IBM SPSS Modeler - Guide CRISP-DM

Remarque

Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations figurant à la section «Remarques», à la page 39.

Certaines illustrations de ce manuel ne sont pas disponibles en français à la date d'édition.

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.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Ces informations s'appliquent à la version 17.0.0 d'IBM(r) SPSS(r) Modeler et à toutes les éditions et modifications ultérieures, sauf mention contraire dans les nouvelles éditions.

Table des matières

Avis aux lecteurs canadiens v

Préface vii

Chapitre 1. Introduction à CRISP-DM . . 1
Présentation de l'aide CRISP-DM 1
CRISP-DM dans IBM SPSS Modeler 1
Ressources supplémentaires 2

Chapitre 2. Compréhension de l'entreprise 5
Présentation de la compréhension de l'entreprise . . 5
Détermination des objectifs commerciaux 5
Exemple d'eCommerce : définition d'objectifs commerciaux 5
Compilation de l'environnement métier 6
Définition des objectifs commerciaux 6
Critères en matière de réussite commerciale . . . 7
Evaluation de la situation 7
Exemple d'eCommerce : évaluation de la situation 7
Inventaire des ressources 8
Impératifs, hypothèses et contraintes 8
Risques et plans de secours 9
Terminologie 9
Analyse coût-bénéfice 9
Détermination des objectifs de l'exploration de données 9
Objectifs de l'exploration de données 10
Exemple d'eCommerce : objectifs de l'exploration de données 10
Critères de réussite de l'exploration de données 10
Production d'un plan de projet 11
Elaboration du plan du projet 11
Exemple de plan de projet 11
Outils et techniques d'évaluation 11
Prêt pour la prochaine étape ? 12

Chapitre 3. Compréhension des données 13
Présentation de la compréhension des données . . 13
Collecte des données initiales 13
Exemple d'eCommerce : collecte initiale de données 13
Elaboration d'un rapport sur la collecte des données 14
Description des données 14
Exemple d'eCommerce : description de données 14
Elaboration d'un rapport de description des données 15
Exploration des données 15
Exemple d'eCommerce : exploration des données 15
Elaboration d'un rapport sur l'exploration des données 16
Vérification de la qualité des données 16
Exemple d'eCommerce : vérification de la qualité des données 17
Elaboration d'un rapport sur la qualité des données 17
Prêt pour la prochaine étape ? 17

Chapitre 4. Préparation des données 19
Présentation de la préparation des données. . . . 19
Sélection de données 19
Exemple d'eCommerce : sélection de données . . 19
Inclusion ou exclusion de données. 19
Nettoyage des données 20
Exemple d'eCommerce : nettoyage des données 20
Elaboration d'un rapport sur le nettoyage des données 21
Construction de nouvelles données 21
Exemple d'eCommerce : construction de données 21
Calcul d'attributs 21
Intégration des données 22
Exemple d'eCommerce : intégration de données 22
Tâches d'intégration 22
Formatage de données. 23
Prêt pour la modélisation ? 23

Chapitre 5. Modélisation 25
Présentation de la modélisation. 25
Sélection de techniques de modélisation 25
Exemple d'eCommerce : techniques de modélisation 25
Choix des techniques de modélisation appropriées 26
Hypothèses de modélisation 26
Génération d'une conception de test 26
Elaboration d'une conception de test 27
Exemple d'eCommerce : conception de test. . . 27
Création des modèles 27
Exemple d'eCommerce : création de modèles . . 28
Valeurs des paramètres 28
Exécution des modèles 28
Description des modèles 28
Evaluation du modèle 29
Evaluation complète d'un modèle 29
Exemple d'eCommerce : évaluation des modèles 29
Suivi des paramètres révisés. 30
Prêt pour la prochaine étape ? 30

Chapitre 6. Evaluation 31
Présentation de l'évaluation 31
Evaluation des résultats 31
Exemple d'eCommerce : évaluation des résultats 31
Processus de révision 32
Exemple d'eCommerce : rapport de révision . . 32
Détermination des étapes suivantes 33
Exemple d'eCommerce : étapes suivantes . . . 33

Chapitre 7. Déploiement 35
Présentation du déploiement 35
Planification pour le déploiement 35
Exemple d'eCommerce : planification du déploiement 35
Planification de la surveillance et de la maintenance 36
Exemple d'eCommerce : surveillance et maintenance 36
Production d'un rapport final 37
Préparation d'une présentation finale 37
Exemple d'eCommerce : rapport final 37
Exécution d'une révision de projet finale. 38
Exemple d'eCommerce : révision finale 38

Remarques 39
Marques 40

Index 43

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Préface

IBM® SPSS Modeler est le puissant utilitaire d'exploration de données de IBM Corp.. SPSS Modeler aide les entreprises et les organismes à améliorer leurs relations avec les clients et les citoyens grâce à une compréhension approfondie des données. A l'aide des connaissances plus précises obtenues par le biais de SPSS Modeler, les entreprises et les organismes peuvent conserver les clients rentables, identifier les opportunités de vente croisée, attirer de nouveaux clients, détecter les éventuelles fraudes, réduire les risques et améliorer les services gouvernementaux.

L'interface visuelle de SPSS Modeler met à contribution les compétences professionnelles de l'utilisateur, ce qui permet d'obtenir des modèles prédictifs plus efficaces et de trouver des solutions plus rapidement. SPSS Modeler dispose de nombreuses techniques de modélisation, telles que les algorithmes de prévision, de classification, de segmentation et de détection d'association. Une fois les modèles créés, l'utilisateur peut utiliser IBM SPSS Modeler Solution Publisher pour les remettre aux responsables, où qu'ils se trouvent dans l'entreprise, ou pour les transférer vers une base de données.

A propos d'IBM Business Analytics

Le logiciel IBM Business Analytics propose des informations complètes, cohérentes et précises auxquelles les preneurs de décisions peuvent se fier pour améliorer les performances de leur entreprise. Un porte-feuilles étendu de veille économique, d'analyses prédictives, de gestion des performances et de stratégie financières et d'applications analytiques vous offre des informations claires, immédiates et décisionnelles sur les performances actuelles et vous permet de prévoir les résultats futurs. Ce logiciel intègre des solutions dédiées à l'industrie, des pratiques éprouvées et des services professionnels qui permettent aux organisations de toute taille de maximiser leur productivité, d'automatiser leurs décisions sans risque et de proposer de meilleurs résultats.

Ce porte-feuilles intègre le logiciel IBM SPSS Predictive Analytics qui aide les organisations à prévoir les événements à venir et à réagir en fonction des informations afin d'améliorer leurs résultats. Des clients des secteurs du commerce, de l'éducation et des administrations du monde entier font confiance à la technologie IBM SPSS qui offre un avantage concurrentiel en attirant et fidélisant les clients et en améliorant la base de données de la clientèle tout en diminuant la fraude et en réduisant les risques. En utilisant le logiciel IBM SPSS dans leurs opérations quotidiennes, les organisations deviennent des entreprises prédictives, capables de diriger et d'automatiser les décisions pour répondre aux objectifs commerciaux et obtenir un avantage concurrentiel mesurable. Pour des informations supplémentaires ou pour joindre un représentant, consultez http://www.ibm.com/spss.

Assistance technique

L'assistance technique est disponible pour les clients du service de maintenance. Les clients peuvent contacter l'assistance technique pour obtenir de l'aide concernant l'utilisation des produits IBM Corp. ou l'installation dans l'un des environnements matériels pris en charge. Pour contacter l'assistance technique, rendez-vous sur le site Web IBM Corp. à l'adresse http://www.ibm.com/support. Lorsque vous contactez l'assistance technique, soyez prêt à indiquer votre identité, le nom de votre société et votre contrat d'assistance.

Chapitre 1. Introduction à CRISP-DM

Présentation de l'aide CRISP-DM

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining, est une méthode mise à l'épreuve sur le terrain permettant d'orienter vos travaux d'exploration de données.

- En tant que **méthodologie**, CRISP-DM comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.
- En tant que **modèle de processus**, CRISP-DM offre un aperçu du cycle de vie de l'exploration de données.

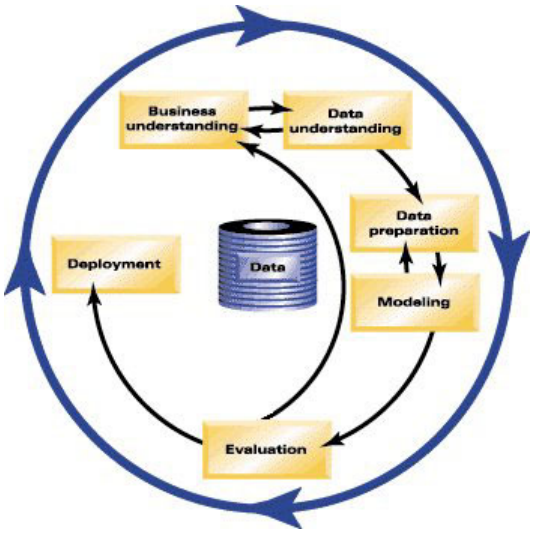

Figure 1. Le cycle de vie de l'exploration des données

Le modèle de cycle de vie comporte six phases dotées de flèches indiquant les dépendances les plus importantes et les plus fréquentes entre les phases. La séquence des phases n'est pas strictement établie. De fait, les projets, pour la plupart, passent d'une phase à l'autre en fonction des besoins.

Adaptable, le modèle CRISP-DM peut être aisément personnalisé. Ainsi, si votre entreprise cherche à repérer un blanchiment d'argent, vous examinerez certainement une grande quantité de données sans objectif précis concernant la modélisation. Votre travail sera ciblé non sur la modélisation, mais sur l'exploration et la visualisation de données avec pour objectif de découvrir des configurations suspectes parmi les données financières. CRISP-DM vous permet de créer un modèle d'exploration de données adapté à vos besoins.

