquer
- Modification interactive des groupes et sauvegarde des définitions

Nœud Générateur de règles

- Création de règles et de stratégies ponctuelles basées sur les données
- Définition interactive de la valeur de la variable de sortie

Substitution de valeurs (enrichissement) et traitement des valeurs manquantes

- Mesures de la tendance centrale
- Distribution empirique
- Arbre de décision et avec substitutions
- Espacement mi-moyen,
- M-estimateurs robustes ou valeurs fixes
- Editeur de remplacement :
  - définition de nouvelles valeurs pour les variables qualitatives,
  - affectation de valeurs de remplacement pour les valeurs manquantes,
  - plafonnement interactif des valeurs continues dépassant un seuil.

Statistiques descriptives et représentations graphiques

- Statistiques univariées :
  - pour les variables continues : effectif, moyenne, médiane, min, max, écart-type, écart moyen à l'échelle et pourcentage de valeurs manquantes,
  - pour les variables qualitatives : nombre de catégories, effectifs, classe modale, et pourcentage de valeurs manquantes,
  - distributions,
  - ventilation statistique par valeur de la variable qualitative à expliquer.
- Statistiques bivariées :
  - test de corrélation ordonné de Pearson et de Spearman,
  - test du Khi-2 ordonné, découpage des variables continues en n classes,
  - coefficient de variation.

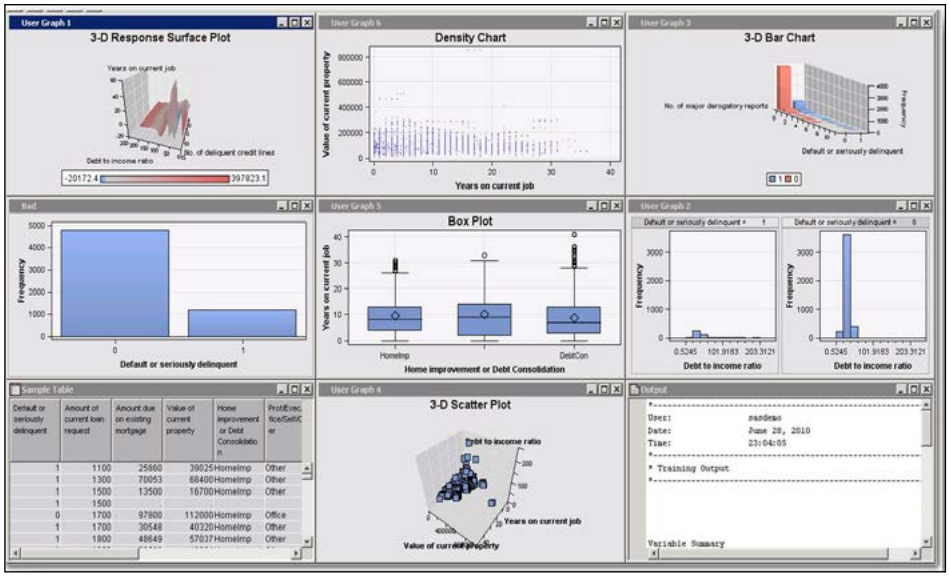

Figure 5 : Explorez vos données de manière interactive, à l'aide de différents types de graphiques.

- Sélection de variables par la statistique de LOGWORTH
- Distribution des variables en fonction de leur valeur dans la variable à expliquer
- Distribution des variables qualitatives sur la variable à expliquer ou sur un segment

Graphiques/visualisation

- Graphiques interactifs ou en batch : nuages de points, matriciels, boîte à moustaches, en constellation, courbe de niveau, diagrammes en bâtons, treillis, graphiques de densité et graphiques multidimensionnels ; diagrammes 3D et circulaires et graphiques en aires ; histogrammes
- Profil interactif des segments créés par la classification et la modélisation.
- Identification des variables caractérisant les profils et les différences entre groupes
- Création de titres et de notes de renvoi, choix entre plusieurs modèles de couleurs, modification aisée de l'échelle des axes
- Application d'une clause WHERE (filtre)
- Liaisons interactives entre graphiques et tables, permettant notamment de balayer et de regrouper
- Copier-coller de données et de graphiques dans d'autres applications ou enregistrement sous forme de fichiers .bmp
- Enregistrement automatique des graphiques interactifs dans la fenêtre de résultats du nœud

Classification et cartes auto-adaptatives

- Classification :
  - définie par l'utilisateur ou choix automatique des meilleures classes,
  - stratégies multiples pour le codage de variables qualitatives dans l'analyse,
  - traitement des valeurs manquantes,
  - représentation graphique du profil de chaque classe montrant la distribution des variables en entrée et des autres facteurs,
  - profil par arbre de décision pour prédire l'appartenance aux classes,
  - code PMML de scoring.
- Cartes auto-adaptatives :
  - calcul des cartes avec un lissage de Nadaraya-Watson ou un lissage linéaire local,
  - réseaux de Kohonen,
  - superposition à la distribution d'autres variables sur la carte,
  - gestion des valeurs manquantes.