Dans une telle situation, les phases de modélisation, d'évaluation et de déploiement peuvent s'avérer d'un intérêt moindre que les phases de préparation et de compréhension des données. Toutefois, certaines des questions soulevées durant ces dernières phases sont tout de même à prendre en considération pour les planifications à long terme et les futurs objectifs d'exploration de données.

CRISP-DM dans IBM SPSS Modeler

IBM SPSS Modeler utilise la méthodologie CRISP-DM de deux manières afin d'offrir une prise en charge unique et efficace de l'exploration de données.

© Copyright IBM Corp. 1994, 2015

- L'outil de projet CRISP-DM vous aide à organiser les flux, les sorties et les annotations des projets en fonction des phases d'un projet d'exploration de données standard. Vous pouvez créer des rapports au cours du projet en utilisant les notes des flux et les phases CRISP-DM.
- L'aide de CRISP-DM vous guide à travers le processus de réalisation d'un projet d'exploration de données. Le système d'aide comprend des listes de tâches pour chaque étape, ainsi que des exemples du fonctionnement de CRISP-DM dans la réalité. Vous pouvez accéder à l'aide de CRISP-DM en choisissant **Aide CRISP-DM** dans la fenêtre principale du menu Aide.

Outil de projet CRISP-DM

L'outil de projet CRISP-DM offre une approche structurée de l'exploration de données, garante de la réussite de votre projet. Il s'agit avant tout d'une extension de l'outil de projet standard de IBM SPSS Modeler. En fait, vous pouvez basculer entre la vue CRISP-DM et la vue Classes standard pour visualiser les flux et les sorties organisés par types ou par phases CRISP-DM.

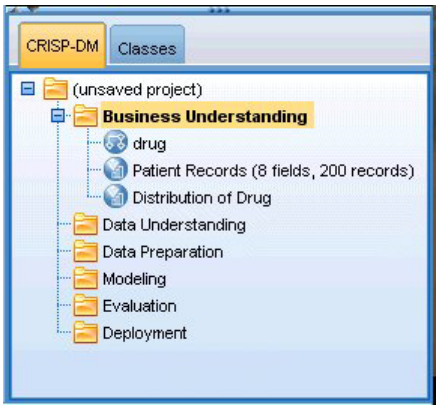

Figure 2. Outil de projet CRISP_DM

A l'aide de la vue CRISP-DM de l'outil de projet, vous pouvez :

- organiser les flux et les sorties d'un projet en fonction des phases d'exploration de données.
- prendre des notes sur les objectifs de l'entreprise pour chaque phase.
- créer des info-bulles personnalisées pour chaque phase.
- prêter attention aux conclusions tirées d'un graphique ou d'un modèle particulier.
- générer un rapport ou une mise à jour au format HTML à remettre à l'équipe chargée du projet.

Aide de CRISP-DM

IBM SPSS Modeler propose un guide en ligne pour le modèle de processus CRISP-DM non exclusif. Ce guide est organisé par phases de projet et comporte les éléments suivants :

- Présentation de chaque phase de CRISP-DM et liste des tâches correspondantes
- Aide concernant la création de rapports pour différentes étapes importantes
- Exemples issus de situations réelles montrant l'apport de CRISP-DM en matière d'exploration de données pour une équipe chargée d'un projet
- Liens vers des ressources supplémentaires liées à CRISP-DM

Vous pouvez accéder à l'aide de CRISP-DM en choisissant **Aide CRISP-DM** dans la fenêtre principale du menu Aide.

Ressources supplémentaires

En plus de l'aide CRISP-DM dans IBM SPSS Modeler, vous pouvez améliorer votre compréhension des processus d'exploration de données de plusieurs manières.

- Visitez le site Web du consortium CRISP-DM à l'adresse www.crisp-dm.org

- Lisez le manuel CRISP-DM, créé par le consortium CRISP-DM et fourni avec cette version.
- Lisez l'ouvrage *Data Mining with Confidence*, © SPSS Inc., 2002., ISBN 1-56827-287-1.

Chapitre 2. Compréhension de l'entreprise

Présentation de la compréhension de l'entreprise

Avant d'utiliser IBM SPSS Modeler, prenez le temps de vous pencher sur les bénéfices que votre société souhaite tirer de l'exploration de données. Cette consultation doit englober le plus grand nombre de personnes possible. Compilez les résultats obtenus. L'étape finale de cette phase CRISP-DM concerne la production d'un plan de projet à l'aide des informations ainsi recueillies.

Bien que cette étude puisse paraître superflue, elle s'avère au contraire indispensable. La compréhension des objectifs de votre société en matière d'exploration de données garantit une approche homogène indispensable avant la mise en oeuvre de ressources précieuses.

Détermination des objectifs commerciaux

Votre première tâche consiste à obtenir la meilleure compréhension possible des objectifs commerciaux en matière d'exploration de données. Cette tâche peut être plus compliquée qu'il n'y paraît, mais vous pouvez réduire au minimum les risques susceptibles de survenir par la suite en clarifiant les problèmes, et en définissant les objectifs et les ressources.

La méthodologie CRISP-DM propose une approche structurée vous permettant de mener à bien cette tâche.

Liste des tâches

- Commencez à recueillir des informations de fond sur la situation actuelle de l'entreprise.
- Apportez des informations précises sur des objectifs commerciaux spécifiques définis par les décideurs les plus importants.
- Convenez des critères utilisés pour la détermination des succès en matière d'exploration de données d'un point de vue commercial.

Exemple d'eCommerce : définition d'objectifs commerciaux

Scénario de Web-Mining à l'aide de CRISP-DM

A mesure que le nombre de sociétés s'attaquant au commerce via l'Internet augmente, les entreprises de commerce électronique déjà présentes sur le Web dans le domaine de l'informatique/de l'électronique sont soumises à une concurrence croissante de la part de ces nouveaux sites. Puisqu'il s'avère que le nombre de commerçants installés sur le Web augmente bien plus rapidement que celui des nouveaux clients adeptes de l'Internet, ces entreprises doivent trouver le moyen de préserver leurs bénéfices malgré l'augmentation des coûts liés à l'acquisition d'une clientèle. L'une des solutions proposées est l'entretien des relations existantes avec la clientèle afin de maximiser la valeur de chacun des clients actuels de l'entreprise.

L'entreprise demande alors la réalisation d'une étude ayant les objectifs suivants :

- Augmenter les ventes croisées de produit en améliorant les recommandations.
- Augmenter la fidélité des clients grâce à une meilleure personnalisation du service.

Cette étude sera considérée comme une réussite si :

- Les ventes de produits associés augmentent de 10 %.
- Les clients passent plus de temps sur le site et consultent davantage de pages à chaque visite.
- L'étude est réalisée dans les délais et sans dépassement de budget.

Compilation de l'environnement métier

La compréhension de la situation commerciale de votre entreprise vous permet de connaître les facteurs en jeu, notamment :

- Les ressources disponibles (personnel et matériel)
- Les problèmes
- Les objectifs

Vous aurez à faire quelques recherches sur la situation commerciale actuelle afin de découvrir des réponses valables aux questions exerçant une influence décisive sur l'issue du projet d'exploration de données.

Tâche 1 : détermination de la structure organisationnelle

- Créez des organigrammes illustrant les services, départements et groupes de travail. Veillez à inclure les noms et fonctions des responsables.
- Identifiez les personnes-clés de l'entreprise.
- Identifiez un commanditaire interne qui fournira un soutien financier et/ou des compétences spécifiques.
- Déterminez l'existence d'un comité d'organisation et procurez-vous la liste de ses membres.
- Identifiez les unités de l'entreprise concernées par le projet d'exploration de données.

Tâche 2 : description du contexte de la problématique

- Identifiez l'environnement de la problématique : par exemple, le marketing, l'assistance clientèle ou la prospection de clientèle.
- Décrivez le problème de manière générale.
- Définissez les conditions préalables du projet. Quels sont les motifs qui le sous-tendent ? L'entreprise utilise-t-elle déjà l'exploration de données ?
- Vérifiez le statut du projet d'exploration de données au sein de l'entreprise. Cet effort a-t-il été approuvé ou faut-il assurer la « promotion » de l'exploration de données en soulignant le caractère essentiel de cette technologie pour l'entreprise ?
- Au besoin, organisez des présentations d'informations sur l'exploration de données à l'intention de votre société.

Tâche 3 : description de la solution actuelle

- Décrivez toute solution actuellement utilisée pour résoudre le problème métier.
- Décrivez les avantages et les inconvénients de la solution actuelle. Indiquez aussi le degré d'acceptation rencontré par cette solution au sein de l'entreprise.

Définition des objectifs commerciaux

Nous abordons à présent l'aspect plus concret de la question. Suite à vos recherches et à votre étude, vous devez établir un objectif principal concret qui fasse l'unanimité des commanditaires du projet et des autres unités de l'entreprise concernées par les résultats du projet. Cet objectif, dont la formulation peut être aussi confuse que « réduction du score d'attrition de la clientèle », se traduira par la suite par des objectifs d'exploration de données spécifiques qui guideront votre analyse.

Liste des tâches

Veillez à prendre des notes sur les points suivants afin de les intégrer par la suite au plan du projet. N'oubliez pas que les objectifs doivent rester réalistes.

- Décrivez le problème à résoudre à l'aide de l'exploration de données.
- Enoncez le plus précisément possible toutes les questions d'ordre commercial.