Analyse d'associations et du panier de consommation (market basket)

- Etude d'associations et recherche de séquences :
  - grille des règles classées selon l'indicateur de confiance,
  - courbes de lift, de la confiance, de la confiance attendue, et de la tolérance en fonction des règles,
  - diagramme croisé du nombre de règles en fonction du Support et de la Confiance,
  - graphique croisant la confiance obtenue avec la confiance attendue,
  - table de description des règles,
  - représentation des règles sous forme de réseaux.
- Définition interactive d'un sous-ensemble de règles basées sur le lift, la confiance, le support, la longueur de chaîne, etc.
- Intégration des règles en tant que variable explicative pour enrichir une modélisation prédictive
- Associations hiérarchiques :
  - dérivation de règles à plusieurs degrés,
  - définition des correspondances
  - parent/enfant pour la table d'entrée dimensionnelle.

Analyse des chemins de navigation web

- Data mining évolutif et efficace des chemins les plus fréquemment empruntés à partir des données de parcours de navigation
- Analyse fréquente de sous-séquences à partir de tout type de données séquentielles

Analyse des liens

- Les données sont converties sous la forme d'un ensemble d'objets (ou d'entités) reliés et interconnectés qui peuvent être visualisés sous forme d'un réseau d'effets.
- Représentation graphique d'un modèle figurant la relation entre deux niveaux de variables provenant de données relationnelles, ou la cooccurrence entre deux éléments provenant de données transactionnelles
- Indicateurs de centralité et informations décrivant les communautés pour appréhender les graphiques relationnels.
- Statistiques de confiance pondérée fournissant des informations sur la meilleure offre à venir
- Scores de classes générés pour la réduction de données et la classification

Réduction des dimensions

- Sélection de variables :
  - sur la base des critères Khi-2 ou R2 (selon la nature de la variable), suppression des variables non

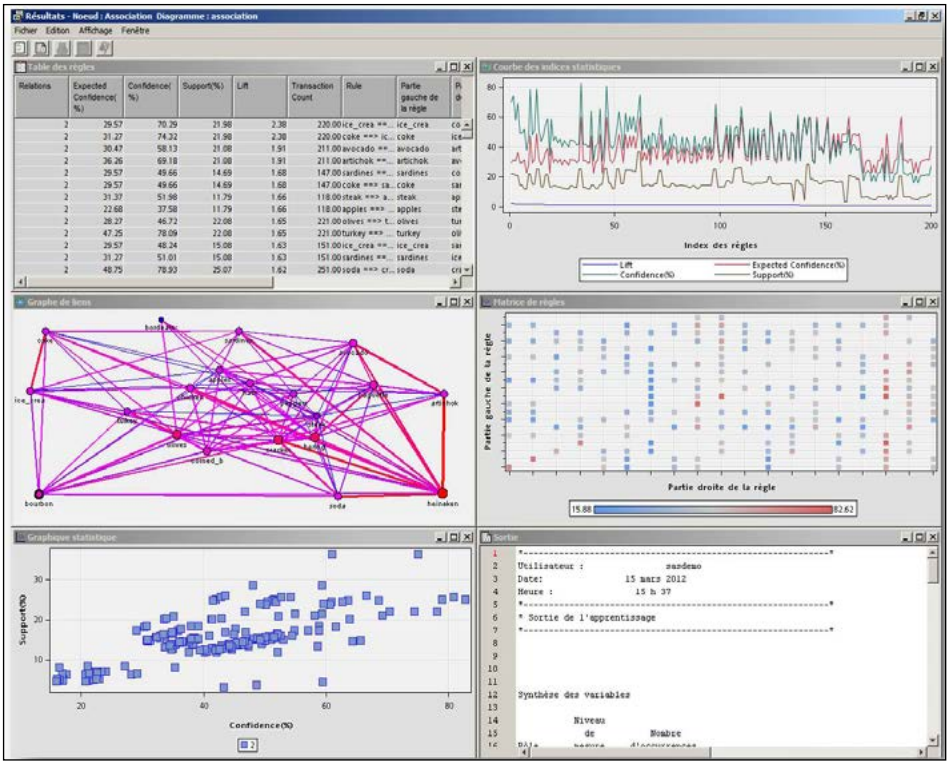

Figure 6 : Visualisez les profils d'associations. Analysez de manière interactive un sous-ensemble de règles sélectionnées selon la valeur du lift, la confiance, le support, la longueur de la chaîne, etc.

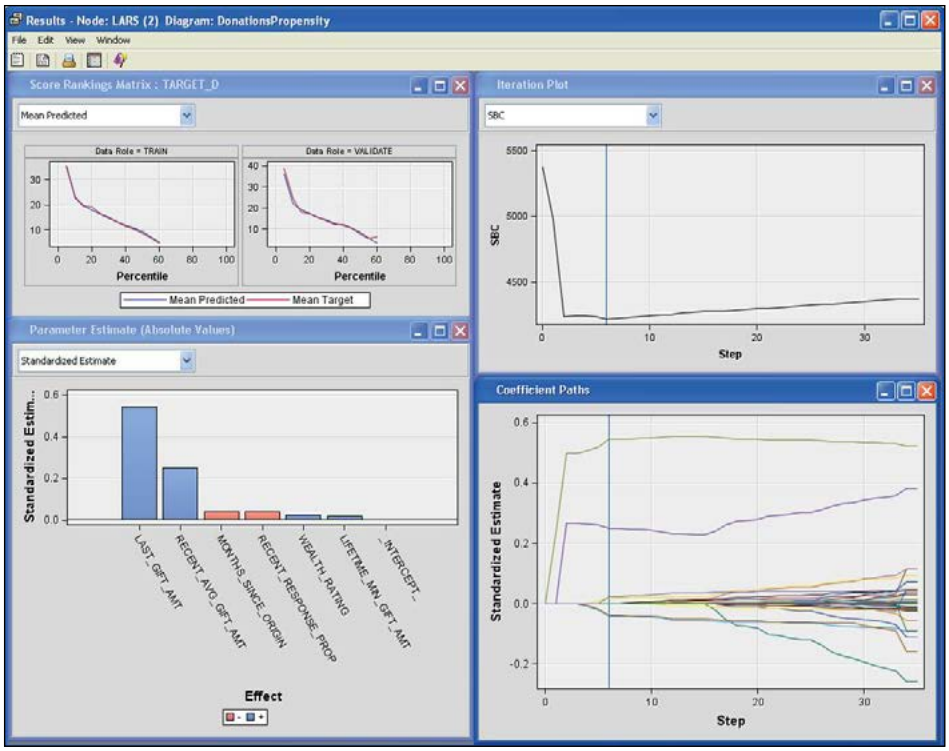

Figure 7 : Améliorez la sélection des variables, la validation croisée, et renforcez la stabilité de vos modèles grâce à la méthode LARS/LASSO

corrélées à la variable à expliquer, ou traduisant des liens hiérarchiques ou comportant de nombreuses valeurs manquantes,
  - réduction des variables nominales comportant un trop grand nombre de modalités.
  - Découpage des variables continues pour l'identification de relations non linéaires
  - Détection des interactions
- Sélection automatique de variables - méthode LARS (régression aux moindres angles) :
  - AIC, SBC, Cp de Mallows, validation croisée et autres critères de sélection,
  - graphiques représentant les paramètres, les itérations, le classement des scores, etc.,
  - généralisation à la méthode LASSO (Least Absolute Shrinkage and Selection Operator).
- Par composantes principales:
  - Calcul des valeurs et vecteurs propres à partir des matrices de corrélation et de covariance,
  - représentation graphique des coefficients et matrices des composantes principales, valeurs propres, logarithme de valeurs propres et valeurs propres proportionnelles cumulées,
  - choix interactif du nombre de composantes à conserver,
  - utilisation des composantes principales sélectionnées dans les algorithmes de modélisation prédictive.
- Classification des variables :
  - regrouper les variables en classes disjointes ou hiérarchiques,
  - apprentissage sur les valeurs propres et composantes principales,
  - support des variables nominales,
  - dendogramme de classes,
  - table de variables sélectionnées avec statistiques de classe et de corrélation,
  - réseau de classification et graphique R2,
  - remplacement interactif des variables sélectionnées.
- Préparation de séries chronologiques :
  - agrégation des données transactionnelles en séries chronologiques à l'aide de plusieurs techniques de cumul et transformations,
  - méthodes d'analyses : analyse saisonnière, de tendance,