- Déterminez tout autre impératif commercial (par exemple, éviter de perdre les clients actuels tout en augmentant les possibilités de vente de produits associés).
- Précisez les bénéfices attendus d'un point de vue commercial (par exemple, réduction de 10 % du score d'attrition par les clients importants).

Critères en matière de réussite commerciale

L'objectif à atteindre peut être clairement établi, mais comment saurez-vous qu'il est atteint ? Avant de poursuivre, il est donc important de définir la nature de la réussite commerciale de votre projet d'exploration de données. Les critères de réussite sont de deux sortes :

- **Les critères objectifs.** Il peut s'agir de critères très simples, comme une augmentation donnée de l'exactitude des audits ou une réduction convenue du score d'attrition.
- **Les critères subjectifs.** Les critères subjectifs, tels que la « découverte de clusters de traitements efficaces » sont plus difficiles à cerner, mais vous pouvez convenir de la personne qui prendra la décision finale.

Liste des tâches

- Fournissez des informations aussi précises que possible sur les critères de réussite de votre projet.
- Veillez à ce que chaque objectif commercial soit lié à un critère de réussite.
- Indiquez les arbitres décidant si les critères subjectifs de réussite sont atteints. Si possible, notez leurs attentes.

Evaluation de la situation

Maintenant que l'objectif est clairement établi, passons à l'évaluation de la situation actuelle. Cette étape soulève des questions telles que :

- Quels types de données sont disponibles pour l'analyse ?
- Le personnel nécessaire à la réalisation du projet est-il disponible ?
- Quels sont les plus grands facteurs de risque en jeu ?
- Existe-t-il un plan de secours pour chaque risque ?

Exemple d'eCommerce : évaluation de la situation

Scénario de Web-Mining à l'aide de CRISP-DM

Une société d'eCommerce dans le domaine de l'électronique débute dans le Web-Mining et a donc décidé de requérir l'assistance d'un expert en exploration de données. L'une des premières tâches à laquelle est confronté ce consultant est l'évaluation des ressources de la société en matière d'exploration de données.

Personnel. Il apparaît clairement que la société dispose d'experts en matière de gestion des fichiers logs de serveur et des bases de données de produits et d'achats, mais qu'elle a en revanche peu d'expérience du stockage et du nettoyage des données en vue de l'analyse. La consultation d'un spécialiste des bases de données peut donc être envisagée. Puisque la société espère que les résultats de l'étude s'inscriront dans un processus continu de Web-Mining, la direction doit aussi décider si les postes créés au cours du projet deviendront permanents.

Données. Puisqu'il s'agit d'une société bien établie, l'expert dispose de nombreux logs Web et données sur les achats. En fait, pour cette étude initiale, la société limitera l'analyse aux clients qui se sont « enregistrés » sur le site. En cas de réussite, ce programme pourra être étendu.

Risques. Hormis le coût des consultants et le temps consacré par les employés à l'étude, cette opération ne comporte que peu de risques immédiats. Toutefois, le temps étant un facteur d'importance, ce premier projet est programmé pour durer un trimestre.

Par ailleurs, la société connaissant actuellement beaucoup de décaissements supplémentaires, il est impératif que le budget de cette étude soit respecté. Au cas où l'un de ces deux objectifs serait compromis, les responsables commerciaux ont proposé de réduire la portée du projet.

Inventaire des ressources

Il est indispensable de procéder à un inventaire précis des ressources. Vous gagnerez du temps et vous éviterez beaucoup de soucis en examinant avec réalisme les questions liées au personnel, aux sources de données et au matériel.

Tâche 1 : recherche des ressources en matériel

- Quel est le matériel utilisé ?

Tâche 2 : identification des sources de données et de connaissances

- Quelles sont les sources de données disponibles pour l'exploration de données ? Notez les types et les formats de données.
- Comment les données sont-elles stockées ? Pouvez-vous accéder en direct aux entrepôts de données ou aux bases de données opérationnelles ?
- Projetez-vous d'acquérir des données externes, telles que des informations démographiques ?
- Existe-t-il des problèmes de sécurité empêchant l'accès aux données requises ?

Tâche 3 : identification des ressources en personnel

- Disposez-vous d'experts en matière de données et de métier ?
- Avez-vous identifié les administrateurs de base de données ou le personnel d'assistance technique dont vous pourriez avoir besoin ?

Une fois ces questions posées, joignez la liste des contacts et des ressources au rapport sur cette phase.

Impératifs, hypothèses et contraintes

Vos efforts seront mieux récompensés si vous évaluez avec réalisme les difficultés du projet. Le fait de définir ces risques aussi clairement que possible contribuera à éviter les problèmes ultérieurement.

Tâche 1 : détermination des impératifs

Le premier impératif est l'objectif commercial mentionné précédemment. Toutefois, tenez compte des points suivants :

- Les données ou les résultats du projet sont-ils soumis à des restrictions juridiques ou à des restrictions liées à la sécurité ?
- L'ensemble du personnel concerné est-il conscient des impératifs de la programmation du projet ?
- Le déploiement des résultats est-il soumis à des impératifs particuliers (par exemple, la publication sur le Web ou la lecture des scores dans une base de données) ?

Tâche 2 : définition des hypothèses

- Existe-t-il des facteurs économiques susceptibles d'influer sur le projet (par exemple, honoraires de consultants ou produits concurrentiels) ?
- Existe-t-il des hypothèses relatives à la qualité des données ?
- Comment la direction ou les commanditaires du projet s'attendent-ils à voir les résultats ? En d'autres termes, souhaitent-ils comprendre le modèle utilisé ou simplement consulter les résultats ?

Tâche 3 : vérification des contraintes

- Disposez-vous de tous les mots de passe nécessaires à l'accès aux données ?
- Avez-vous vérifié toutes les contraintes juridiques auxquelles est soumise l'utilisation des données ?

- Toutes les contraintes financières sont-elles couvertes par le budget du projet ?

Risques et plans de secours

La prudence veut aussi que vous examiniez les risques encourus au cours du projet. Ces risques sont notamment liés aux domaines suivants :

- Programmation (Que se passe-t-il si le projet dure plus longtemps que prévu ?)
- Financement (Que se passe-t-il si le commanditaire du projet rencontre des difficultés budgétaires ?)
- Données (Que se passe-t-il si les données sont de mauvaise qualité ou peu représentatives ?)
- Résultats (Que se passe-t-il si les résultats initiaux sont moins spectaculaires que prévu ?)

Une fois ces différents risques évalués, établissez un plan de secours afin d'éviter les désastres.

Liste des tâches

- Evaluez tous les risques possibles avec précision.
- Etablissez un plan de secours pour chaque risque.

Terminologie

Afin de vous assurer que l'équipe commerciale et celle chargée de l'exploration de données « parlent bien le même langage », pensez à compiler un glossaire des termes techniques et des mots-clés nécessitant une définition. Par exemple, si le terme « score d'attrition » possède un sens bien particulier pour votre entreprise, il convient de le rappeler explicitement à l'équipe. De même, il peut être utile pour tout le monde de définir l'utilisation des graphiques de gains.

Liste des tâches

- Dressez la liste des termes posant problème aux membres de l'équipe. Incluez les termes commerciaux et ceux liés à l'exploration de données.
- Pensez à publier cette liste sur le réseau interne de l'entreprise ou dans la documentation du projet.

Analyse coût-bénéfice

Cette étape permet de répondre à la question : **Quel est votre résultat net ?** Dans le cadre de l'évaluation finale, il est essentiel de comparer les coûts du projet aux bénéfices rapportés en cas de succès.

Liste des tâches

Dans votre analyse, incluez une estimation des coûts liés aux éléments suivants :

- Collecte des données et utilisation de données externes
- Déploiement des résultats
- Coûts d'exploitation

Ensuite, prenez en compte les bénéfices apportés par les éléments suivants :

- Réussite de l'objectif principal
- Connaissances supplémentaires engendrées par l'exploration des données
- Avantages éventuels issus d'une meilleure compréhension des données

---

Détermination des objectifs de l'exploration de données

Puisque l'objectif commercial est clairement défini, il convient à présent de le traduire en concepts d'exploration de données. Par exemple, l'objectif commercial « réduction du score d'attrition » peut se traduire en un objectif d'exploration de données comprenant les éléments suivants :

- Identification des clients importants en fonction des données concernant les achats récents

- Création d'un modèle utilisant les données disponibles sur les clients afin de prédire la probabilité d'attrition pour chaque client
- Affectation d'un rang à chaque client en fonction de sa valeur et de sa propension à l'attrition

Si ces objectifs d'exploration de données sont réalisés, l'entreprise peut les utiliser pour réduire le score d'attrition parmi ses clients les plus précieux.

Comme vous le voyez, la technologie et les compétences métier doivent être utilisées de pair pour que les opérations d'exploration de données soient efficaces. Pour obtenir des conseils plus spécifiques sur la détermination des objectifs de l'exploration de données, poursuivez la lecture de l'aide.

Objectifs de l'exploration de données

Lorsque vous travaillez avec des analystes de gestion et des analystes de données sur la définition d'une solution technique au problème métier, pensez à rester dans le concret.

Liste des tâches

- Décrivez le **type** du problème d'exploration de données (par exemple, classification ou prévision).
- Fournissez des informations sur les objectifs techniques en employant des unités de temps précises (par exemple, prévisions valables sur trois mois).
- Dans la mesure du possible, chiffrez les résultats souhaités (par exemple, obtention de scores d'attrition pour 80 % des clients existants).