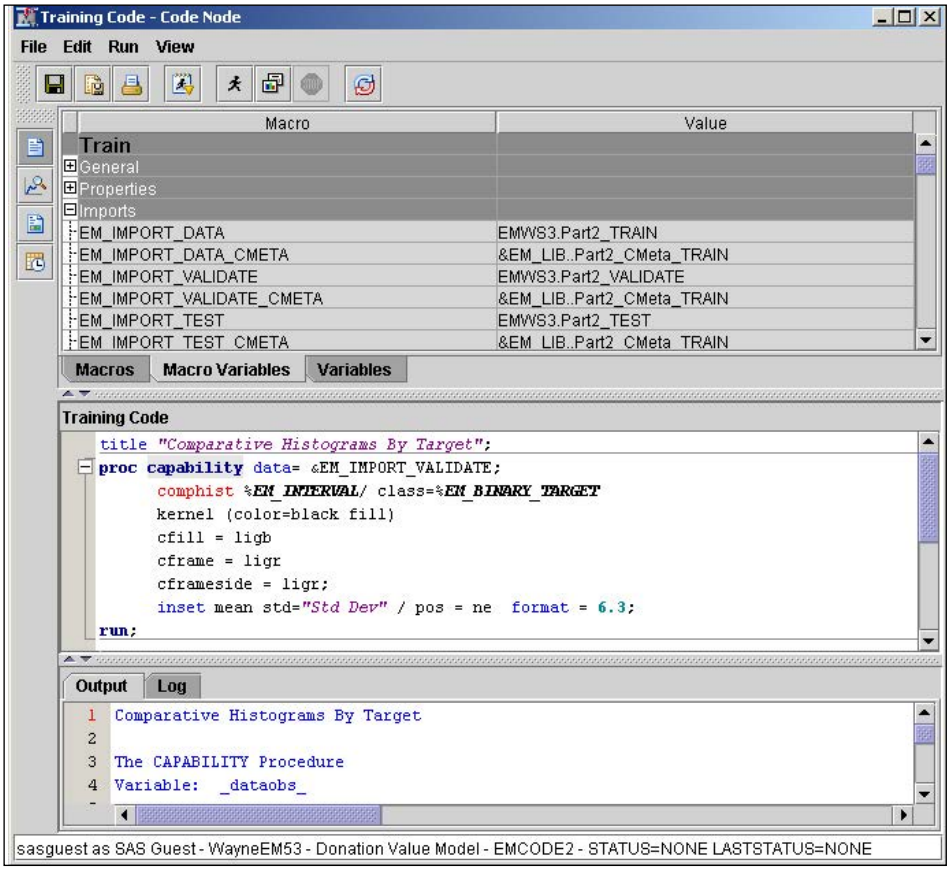

Figure 8 : Intégrez du code SAS personnalisé pour transformer les variables, incorporer des instructions ou des programmes existants, développer de nouveaux nœuds, enrichir les fonctions de scoring, personnaliser les rapports etc…

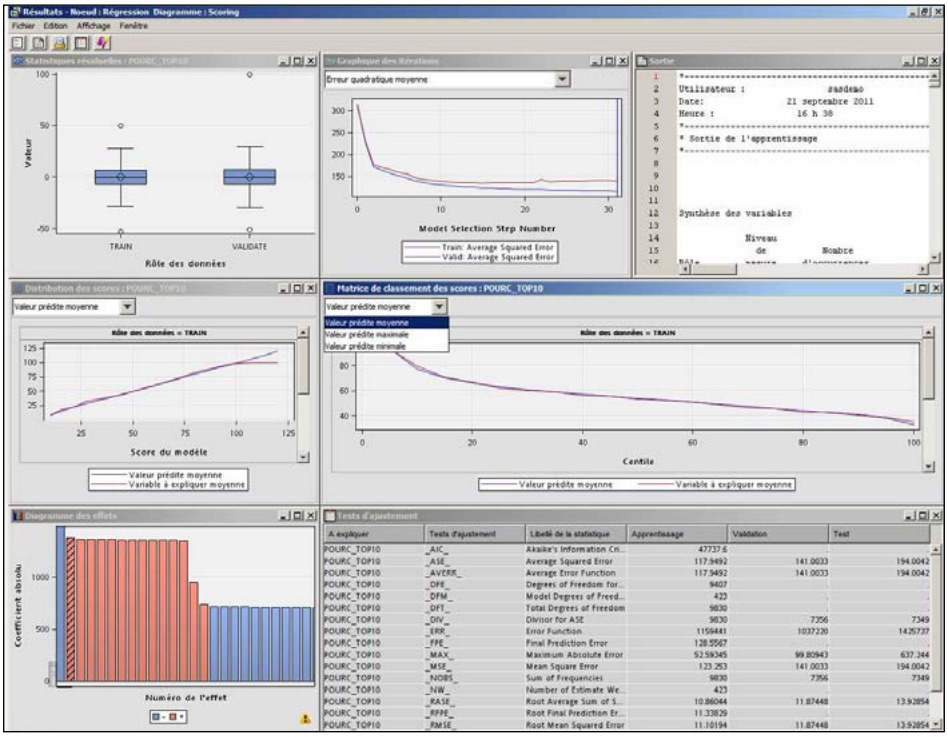

Figure 9 : Elaborez des modèles de régression linéaire et logistique en utilisant des méthodes de sélection adaptées ainsi que les diagnostiques d'interprétation.

horodatage, événementielle,
- exploitation des séries chronologiques obtenues dans les algorithmes de classification et de modélisation prédictive.

Nœud Code SAS

- Ecriture de code SAS pour les opérations de préparation et de transformation de données
- Incorporation de programmes SAS externes
- Elaboration de modèles personnalisés
- Création de nœuds personnalisés dans Enterprise Miner
- Enrichir le script de scoring
- Définition des bibliothèques, tables et options intégrées dans les scripts batch
- Interface de programmation intuitive :
  - utilisation de macro-variables pour le référencement de sources de données, de variables, etc.,
  - gestion distincte des scripts d'apprentissage, de scoring et de reporting,
  - sorties et journaux SAS.
- Création de graphiques

Fonctions pour une modélisation cohérente

- Sélection des modèles sur la base des échantillons d'apprentissage, de validation ou de test à partir d'un certain nombre de critères : bénéfice ou perte, AIC, SBC, erreur quadratique moyenne, taux de mauvaise classification, ROC, Gini ou KS (Kolmogorov-Smirnov)
- Intégration des probabilités a priori dans le processus de développement des modèles
- Identifications du type des variables (explicatives et à expliquer) : binaires, nominales, ordinales et continues
- Accès aisé au script de scoring et à la totalité des sources de données partitionnées
- Centralisation et organisation de résultats dans une même fenêtre pour mieux évaluer les performances
- Définition d'événements cible, de probabilités a priori et de matrices
- bénéfice/perte

Régression

- Linéaire et logistique (méthode de sélection pas à pas, ascendante ou descendante)
- Définition interactive des termes des équations : polynômiaux, interactions, effets hiérarchisés
- Validation croisée

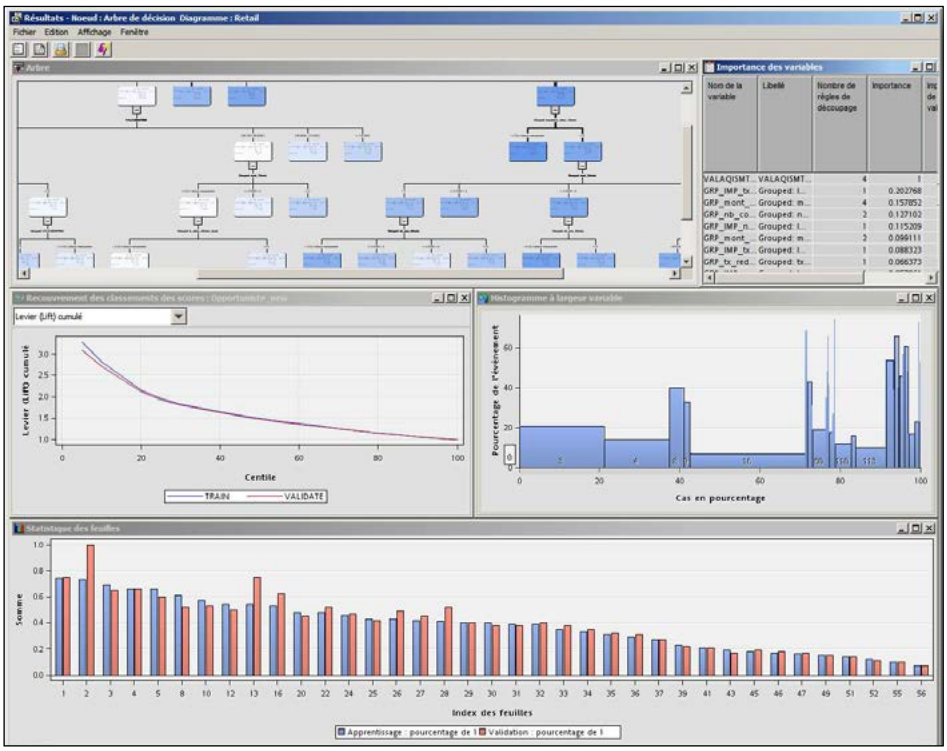

Figure 10 : Développez des arbres de décision de manière interactive. De nombreux graphiques de validation permettent d'évaluer la stabilité des arbres générés.

- Règles définissants les effets hiérarchisés
- Techniques d'optimisation : Gradient conjugué, DBLDOG, Newton-Raphson avec recherche à la ligne ou méthode de Ridge, Quasi-Newton et Région de confiance
- Régression Dmine :
  - régression des moindres carrés rapide, ascendante, pas à pas,
  - discrétisation optionnelle des variables pour détecter les relations non linéaires,
  - réduction des variables qualitatives optionnelle.
- Prise en compte des interactions
- Modélisation directement dans la base de données sur Teradata 13
- Script de scoring PMML

Arbres de décision

- Méthodologies:
  - CHAID, classification et régression, Bagging et Boosting, forêt décisionnelle et forêt aléatoire
  - Sélection de l'arbre et élagage en fonction des objectifs de bénéfice ou de lift
  - Validation croisée k-fold
- Critères de découpe : test du Khi-2, test de probabilité de Fisher, Gini, Entropie ou réduction de variance
- Changement de variables à expliquer pour l'élaboration de stratégies de segmentation multiobjectifs
- Production automatique d'identifiants de feuilles pour la modélisation et le traitement en groupe
- Affichage des règles
- Calcul de l'importance des variables à des fins de sélection préliminaire et d'interprétation des modèles
- Affichage de la précision des variables aux points de découpe des branches et des feuilles.
- Représentation consolidée de l'arbre
- Déploiement interactif de l'arbre :
  - croissance/élagage interactif des arbres ; développement/réduction des nœuds de l'arbre,
  - validation des données pour évaluer la stabilité de l'arbre,
  - définition de points de découpe personnalisés, binaires ou multiples sur toute variable candidate,
  - copier-coller des découpes,
  - les tables et les graphiques sont liés dynamiquement permettant ainsi une meilleure évaluation de la performance,
  - impression des arbres sur une ou plusieurs pages.
- Sélection interactive d'une partie d'un arbre
- Affichage de commentaires et de statistiques personnalisées sur un nœud
- Taille d'échantillon au sein d'un arbre contrôlée par l'utilisateur
- Utilise la procédure rapide ARBORETUM
- Script de scoring PMML

Machine à support vectoriel SVM

- Classificateur de marge maximale, pour les problèmes à plusieurs variables
- Prise en charge des variables à expliquer binaires
- Méthodes d'estimation : quadratique complète, quadratique décomposée, de Lagrange et des moindres carrés
- Fonctions linéaire, polynomiale, de base radiale et paramètres sigmoïdes
- Echantillonnage optionnel et validation croisée
- Scoring utilisant des procédures Base® SAS

Réseaux de neurones

- Architectures réseaux souples avec fonctions de combinaison et d'activation
- 10 techniques d'apprentissage
- Optimisation préliminaire
- Standardisation automatique des données en entrée
- Prise en charge des connexions directionnelles
- Nœud avec méthode automatique :
  - perceptron multicouche automatisé, avec recherche de la configuration optimale,
  - sélection du type et de la fonction d'activation à partir de quatre modèles d'architecture différents.
- Script de scoring PMML
- Nœud DMNeural :
  - modèles construits en utilisant la réduction de dimensions et la sélection de fonction,
  - apprentissage accéléré ; estimation linéaire et non linéaire.