Exemple d'eCommerce : objectifs de l'exploration de données

Scénario de Web-Mining à l'aide de CRISP-DM

Avec l'aide d'un consultant en exploration de données, la société d'eCommerce a traduit ses objectifs commerciaux en objectifs d'exploration de données. Les objectifs de l'étude initiale à réaliser ce trimestre sont les suivants :

- Utiliser l'historique des précédents achats pour générer un modèle liant les articles « associés ». Lorsque les utilisateurs consultent la description d'un article, fournir des liens vers d'autres articles du groupe associé (**analyse de panier d'achat**).
- Utiliser les logs Web pour déterminer ce que recherchent les différents clients, puis concevoir différemment le site afin de mettre ces articles en relief. Chacun des différents « types » de client verra apparaître une page d'accueil différente pour le site (**réalisation de profils**).
- Utiliser les logs Web pour tenter de prévoir les pages que consultera l'utilisateur en vous basant sur la page qu'il vient de quitter et les pages du site consultées (**analyse de la séquence**).

Critères de réussite de l'exploration de données

La réussite doit aussi être définie en termes techniques afin que vos efforts d'exploration de données conservent leur pertinence. Utilisez l'objectif d'exploration de données déterminé précédemment pour élaborer un point de référence pour la réussite. IBM SPSS Modeler fournit des outils tels que les noeuds Evaluation et Analysis pour vous aider à analyser l'exactitude et la validité de vos résultats.

Liste des tâches

- Décrivez les méthodes d'évaluation des modèles (par exemple, exactitude, performances, etc.).
- Définissez des points de référence pour l'évaluation de la réussite. Indiquez des chiffres précis.
- Définissez les mesures subjectives de votre mieux et déterminez l'arbitre en matière de réussite.
- Déterminez si le déploiement réussi des résultats du modèle fait partie de la réussite de l'exploration de données. Commencez dès à présent à planifier le déploiement.

Production d'un plan de projet

A présent, vous êtes prêt à créer le plan du projet d'exploration de données. Les questions posées jusqu'ici, ainsi que les objectifs d'exploration de données et les objectifs commerciaux formulés, formeront la base de ce plan.

Elaboration du plan du projet

Le plan du projet est le document principal régissant tout votre travail d'exploration de données. S'il est bien créé, il permettra d'informer toutes les personnes associées au projet des objectifs, des ressources, des risques et du programme de toutes les phases de l'exploration de données. Vous pouvez publier ce plan, ainsi que la documentation recueillie lors de cette phase, sur le réseau interne de la société.

Liste des tâches

Lors de l'élaboration du plan, vérifiez que vous avez bien répondu aux questions suivantes :

- Avez-vous discuté des tâches du projet et du plan proposé avec toutes les personnes concernées ?
- Le plan comprend-il une estimation des dates pour toutes les phases ou tâches ?
- Avez-vous inclus dans le plan les efforts et les ressources nécessaires au déploiement des résultats ou de la solution commerciale ?
- Les demandes de révision et les points de décision sont-ils mis en évidence dans le plan ?
- Avez-vous signalé les phases comprenant généralement des itérations multiples, telles que la modélisation ?

Exemple de plan de projet

Le plan d'ensemble de l'étude se présente comme le tableau suivant :

Tableau 1. Plan d'ensemble du projet Exemple

Phase	Temps	Resources	Risques
Compréhension de l'entreprise	1 semaine	Tous les analystes	Changement économique
Compréhension des données	3 semaines	Tous les analystes	Problèmes de données, problèmes technologiques
Préparation des données	5 semaines	Consultant en exploration de données, analyste en bases de données (quelques heures)	Problèmes de données, problèmes technologiques
Modélisation	2 semaines	Consultant en exploration de données, analyste en bases de données (quelques heures)	Problèmes technologiques, incapacité à trouver un modèle adéquat
Evaluation	1 semaine	Tous les analystes	Changement économique, incapacité à mettre en oeuvre les résultats
Déploiement	1 semaine	Consultant en exploration de données, analyste en bases de données (quelques heures)	Changement économique, incapacité à mettre en oeuvre les résultats

Outils et techniques d'évaluation

Puisque vous avez choisi d'utiliser IBM SPSS Modeler comme outil de la réussite de votre opération d'exploration de données, recherchez au cours de cette étape les techniques d'exploration de données

répondant le mieux aux besoins de votre société. IBM SPSS Modeler offre une gamme complète d'outils pour chaque phase de l'exploration de données. Pour décider du moment où il convient d'utiliser ces différentes techniques, consultez la section relative à la modélisation dans l'aide en ligne.

Prêt pour la prochaine étape ?

Avant d'explorer les données et de commencer à utiliser IBM SPSS Modeler, vérifiez que vous avez bien répondu aux questions suivantes.

Du point de vue commercial :

- Quels sont les bénéfices espérés par l'entreprise à l'issue de ce projet ?
- Comment définirez-vous la réussite des efforts mis en oeuvre ?
- Disposez-vous du budget et des ressources nécessaires à la réussite des objectifs ?
- Avez-vous accès à toutes les données nécessaires au projet ?
- Avez-vous débattu avec votre équipe des risques et des plans de secours associés au projet ?
- Ce projet est-il valable, au vu des résultats de l'analyse coût-bénéfice ?

Après avoir répondu aux questions ci-dessus, avez-vous traduit les réponses en un objectif d'exploration de données ?

Du point de vue de l'exploration de données :

- Quelle aide spécifique vous apportera l'exploration de données dans la réussite de vos objectifs commerciaux ?
- Connaissez-vous les techniques d'exploration de données susceptibles de produire les meilleurs résultats ?
- Comment saurez-vous que vos résultats sont suffisamment précis ou efficaces ? (*Avez-vous défini une mesure de la réussite en matière d'exploration de données ?*)
- Comment les résultats de la modélisation seront-ils déployés ? Avez-vous tenu compte du déploiement dans votre plan de projet ?
- Le plan du projet comprend-il toutes les phases CRISP-DM ?
- Les dépendances et les risques sont-ils énumérés dans le plan ?

Si la réponse à toutes ces questions est Oui, vous êtes prêt à vous pencher sur les données.

Chapitre 3. Compréhension des données

Présentation de la compréhension des données

La phase de compréhension des données de CRISP-DM implique l'étude des données disponibles pour l'exploration de données. Cette étape revêt une importance vitale, car elle permet d'éviter les problèmes inattendus au cours de la phase suivante, la préparation des données, phase généralement la plus longue d'un projet.

La compréhension des données implique l'accès aux données et leur exploration à l'aide de tables et de graphiques pouvant être organisés dans IBM SPSS Modeler grâce à l'outil de projet CRISP-DM. Vous pouvez ainsi déterminer la qualité des données et décrire les résultats de ces étapes dans la documentation du projet.

Collecte des données initiales

Cette phase de l'utilisation de CRISP-DM implique à présent l'accès aux données et leur insertion dans IBM SPSS Modeler. Les données proviennent de sources variées, telles que :

- **Données existantes**. Cette catégorie comprend plusieurs types de données, telles que les données transactionnelles, les données issues d'enquêtes, les logs Web, etc. Evaluez ces données existantes pour voir si elles suffisent à répondre à vos besoins.
- **Données acquises.** Votre société utilise-t-elle des données d'appoint, telles que des données démographiques ? Si la réponse est non, peut-être faut-il envisager leur utilisation.
- **Autres données. Si les sources ci-dessus ne répondent pas à vos besoins, vous devrez peut-être mener des enquêtes ou effectuer davantage de suivis afin de compléter les magasins de données existants.**

Liste des tâches

Examinez les données de IBM SPSS Modeler et étudiez les questions suivantes. Veillez à noter vos résultats. Pour plus d'informations, voir la rubrique «Elaboration d'un rapport sur la collecte des données», à la page 14.

- Quels sont les attributs (colonnes) de la base de données qui semblent les plus prometteurs ?
- Quels sont les attributs qui semblent sans intérêt et peuvent être exclus ?
- Le nombre de données permet-il de tirer des conclusions pouvant être généralisées ou d'effectuer des prévisions précises ?
- Les attributs sont-ils trop nombreux pour la méthode de modélisation choisie ?
- Opérez-vous la fusion de données issues de plusieurs sources ? Si oui, certains points risquent-ils de poser problème lors de la fusion ?
- Avez-vous envisagé le mode de traitement des valeurs manquantes dans chacune de vos sources de données ?

Exemple d'eCommerce : collecte initiale de données

Scénario de Web-Mining à l'aide de CRISP-DM

La société d'eCommerce de notre exemple utilise plusieurs sources de données importantes, notamment :

Les logs Web. Les journaux d'accès bruts contiennent l'ensemble des informations sur la manière dont les clients naviguent sur le site Web. Dans le cadre du processus de préparation des données, vous aurez à supprimer des logs Web les références aux fichiers image et toute entrée dépourvue d'informations.

© Copyright IBM Corp. 1994, 2015

La base de données des achats. Lorsqu'un client soumet une commande, toutes les informations la concernant sont enregistrées. Les commandes figurant dans la base de données des achats doivent être mappées avec les sessions correspondantes des logs Web.

La base de données des produits. Les attributs des produits peuvent être utiles pour la détermination de produits « associés ». Les informations sur les produits doivent être mappées avec les commandes correspondantes.

La base de données des clients. Cette base de données contient des informations supplémentaires recueillies auprès des clients enregistrés. Ces enregistrements sont loin d'être complets, car beaucoup de clients ne remplissent pas les questionnaires. Les informations sur les clients doivent être mappées avec les sessions et les achats correspondants des logs Web.