Nœud Moindres carrés partiels

- Particulièrement utile pour extraire des facteurs à partir d'un nombre conséquent de variables éligibles

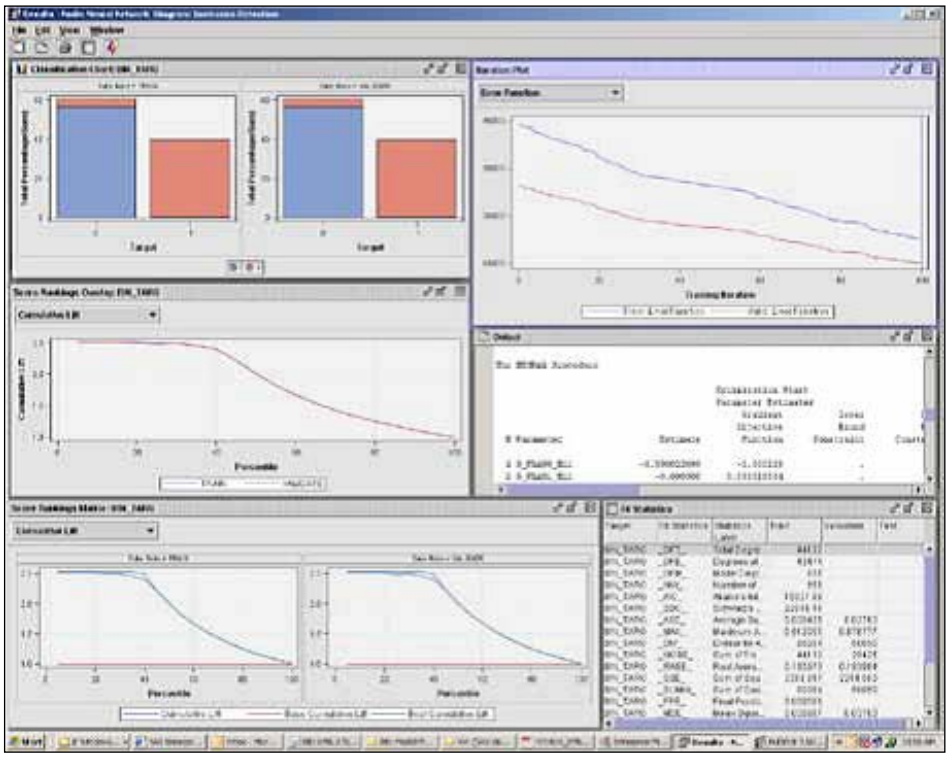

Figure 11 : Détectez les relations non linéaires complexes au moyen des Réseaux de neurones.

corrélées
- Régression sur composantes principales et régression de rang réduit
- Sélection du nombre de facteurs soit par l'utilisateur, soit automatique
- Gestion de la sélection de variables

Induction de règles

- Technique de modélisation prédictive récursive
- Particulièrement utile pour la modélisation d'événements rares

Modélisation en deux étapes (Two Stage)

- Modélisation séquentielle ou simultanée d'une variable cible qualitative ou continue
- Choix d'un arbre de décision, d'un modèle de régression ou de réseau de neurones à chaque étape
- Contrôle de l'application de la prédiction nominale à la prédiction continue
- Estimation précise de la valeur client

Raisonnement à base de cas (MBR)

- Technique basée sur les k plus proches voisins pour catégoriser ou prédire des observations
- Méthodes de l'arbre à dimensionnalité réduite et Scan brevetées

Ensembles de modèles

- Combinaison des prédictions de différents modèles pour constituer une solution plus robuste
- Méthodes utilisées : Moyenne, Par vote, et Maximum

Modèles de réponses incrémentales / net lift

- Comparaison des modèles de traitement net et de contrôle
- Variables cibles binaires ou continues
- Sélection de variables selon la méthode stepwise
- Calcul de revenu fixe ou variable
- Sélection de variables en fonction de la net information value (NIV)
- L'utilisateur peut définir le niveau de traitement sur la variable de traitement
- L'utilisateur peut définir une variable de coût ainsi qu'un coût fixe.
- Prise en compte possible sur le critère de la PNIV (Penalized Net Information Value)
- Options de sélection de modèle distinctes pour le modèle incrémental de ventes (seconde variable à expliquer)

Time series data mining

- Préparation de données sur des séries chronologiques :
  - Agrégation, transformation et synthèse de données transactionnelles et séquentielles
  - Transposition automatique des séries chronologiques à l'appui de l'analyse de similarité, de la classification et de la modélisation prédictive
  - traitement des données avec ou sans identifiant temporel
- Analyse de similarité :
  - Utile pour des prévisions sur de nouveaux produits, sur de courtes durées ou pour l'identification de tendances
  - Calcul de mesures de similarité entre les séries cible et celles en entrée ou parmi les séries en entrée
  - Matrice de similarité pour toutes les combinaisons des séries
  - Classification hiérarchique basée sur la matrice de similarité et les résultats du dendogramme
  - Diagramme en constellation pour l'évaluation des classes
- Lissage exponentiel :
  - Contrôle de décroissance pondérée avec un ou plusieurs paramètres de lissage
  - Sélection automatique de la méthode de lissage la mieux adaptée
- Analyse d'une série chronologique réduite au moyen de techniques descriptives

Analyse de survie

- Mise en œuvre de régressions logistiques multinomiales additives évaluant la réalisation d'un phénomène (temps discret)
- La probabilité de survenue de l'événement pour l'effet temps est modélisée avec une fonction spline cubique
- Les fonctions spline cubiques peuvent être prises en compte dans la méthode de sélection de variables stepwise en plus des effets principaux
- Intervalles de temps définis par les utilisateurs pour spécifier comment les analyser les données et gérer la censure
- Extension automatique des données et échantillonnage optionnel
- Prise en charge des covariables indépendantes du temps

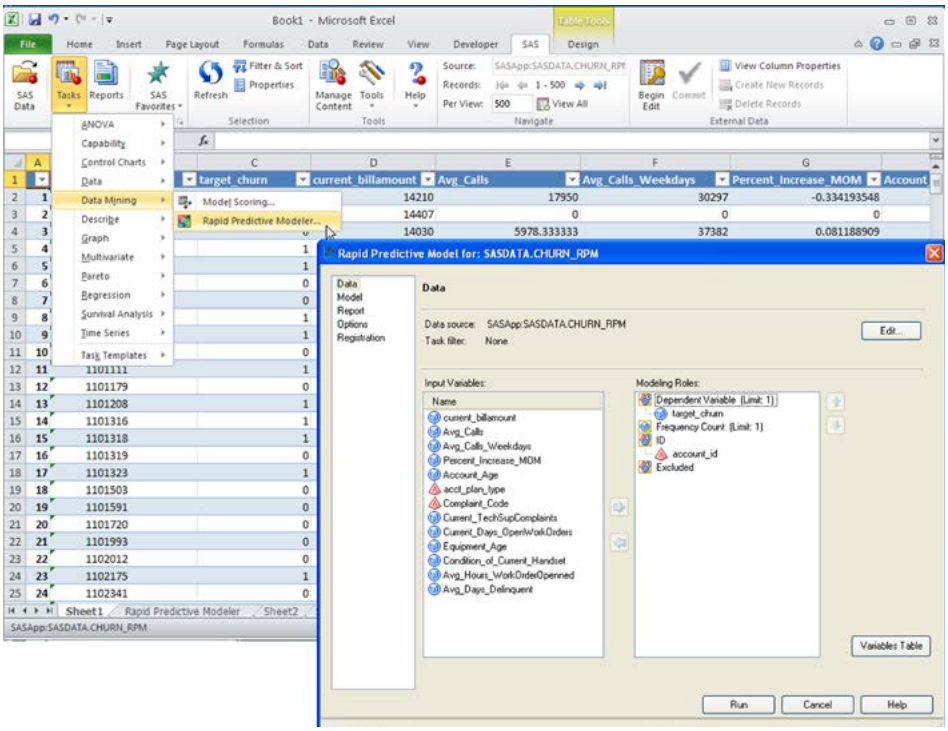

Figure 12 : Générez automatiquement des modèles prédictifs applicables à diverses problématiques métier via SAS Rapid Predictive Modeler (dans SAS Enterprise Guide ou dans Excel via SAS Add-In for Microsoft Excel).

- Fonction de survie générée validée sur les données exclues
- Calcul des risques compétitifs ou des sous-risques
- Script de scoring incluant le calcul de la durée de vie résiduelle moyenne
- Covariables dépendantes du temps inclues dans l'analyse – formats définis par l'utilisateur (standard, temps variable, étendu)
- Possibilité pour les utilisateurs de définir les troncatures à gauche et les date de censure

Tarification (assurances)

- Modélisation de la fréquence des dommages, de leur gravité et de la prime pure
- Sélection automatique d'une fonction de distribution et de liaison appropriée, avec ajustement manuel
  - distributions : de Poisson, binomiale négative, gamma, normale, gaussienne inverse, Tweedie et de Poisson,
  - fonctions de liaison : logarithmique, inverse, puissance quadratique, puissance cubique et log-log.
- Rapidité de modélisation en mémoire avec pré-classification des effets
- Générateur d'interactions
- Contrôle par l'utilisateur de la définition du modèle ZIP pour gérer la sur-dispersion
- Statistiques standard : AIC, SBC, Khi-2 et déviance de Pearson
- Calcul de relativités pour chaque effet de modélisation
- Représentation graphique en bandes de la relativité affichant les bornes basses et hautes de l'intervalle de confiance
- Script de scoring SAS

Traitement en groupe avec les nœuds Début et Fin de traitement en groupe

- Répétition du traitement d'un segment du diagramme
- Utilisations : modélisation stratifiée, Bagging et Boosting, plusieurs variables à expliquer, validation croisée