Actuellement, la société ne projette pas d'acquérir de bases de données externes ni d'engager des frais pour la réalisation d'enquêtes, car ses analystes sont occupés à exploiter les données déjà en leur possession. Toutefois, l'avenir les amènera certainement à envisager un déploiement plus vaste des résultats de l'exploration de données, auquel cas l'acquisition de données démographiques supplémentaires concernant les clients non enregistrés pourrait être utile. La possession d'informations démographiques permet également de voir les divergences entre la clientèle de la société d'eCommerce et le client moyen effectuant ses achats sur le Web.

Elaboration d'un rapport sur la collecte des données

A l'aide des éléments obtenus à l'étape précédente, vous pouvez commencer à rédiger un rapport sur la collecte des données. Une fois terminé, ce rapport peut être ajouté au site Web du projet ou remis à vos collaborateurs. Il peut aussi être combiné aux rapports qui seront rédigés au cours des étapes suivantes : description et exploration des données, puis vérification de la qualité des données. Ces rapports guideront votre travail tout au long de la phase de préparation des données.

Description des données

S'il existe de nombreuses manières de décrire des données, la plupart des descriptions sont axées sur la quantité et la qualité des données : le volume de données disponibles et l'état de ces données. Le paragraphe ci-dessous répertorie certaines caractéristiques-clés à prendre en compte pour la description des données.

- **Quantité de données.** Dans la majorité des techniques de modélisation, le volume des données entraîne des compromis. Les grands jeux de données peuvent produire des modèles plus précis, mais augmentent également le temps de traitement. Voyez s'il est possible d'utiliser un sous-ensemble de données. Lors de la prise de notes en vue du rapport final, veillez à inclure des statistiques sur la taille de tous les jeux de données et à prendre en compte le nombre d'enregistrements et les champs (attributs) lors de la description des données.
- **Types de valeur.** Les données peuvent se présenter sous plusieurs formats, tels que le format **numérique**, **catégoriel** ou **booléen** (true/false (vrai/faux)). La prise en compte du type de valeur permet d'éviter certains problèmes durant la modélisation.
- **Méthodes de codage.** Les valeurs d'une base de données représentent souvent des caractéristiques, telles que le sexe ou le type de produit. Par exemple, un jeu de données peut utiliser les lettres *M* et *F* pour signifier *masculin* et *féminin*, tandis qu'un autre emploiera les valeurs numériques *1* et *2*. Notez tous les conflits entre les méthodes de codage dans le rapport sur les données.

Fort de ces connaissances, vous êtes maintenant prêt à rédiger le rapport de description des données et à faire part de vos constatations à un public plus important.

Exemple d'eCommerce : description de données

Scénario de Web-Mining à l'aide de CRISP-DM

Une application d'exploration du Web est amenée à traiter de nombreux enregistrements et attributs. Bien que la société d'eCommerce réalisant le projet d'exploration de données ait limité la portée de son étude initiale aux 30 000 clients environ qui se sont enregistrés sur le site, les logs Web contiennent encore des millions d'enregistrements.

La plupart des valeurs des sources de données sont de type symbolique, qu'il s'agisse de dates et d'heures, de pages Web consultées ou de réponses à des questions à choix multiples issues du questionnaire d'enregistrement. Certaines de ces variables serviront à créer des variables numériques, telles que le nombre de pages Web consultées et le temps passé sur le site Web. Les rares variables numériques figurant dans les sources de données comprennent la quantité commandée pour chaque produit, le montant dépensé en un achat, et les informations sur le poids et les dimensions des produits issues de la base de données des produits.

Le chevauchement entre les méthodes de codage des différentes sources de données est peu fréquent car les attributs de ces sources sont très différents. Les seules variables qui se chevauchent sont les « clés », telles que les ID client et les codes des produits. Les méthodes de codage de ces variables doivent être identiques d'une source de données à l'autre, faute de quoi la fusion des sources de données est impossible. Les données nécessitent donc une préparation supplémentaire afin que vous puissiez recoder ces champs-clés en vue de la fusion.

Elaboration d'un rapport de description des données

Afin de poursuivre de manière efficace le projet d'exploration de données, imaginez la valeur que représenterait un rapport de description des données précis produit à l'aide des mesures suivantes :

Quantité de données

- Quel est le format des données ?
- Identifiez la méthode utilisée pour capturer les données (par exemple, ODBC).
- Quelle est la taille de la base de données (en nombre de lignes et de colonnes) ?

Qualité des données

- Les données comprennent-elles des caractéristiques pertinentes pour la problématique métiere ?
- Quels sont les types de données présents (symbolique, numérique, etc.) ?
- Avez-vous calculé des statistiques de base pour les attributs-clés ? En quoi cela a-t-il permis d'éclaircir la problématique métier ?
- Pouvez-vous classer les attributs pertinents par ordre de priorité ? Si tel n'est pas le cas, pouvez-vous recourir à des analystes en informatique de gestion afin d'obtenir des éclaircissements ?

Exploration des données

Utilisez cette phase de CRISP-DM pour explorer les données à l'aide des tableaux, des graphiques et des autres outils de visualisation disponibles dans IBM SPSS Modeler. De telles analyses peuvent vous aider à atteindre l'objectif d'exploration de données instauré durant la phase de compréhension de l'entreprise. Elles permettent également de formuler des hypothèses et d'élaborer les tâches de transformation des données réalisées durant la préparation des données.

Exemple d'eCommerce : exploration des données

Scénario de Web-Mining à l'aide de CRISP-DM

Bien que la méthodologie CRISP-DM suggère d'effectuer à ce stade une exploration initiale des données, il est difficile, voire impossible, d'effectuer cette tâche avec des logs Web bruts, comme l'a constaté notre société d'eCommerce. Généralement, les données des logs Web doivent tout d'abord être traitées lors de la phase de préparation des données pour produire des données pouvant être explorées de manière utile. Cet écart par rapport à la méthodologie CRISP-DM souligne le fait que le traitement peut et doit être

personnalisé en fonction de vos besoins en matière d'exploration de données. CRISP-DM opère de manière cyclique, tandis que les data miners effectuent des va-et-vient entre les phases.

Si les logs Web doivent être traités avant l'exploration, les autres sources de données dont disposent la société d'eCommerce sont plus simples à explorer. L'exploration de la base de données des achats permet d'obtenir des récapitulatifs intéressants sur les clients, tels que leur provenance, les montants qu'ils dépensent et le nombre d'articles acquis par achat. Les récapitulatifs de la base de données des clients révèlent la répartition des réponses aux questions du questionnaire d'enregistrement.

L'exploration permet également de rechercher les erreurs dans les données. Si la plupart des sources de données sont générées automatiquement, les informations de la base de données des produits sont saisies manuellement. Certains récapitulatifs rapides des dimensions des produits répertoriés permettent de découvrir des fautes de saisie, telles que « 119 centimètres » (au lieu de « 19 centimètres »).

Elaboration d'un rapport sur l'exploration des données

A mesure que vous créez des graphiques et que vous calculez des statistiques à partir des données disponibles, commencez à formuler des hypothèses sur la manière dont ces données peuvent permettre d'atteindre les objectifs techniques et commerciaux.

Liste des tâches

Notez vos constatations afin de les inclure dans le rapport sur l'exploration des données. Prenez soin de répondre aux questions suivantes :

- Quels types d'hypothèse avez-vous formulés au sujet des données ?
- Quels attributs semblent prometteurs en vue d'une future analyse ?
- Vos explorations ont-elles révélé de nouvelles caractéristiques des données ?
- Comment ces explorations ont-elles modifié votre hypothèse initiale ?
- Pouvez-vous identifier des sous-jeux de données particuliers à utiliser ultérieurement ?
- Examinez à nouveau vos objectifs d'exploration de données. L'exploration a-t-elle entraîné une modification de ces objectifs ?

Vérification de la qualité des données

Les données sont rarement parfaites. En fait, la plupart des données contiennent des erreurs de codage, des valeurs manquantes ou d'autres types d'incohérence qui compliquent parfois l'analyse. Pour éviter les pièges, menez une analyse approfondie de la qualité des données disponibles avant la modélisation.

Les outils de création de rapports de IBM SPSS Modeler (tels que Data Audit, Table et autres noeuds de sortie) peuvent vous aider à rechercher les types de problème suivants :

- Les **données manquantes** comprennent les valeurs vides ou codées comme une absence de réponse (telles que *$null$*, *?* ou *999*).
- Les **erreurs de données** sont généralement des erreurs typographiques faites lors de la saisie des données.
- Les **erreurs de mesure** représentent notamment les données saisies correctement, mais basées sur une méthode de mesure erronée.
- Les **incohérences de codage** concernent généralement les unités de mesure non standard ou les incohérences dans les valeurs, telles que l'utilisation de *M* et de *masculin* pour le sexe.
- Les **métadonnées erronées** représentent notamment les discordances entre la signification apparente d'un champ et celle énoncée dans le nom ou la définition du champ.

Veillez à noter ces problèmes de qualité. Pour plus d'informations, voir la rubrique «Elaboration d'un rapport sur la qualité des données», à la page 17.

Exemple d'eCommerce : vérification de la qualité des données

Scénario de Web-Mining à l'aide de CRISP-DM

La vérification de la qualité des données est souvent effectuée au cours des processus de description et d'exploration. La société d'eCommerce a notamment rencontré les problèmes suivants :

Données manquantes. Les données manquantes connues comprennent les questionnaires non remplis par certains des utilisateurs enregistrés. Sans les informations supplémentaires issues du questionnaire, ces clients devront peut-être être exclus d'une partie des modèles ultérieurs.

Erreurs de données. La plupart des sources de données étant générées automatiquement, ce type d'erreur ne pose pas de problème majeur. Les erreurs typographiques figurant dans la base de données des produits peuvent être repérées durant le processus d'exploration.

Erreurs de mesure. La plus grande source potentielle des erreurs de mesure est le questionnaire. Si certains des éléments sont incorrects ou mal formulés, ils risquent de ne pas offrir les informations que la société d'eCommerce souhaite obtenir. A nouveau, durant le processus d'exploration, il est nécessaire de prêter une attention particulière aux éléments dont les réponses sont réparties singulièrement.

Elaboration d'un rapport sur la qualité des données

Suite à l'exploration et à la vérification de la qualité des données, vous êtes maintenant prêt à élaborer un rapport qui orientera la prochaine phase de CRISP-DM. Pour plus d'informations, voir la rubrique «Vérification de la qualité des données», à la page 16.

Liste des tâches

Comme mentionné précédemment, il existe plusieurs types de problème de qualité des données. Avant de passer à l'étape suivante, étudiez les problèmes de qualité suivants et prévoyez une solution. Apportez des réponses complètes dans le rapport sur la qualité des données.

- Avez-vous identifié des attributs manquants et des champs vides ? Si oui, ces valeurs manquantes ont-elles une signification ?
- L'orthographe présente-t-elle des incohérences pouvant engendrer des problèmes lors de fusions ou de transformations ultérieures ?
- Avez-vous exploré les écarts afin de déterminer s'il existe des « parasites » ou des phénomènes à analyser plus en profondeur ?
- Avez-vous vérifié la plausibilité des valeurs ? Relevez les conflits apparents (par exemple, des adolescents à revenus élevés).
- Avez-vous envisagé d'exclure les données sans impact sur vos hypothèses ?
- Les données sont-elles conservées dans des fichiers non hiérarchiques ? Si oui, les délimiteurs des différents fichiers sont-ils cohérents ? Chaque enregistrement contient-il le même nombre de champs ?

Prêt pour la prochaine étape ?

Avant de préparer les données en vue de la modélisation dans IBM SPSS Modeler, réfléchissez aux points suivants :

Quel est votre degré de compréhension des données ?

- Toutes les sources de données sont-elles clairement identifiées et consultées ? Avez-vous connaissance de problèmes ou de restrictions ?
- Avez-vous identifié des attributs-clés parmi les données disponibles ?
- Ces attributs vous ont-ils aidé à formuler des hypothèses ?
- Avez-vous noté la taille de toutes les sources de données ?

- Etes-vous en mesure d'utiliser un sous-ensemble de données si nécessaire ?
- Avez-vous calculé des statistiques de base pour chaque attribut intéressant ? En avez-vous tiré des informations intéressantes ?
- Avez-vous utilisé des graphiques exploratoires pour clarifier les attributs-clés ? Cet approfondissement a-t-il entraîné une reformulation de certaines de vos hypothèses ?
- Quels sont les problèmes de qualité des données de ce projet ? Avez-vous un plan pour résoudre ces problèmes ?
- Les étapes de la préparation des données vous apparaissent-elles clairement ? Par exemple, savez-vous quels sont les attributs à filtrer ou à sélectionner, et les sources de données à fusionner ?

Fort de cette compréhension de l'entreprise et des données, vous êtes prêt à utiliser IBM SPSS Modeler afin de préparer vos données en vue de la modélisation.

Chapitre 4. Préparation des données

Présentation de la préparation des données

La préparation des données est l'un des aspects les plus importants et les plus coûteux en temps de l'exploration de données. En fait, la préparation des données représente, selon les estimations, de 50 à 70 % du temps et des efforts consacrés à un projet. Le fait de consacrer une énergie suffisante aux phases initiales de compréhension de l'entreprise et de compréhension des données permet de réduire cette étape, mais la préparation et l'intégration des données en vue de l'exploration de données requièrent encore beaucoup d'efforts.

En fonction du type de société et de ses objectifs, la préparation des données comporte généralement les tâches suivantes :

- Fusion des ensembles et/ou des enregistrements de données
- Sélection d'un sous-ensemble de données exemple
- Agrégation des enregistrements
- Calcul de nouveaux attributs
- Tri des données en vue de la modélisation
- Suppression ou remplacement des blancs ou des valeurs manquantes
- Fractionnement en sous-ensembles d'apprentissage et de test

Sélection de données

En fonction de la collecte initiale de données réalisée dans la phase CRISP-DM précédente, vous pouvez commencer à choisir les données pertinentes pour vos objectifs d'exploration de données. En général, les données peuvent être sélectionnées de deux manières :

- **Sélection des enregistrements (lignes)** : implique des décisions concernant les comptes, les produits ou les clients à inclure.
- **Sélection des attributs ou des caractéristiques (colonnes)** : implique des décisions concernant l'utilisation de caractéristiques telles que le montant des transactions ou le revenu des ménages.

Exemple d'eCommerce : sélection de données

Scénario de Web-Mining à l'aide de CRISP-DM

La plupart des décisions concernant les données à sélectionner ont déjà été prises au cours des phases précédentes du processus d'exploration de données.

Sélection des éléments. La portée de l'étude initiale sera limitée aux 30 000 clients (environ) qui se sont enregistrés sur le site. Il convient donc de définir des filtres afin d'exclure les achats et les logs Web des clients non enregistrés. D'autres filtres doivent être instaurés afin de supprimer des logs Web les appels de fichiers image, ainsi que toute entrée dépourvue d'informations.

Sélection d'attributs. La base de données des achats contient des informations confidentielles sur les clients de la société d'eCommerce. Il est donc important de filtrer les attributs tels que le nom, l'adresse, le numéro de téléphone et le numéro de carte de crédit des clients.

Inclusion ou exclusion de données

Lorsque vous décidez des sous-jeux de données à inclure ou à exclure, veillez à fournir des informations sur la logique qui sous-tend vos décisions.

© Copyright IBM Corp. 1994, 2015

Questions à prendre en considération

- Un attribut donné est-il pertinent compte tenu des objectifs d'exploration de données ?
- La qualité d'un jeu de données ou d'un attribut particulier peut-elle nuire à la validité des résultats ?
- Ces données peuvent-elles être récupérées ?
- Existe-t-il des contraintes liées à l'utilisation de champs tels ceux indiquant le *sexe* ou la *race* ?

Les décisions prises ici diffèrent-elles des hypothèses formulées dans la phase de compréhension des données ? Si oui, veillez à fournir un raisonnement complet dans le rapport du projet.

Nettoyage des données

Lorsque vous nettoyez les données, vous devez examiner en profondeur les problèmes des données que vous avez choisi d'inclure dans l'analyse. Les noeuds d'opérations sur les enregistrements et sur les champs de IBM SPSS Modeler permettent de nettoyer les données de plusieurs manières.

Tableau 2. Nettoyage des données

Problème posé par les données	**Solution possible**
Données manquantes	Excluez les lignes ou les caractéristiques, ou insérez une valeur estimée dans les blancs.
Erreurs dans les données	Procédez de manière logique pour découvrir manuellement les erreurs et les corriger, ou excluez les caractéristiques.
Codage des incohérences	Décidez d'une méthode de codage unique, puis convertissez et remplacez les valeurs.
Métadonnées erronées ou manquantes	Examinez manuellement les champs suspects et recherchez la signification correcte.

Le rapport sur la qualité des données préparé au cours de la phase de compréhension des données contient des informations sur les types de problème propres à vos données. Vous pouvez l'utiliser comme point de départ pour la manipulation des données dans IBM SPSS Modeler.

Exemple d'eCommerce : nettoyage des données

Scénario de Web-Mining à l'aide de CRISP-DM

La société d'eCommerce utilise le processus de nettoyage des données pour résoudre les problèmes indiqués dans le rapport sur la qualité des données.

Données manquantes. Les clients qui n'ont pas rempli le questionnaire en ligne risquent d'être exclus de certains modèles par la suite. La société d'eCommerce pourrait redemander à ces clients de remplir le questionnaire, mais cela représenterait pour elle un investissement en temps et en argent qu'elle ne peut pas se permettre. Par contre, la société peut modéliser les divergences des achats entre les clients qui ont répondu et ceux qui n'ont pas répondu au questionnaire. Si ces deux ensembles de clients ont des habitudes d'achat similaires, l'absence d'une partie des questionnaires est moins problématique.

Erreurs de données. Les erreurs relevées durant le processus d'exploration peuvent être corrigées ici. Toutefois, la nécessité de saisir des données correctes est généralement appliquée sur le site Web avant l'envoi d'une page à la base de données d'arrière-plan par le client.

Erreurs de mesure. La mauvaise formulation d'éléments du questionnaire peut grandement affecter la qualité des données. Comme pour les questionnaires manquants, il s'agit là d'un problème délicat, car l'argent ou le temps nécessaire à la collecte de réponses à une question de substitution peuvent manquer. Pour les éléments problématiques, la meilleure solution consiste parfois à revenir au processus de sélection pour ôter ces éléments des analyses ultérieures.

Elaboration d'un rapport sur le nettoyage des données

L'élaboration d'un rapport sur les tâches de nettoyage des données effectuées est essentielle au suivi des modifications des données. Vous bénéficierez ainsi des informations disponibles sur votre travail pour les prochains projets d'exploration de données.

Liste des tâches

Lors de l'élaboration du rapport, tenez compte des questions suivantes :

- Quels sont les types de parasite qui se produisent dans les données ?
- Quelles approches avez-vous utilisées pour supprimer ces parasites ? Quelles techniques se sont révélées concluantes ?
- Existe-t-il des cas ou des attributs qui n'ont pas pu être récupérés ? Veillez à noter les données exclues à cause de parasites.

Construction de nouvelles données

Vous serez souvent amené à construire de nouvelles données. Par exemple, il peut être utile de créer une nouvelle colonne signalant l'achat d'une extension de garantie pour chaque transaction. Ce nouveau champ, *garantie_achetée*, peut facilement être généré à l'aide d'un noeud Binariser dans IBM SPSS Modeler.