SAS® Rapid Predictive Modeler, outil personnalisé dans SAS® Enterprise Guide® ou SAS® Add-In for Microsoft (Excel)

- Génération automatique de modèles prédictifs applicables à diverses problématiques métier
- Ouverture, enrichissement et modification des modèles dans SAS Enterprise Miner
- Rapports concis et simples à analyser (courbe de lift, courbes ROC et scorecard)
- Possibilité d'évaluer les tables d'apprentissage, et de sauvegarder une table scorée

Procédures data-mining haute performance

- Ces procedures multithreaded peuvent s'exécuter simultanément de manière distribuée sur les processeurs disponibles de votre serveur SMP, accélérant ainsi les traitements :
  - HPREDUCE (high-performance variable reduction).
  - HPNEURAL (high-performance neural networks).
  - HPFOREST (high-performance forests).
  - HP4SCORE (high-performance 4Score).
  - HPDECIDE (high-performance decide).
  - HPDS2 (high-performance DS2).
  - HPDMDB (high-performance data mining database).
  - HPSAMPLE (high-performance sampling).
  - HPSUMMARY (high-performance data summarization).
  - HPIMPUTE (high-performance imputation).
  - HPBIN (high-performance binning).
  - HPCORR (high-performance correlation).
- Nœuds haute performance dans SAS Enterprise Miner :
  - HP Data Partition, HP Explore,
  - HP Transform, HP Variable Selection, HP Regression, HP Neural Network, HP Tree, HP Forest and HP Impute.

Import de modèles

- Enregistrement des modèles en vue de leur réutilisation dans d'autres diagrammes ou projets
- Import et évaluation de modèles externes

Figure 13 : Evaluez simultanément plusieurs modèles au sein d'un environnement facile à interpréter à l'aide du nœud de comparaison des modèles.

Evaluation de modèles

- Un environnement unique pour comparer les performances de tous les modèles testés
- Sélection automatique du modèle le plus performant en fonction de critères définis par l'utilisateur (AIC, erreur quadratique moyenne, ROC, Gini, KS, taux de mauvaise classification, taux de vrais positifs, etc.)
- Possibilité pour l'utilisateur de forcer la sélection
- Statistiques très complètes d'ajustement et de diagnostic
- Courbes de lift, courbes ROC
- Graphiques bénéfice/perte avec choix décisionnels ; matrice de confusion (classification)
- Graphique de distribution des scores selon des probabilités de classe ; matrice de classement des scores
- Classement des scores et distribution des variables continues
- Nœud Seuil permettant de définir le(s) seuil(s) de probabilité pour les cibles binaires
- Statistiques : KS Max, coût minimal d'erreur de classification, profil cumulatif max, taux max de vrais positifs, précision max d'un événement à partir des entraînements antérieurs, Event Precision Equal Recall
- Accès à un environnement graphique interactif pour :
  - comparer et opposer des modèles concurrents issus de familles identiques ou différentes,
  - évaluer l'importance des variables et leurs effets sur les réponses prévues,
  - identifier les paramètres de fonctions optimaux par rapport au résultat mesuré, simulations incluses.

Nœud Générateur de rapports

- Génération automatique de document PDF ou RTF à partir d'un flux de processus
- Documentation du processus d'analyse et partage des résultats facilité
- La documentation peut être intégrée dans les packages générés
- Intègre l'image du diagramme de flux de processus
- Intègre les notes définies par l'utilisateur

Scoring

- Nœud mis en place pour appliquer un scoring interactif via l'interface
- Création par défaut d'un script de scoring optimisé, éliminant les variables inutilisées
- Génération automatisée du script de scoring dans SAS, C, Java et PMML (version 3.1)
- Le code de scoring (SAS, C et Java) inclut les étapes de modélisation, classification, transformation et de traitement des valeurs manquantes
- La procédure PSCORE permet d'exécuter des scripts de scoring au format PMML sur les données (expérimental). : régression, arbres de décision, classification et réseaux neuronaux
- Scoring des modèles SAS Enterprise Miner dans les bases de données (Aster Data, Teradata, Pivotal (anciennement GreenPlum), IBM DB2 ou Netezza) avec SAS Scoring Accelerator

Enregistrement et gestion des modèles

- Enregistrement des modèles SAS Enterprise Miner sur le serveur de métadonnées SAS® Metadata Server
- Enregistrement des modèles développés en langage SAS sur le serveur de métadonnées SAS Metadata Server via la macro %AA_MODEL_REGISTER
- Intégration avec SAS Model Manager, autorisant la gestion des versions du script de scoring, la gestion du cycle de vie des modèles, du développement jusqu'à la production, et le suivi des modèles
- Intégration avec SAS® Enterprise Guide®, SAS® Add-In for Microsoft Office et SAS® Data Integration Studio pour le scoring de modèle

SAS Institute s. a. s. – DOMAINE DE GREGY - GREGY-SUR-YERRES - 77257 BRIE COMTE ROBERT - FRANCE
TEL. : +33(0) 1 60 62 11 11 FAX : +33(0) 1 60 62 11 99 www.sas.com/france

02133FCE0214