Les nouvelles données peuvent être construites de deux manières :

- Calcul des attributs (colonnes ou caractéristiques)
- Génération des enregistrements (lignes)

IBM SPSS Modeler propose un grand nombre de méthodes de construction de données mettant en oeuvre ses noeuds d'opérations sur les champs et sur les enregistrements.

Exemple d'eCommerce : construction de données

Scénario de Web-Mining à l'aide de CRISP-DM

Le traitement de logs Web peut créer de nombreux attributs. Pour les événements enregistrés dans les journaux, la société d'eCommerce peut souhaiter créer des horodatages, identifier les visiteurs et les sessions, et noter les pages consultées et le type d'activité représenté par l'événement. Certaines de ces variables seront utilisées pour créer d'autres attributs, tel que le temps écoulé entre les événements d'une même session.

D'autres attributs peuvent être créés suite à une fusion ou à une autre restructuration des données. Par exemple, lorsque les logs Web recensant les événements par ligne sont « agrégés » afin que chaque ligne corresponde à une session, de nouveaux attributs enregistrant le nombre total d'actions, le total du temps passé et le total des achats effectués durant la session sont créés. Lors de la fusion des logs Web avec la base de données des clients, de sorte à faire correspondre chaque ligne à un client, de nouveaux attributs enregistrant le nombre de sessions, le nombre total d'actions, le total du temps passé et le total des achats effectués par chaque client sont créés.

Une fois les données construites, la société d'eCommerce entreprend un processus d'exploration afin de s'assurer que la création des données a été effectuée correctement.

Calcul d'attributs

Dans IBM SPSS Modeler, vous pouvez calculer de nouveaux attributs à l'aide des noeuds d'opérations sur les champs suivants :

- Créez des champs calculés à partir de champs existants à l'aide d'un **noeud Calculer**.
- Créez un champ indicateur à l'aide d'un **noeud Binariser**.

Liste des tâches

- Lors du calcul des attributs, tenez compte des impératifs en matière de données pour la modélisation. L'algorithme de modélisation nécessite-t-il un type de données particulier, tel que des données numériques ? Si oui, effectuez les transformations nécessaires.
- Les données doivent-elles être normalisées avant la modélisation ?
- Les attributs manquants peuvent-ils être construits par agrégation, utilisation d'une moyenne ou induction ?
- D'après vos connaissances du contexte, existe-t-il des faits importants (tels que le temps passé sur le site Web) pouvant être calculés à partir des champs existants ?

Intégration des données

Il n'est pas inhabituel d'avoir plusieurs sources de données pour un même ensemble de questions d'ordre commercial. Par exemple, pour un même ensemble de clients, vous pouvez avoir accès à des données sur les prêts hypothécaires, ainsi qu'à des données démographiques achetées. Si ces jeux de données utilisent un identifiant unique identique (tel que le numéro de sécurité sociale), vous pouvez les fusionner dans IBM SPSS Modeler en utilisant ce champ-clé.

Deux méthodes principales existent pour l'intégration de données :

- La **fusion** de données, qui implique la fusion de deux jeux de données possédant des enregistrements semblables mais des attributs différents. Ces données sont fusionnées à l'aide d'un même identificateur-clé pour chaque enregistrement (tel que l'ID client). Les données qui en résultent ont un plus grand nombre de colonnes ou de caractéristiques.
- L'**ajout** de données, qui implique l'intégration de plusieurs jeux de données possédant des attributs semblables mais des enregistrements différents. Ces données sont intégrées en fonction d'un champ identique (tel qu'un nom de produit ou une durée de contrat).

Exemple d'eCommerce : intégration de données

Scénario de Web-Mining à l'aide de CRISP-DM

Lorsque plusieurs sources de données existent, la société d'eCommerce peut intégrer ces données de plusieurs manières :

- **Ajout d'attributs de client et de produit aux données d'événement.** Pour modéliser les événements des logs Web à l'aide d'attributs issus d'autres bases de données, tous les ID client, numéros de référence produit et numéros de bon de commande associés à chaque événement doivent être identifiés correctement, et les attributs correspondants fusionnés avec les logs Web traités. Le fichier fusionné réplique les informations concernant le client et le produit chaque fois qu'un client ou un produit est associé à un événement.
- **Ajout d'informations sur les achats et les logs Web aux données des clients.** Pour modéliser la valeur d'un client, les informations sur ses achats et sa session doivent être prélevées dans les bases de données appropriées, puis totalisées et fusionnées avec la base de données des clients. Cette opération implique la création d'attributs, comme mentionné dans le processus de construction de données.

Après l'intégration des bases de données, la société d'eCommerce entreprend un processus d'exploration afin de s'assurer que la fusion des données a été correctement réalisée.

Tâches d'intégration

L'intégration des données peut s'avérer complexe si vous n'avez pas passé suffisamment de temps à acquérir une bonne compréhension de vos données. Réfléchissez aux éléments et aux attributs qui semblent les plus pertinents au vu des objectifs d'exploration de données, puis commencez l'intégration de vos données.

Liste des tâches

- A l'aide des noeuds Fusionner ou Ajouter de IBM SPSS Modeler, intégrez les jeux de données jugés utiles à la modélisation.
- Pensez à enregistrer le résultat avant de passer à la modélisation.
- Après la fusion, les données peuvent être simplifiées par **agrégation** des valeurs. L'agrégation signifie que de nouvelles valeurs sont calculées en récapitulant les informations issues de plusieurs enregistrements et/ou tables.
- Vous aurez peut-être aussi à générer de nouveaux enregistrements (tels que la moyenne des abattements calculée d'après les déclarations de revenus de plusieurs années).

Formatage de données

Avant de commencer la création d'un modèle, il est utile de vérifier si certaines techniques nécessitent l'application d'un format ou d'un ordre particulier aux données. Par exemple, un algorithme de séquence nécessite fréquemment un tri préalable des données avant l'exécution du modèle. Même si le modèle est en mesure de réaliser ce tri pour vous, vous pouvez parfois réduire le temps de traitement en utilisant un noeud Trier avant la modélisation.

Liste des tâches

Tenez compte des questions suivantes lors du formatage des données :

- Quels modèles prévoyez-vous d'utiliser ?
- Ces modèles nécessitent-ils l'application d'un format ou d'un ordre particulier aux données ?

Si des modifications sont de rigueur, les outils de traitement de IBM SPSS Modeler peuvent vous aider à effectuer les manipulations de données nécessaires.

Prêt pour la modélisation ?

Avant de créer des modèles dans IBM SPSS Modeler, assurez-vous d'avoir répondu aux questions suivantes.

- Toutes les données sont-elles accessibles à partir de IBM SPSS Modeler ?
- Votre exploration et votre compréhension initiales vous ont-elles permis de sélectionner des sous-jeux de données pertinents ?
- Avez-vous nettoyé les données de manière efficace ou retiré les éléments irrécupérables ? Fournissez des informations sur toutes vos décisions dans le rapport final.
- Les jeux de données multiples sont-ils correctement intégrés ? La fusion a-t-elle entraîné des problèmes nécessitant un complément d'informations ?
- Avez-vous étudié les impératifs des outils de modélisation que vous prévoyez d'utiliser ?
- Pouvez-vous résoudre certains problèmes de formatage avant la modélisation ? Sont considérés les problèmes liés au formatage requis et les tâches susceptibles de réduire la durée de la modélisation.

Si vous pouvez répondre aux questions ci-dessus, vous êtes prêt à affronter le coeur même de l'exploration de données : la modélisation.

Chapitre 5. Modélisation

Présentation de la modélisation

C'est à ce stade que vos efforts commencent à être récompensés. Les données que vous avez préparées minutieusement sont importées dans les outils d'analyse de IBM SPSS Modeler et les résultats commencent à éclaircir la problématique commerciale posée lors de la compréhension de l'entreprise.

La modélisation est généralement effectuée en utilisant plusieurs itérations. Généralement, les data miners exécutent plusieurs modèles en utilisant les paramètres par défaut, puis affinent ces derniers ou reviennent à la phase de préparation des données pour effectuer les manipulations requises par le modèle de leur choix. Il est rare qu'une question d'exploration de données soit résolue de façon satisfaisante avec un seul modèle et une seule exécution. C'est pourquoi l'exploration de données est si intéressant ; il existe de nombreuses façons de traiter un problème donné et IBM SPSS Modeler offre à cette fin une grande variété d'outils.

Sélection de techniques de modélisation

Vous avez sûrement déjà une idée des types de modélisation pouvant répondre au mieux aux exigences de votre entreprise. Il est maintenant temps de choisir définitivement ceux que vous souhaitez utiliser. Le choix du modèle le plus adéquat sera généralement basé sur les critères suivants :

- **Les types de données disponibles pour l'exploration.** Par exemple, les champs intéressants sont-ils catégoriels (symboliques) ?
- **Vos objectifs d'exploration de données.** Souhaitez-vous simplement obtenir un aperçu des entrepôts de données transactionnelles et découvrir des motifs d'achat intéressants ? Ou voulez-vous produire un score indiquant, par exemple, la tendance à ne pas rembourser un prêt étudiant ?
- **Les exigences de modélisation particulières.** Le modèle requiert-il une taille ou un type de données particuliers ? Avez-vous besoin d'un modèle dont les résultats sont facilement présentables ?

Pour plus d'informations sur les types de modèle de IBM SPSS Modeler et leurs exigences, consultez la documentation IBM SPSS Modeler ou l'aide en ligne.

Exemple d'eCommerce : techniques de modélisation

Les techniques de modélisation utilisées par la société d'eCommerce sont établies en fonction des objectifs d'exploration de données de l'entreprise :

Recommandations améliorées. Dans leur forme la plus simple, elles impliquent la classification des commandes achat pour déterminer les produits les plus souvent achetés ensemble. Des données client, et même des enregistrements de visite, peuvent être ajoutés afin d'obtenir des résultats plus étoffés. Les techniques de classification TwoStep ou du réseau Kohonen sont adaptées à ce type de modélisation. Les clusters peuvent ensuite être profilés à l'aide d'un ensemble de règles C5.0 afin de déterminer les recommandations les mieux adaptées à tel moment de la visite d'un client.

Navigation sur le site améliorée. Pour l'instant, la société d'eCommerce se concentre sur l'identification des pages souvent utilisées mais difficiles d'accès pour les utilisateurs. Cette opération sous-entend l'application d'un algorithme d'ordonnancement aux logs Web de façon à générer les « chemins uniques » suivis par les clients sur le site Web. Il faut ensuite rechercher les sessions comportant un grand nombre de pages visitées sans (ou avant) qu'une action soit entreprise. Plus tard, au cours d'une analyse plus approfondie, les techniques de classification pourront être utilisées pour identifier différents « types » de visite et de visiteur, et le contenu du site pourra être organisé et présenté en fonction d'un type donné.

© Copyright IBM Corp. 1994, 2015

Choix des techniques de modélisation appropriées

De nombreuses techniques de modélisation sont disponibles avec IBM SPSS Modeler. Il arrive souvent que les data miners utilisent plusieurs techniques pour traiter un problème à partir de perspectives différentes.

Liste des tâches

Lorsque vous décidez des modèles à utiliser, étudiez les points suivants pour savoir s'ils ont une incidence sur votre choix :

- Le modèle exige-t-il que les données soient divisées en ensembles de test et d'apprentissage ?
- Avez-vous suffisamment de données pour produire des résultats fiables avec un modèle donné ?
- Le modèle exige-t-il un certain niveau de qualité des données ? Vos données actuelles répondent-elles à ce niveau ?
- Le type de vos données est-il approprié pour un modèle spécifique ? Si ce n'est pas le cas, pouvez-vous effectuer les conversions nécessaires en utilisant des noeuds de manipulation de données ?

Pour plus d'informations sur les types de modèle de IBM SPSS Modeler et leurs exigences, consultez la documentation IBM SPSS Modeler ou l'aide en ligne.

Hypothèses de modélisation

Au fur et à mesure que vous affinez vos outils de modélisation, prenez des notes sur le processus de prise de décision. Apportez toutes les informations nécessaires aux hypothèses de données, ainsi qu'aux manipulations de données effectuées pour répondre aux exigences du modèle.

Par exemple, les noeuds Régression logistique et Réseau de neurones exigent que les types de données soient entièrement **instanciés** (les types de données sont connus) avant l'exécution. Autrement dit, vous devrez ajouter un noeud type au flux et l'exécuter pour explorer les données avant de créer et d'exécuter un modèle. De même, pour les modèles prédictifs, tels que C5.0, il peut être utile de rééquilibrer les données lors de la prévision de règles pour des événements rares. Lorsque vous effectuez ce type de prévision, vous pouvez obtenir de meilleurs résultats en insérant un noeud Equilibrer dans le flux et en incorporant le sous-ensemble le plus équilibré dans le modèle.

N'oubliez pas de fournir des informations sur ces décisions.

Génération d'une conception de test

La dernière étape avant de créer le modèle consiste à considérer de nouveau la façon dont les résultats du modèle seront testés. La génération d'une conception de test complète comporte deux étapes :

- La description des critères de « qualité d'ajustement » d'un modèle
- La définition des données sur lesquelles ces critères seront testés

Les **qualités d'ajustement** d'un modèle peuvent être mesurées de différentes façons. Dans le cas des modèles supervisés, tels que C5.0 et C&RT, les mesures des qualités d'ajustement évaluent généralement le taux d'erreur d'un modèle particulier. Dans le cas des modèles non supervisés, tels que les réseaux de clusters Kohonen, les mesures peuvent inclure des critères tels que la facilité d'interprétation, le déploiement ou le temps de traitement nécessaire.

N'oubliez pas que la création de modèles est un processus itératif. Vous testerez généralement les résultats de plusieurs modèles avant de décider de ceux à utiliser et à déployer.

Elaboration d'une conception de test

La conception de test décrit les étapes que vous suivrez pour tester les modèles produits. Etant donné que la modélisation est un processus itératif, il est important de savoir quand arrêter l'ajustement des paramètres et essayer une autre méthode ou un autre modèle.

Liste des tâches

Lors de la création d'une conception de test, prenez en compte les questions suivantes :

- Quelles données seront utilisées pour tester les modèles ? Avez-vous partitionné les données en ensembles d'apprentissage/de test ? (Il s'agit d'une approche fréquemment utilisée dans la modélisation.)
- Comment pouvez-vous mesurer la réussite des modèles supervisés (C5.0 par exemple) ?
- Comment pouvez-vous mesurer la réussite des modèles non supervisés (réseaux de clusters Kohonen, par exemple) ?
- Combien de fois souhaitez-vous réexécuter un modèle avec des paramètres ajustés avant d'essayer un autre type de modèle ?

Exemple d'eCommerce : conception de test

Scénario de Web-Mining à l'aide de CRISP-DM

Les critères utilisés pour l'évaluation des modèles dépendent des modèles considérés et des objectifs d'exploration de données :

Recommandations améliorées. Tant que les recommandations améliorées ne sont pas présentées aux clients en personne, il n'existe pas de moyen purement objectif pour les évaluer. Néanmoins, la société d'eCommerce peut exiger que les règles générant les recommandations soient aussi simples que possible d'un point de vue commercial. Les règles doivent néanmoins être assez élaborées pour générer différentes recommandations en fonction des différents clients et sessions.

Navigation sur le site améliorée. Après avoir déterminé les pages auxquelles les clients accèdent sur le site Web, la société d'eCommerce peut évaluer de façon objective la nouvelle conception du site en termes de facilité d'accès à des pages importantes. Cependant, tout comme dans le cas des recommandations, il est difficile d'évaluer par avance comment les clients s'adapteront à la nouvelle organisation du site. Si vous disposez de suffisamment de temps et du budget nécessaire, il peut s'avérer utile d'effectuer des tests de convivialité.

Création des modèles

A ce stade, vous devriez être bien préparé pour créer les modèles que vous avez étudiés pendant si longtemps. Prenez le temps de tester plusieurs modèles avant de tirer des conclusions fermes et définitives. La plupart des data miners créent plusieurs modèles et comparent les résultats avant de les déployer ou de les intégrer.

Gardez une trace des données et des paramètres utilisés pour chaque modèle afin de suivre l'évolution des opérations que vous effectuez avec les différents modèles. Ceci vous aidera à discuter des résultats avec d'autres personnes et à retrouver la trace des opérations effectuées, le cas échéant. A la fin du processus de création des modèles, vous disposerez de trois types d'informations à utiliser dans les décisions d'exploration de données :

- Les **valeurs des paramètres**, qui comprennent les notes que vous avez prises concernant les paramètres aboutissant aux meilleurs résultats.
- Les **modèles** réels produits.
- Les **descriptions des résultats du modèle**, qui incluent les problèmes de performances et de données rencontrés lors de l'exécution du modèle et de l'exploration de ses résultats.

Exemple d'eCommerce : création de modèles

Scénario de Web-Mining à l'aide de CRISP-DM

Recommandations améliorées. Les classifications sont produites pour différents niveaux d'intégration des données : de la base de données des achats à l'intégration des informations connexes sur les clients et les sessions. Pour chaque niveau d'intégration, les classifications sont produites en utilisant des valeurs de paramètres différentes pour les algorithmes TwoStep et les algorithmes de réseau Kohonen. Pour chacune de ces classifications, quelques ensembles de règles C5.0 sont générés avec différentes valeurs de paramètres.

Navigation sur le site améliorée. Le noeud de modélisation Séquence est utilisé pour générer les chemins suivis par les clients. L'algorithme autorise la définition d'un critère de prise en charge minimale, ce qui est utile pour se concentrer sur les chemins d'accès client les plus courants. Différentes valeurs ont été essayées avec ce paramètre.

Valeurs des paramètres

La plupart des techniques de modélisation disposent d'un grand nombre de valeurs ou de paramètres pouvant être ajustés pour contrôler le processus de modélisation. Par exemple, les arbres décision peuvent être contrôlés par l'ajustement de la profondeur de l'arbre, des séparations et de nombreuses autres valeurs. En général, les modèles sont créés avec les options par défaut, puis les paramètres sont affinés au cours des sessions suivantes.

Une fois que vous avez déterminé les paramètres produisant les résultats les plus précis, n'oubliez pas d'enregistrer le flux et les noeuds de modèle généré. Il peut également être utile de noter les valeurs optimales lorsque vous décidez d'automatiser ou de recréer le modèle avec de nouvelles données.

Exécution des modèles

Dans IBM SPSS Modeler, l'exécution des modèles est une tâche simple. Une fois que vous avez inséré le noeud du modèle dans le flux et édité certains paramètres, exécutez le modèle pour produire des résultats consultables. Les résultats apparaissent dans le navigateur Modèles générés situé à droite de l'espace de travail. Cliquez avec le bouton droit de la souris sur un modèle pour parcourir les résultats. Avec la plupart des modèles, vous pouvez insérer le modèle généré dans le flux pour mieux évaluer et déployer les résultats. Vous pouvez également enregistrer les modèles dans IBM SPSS Modeler pour les réutiliser plus tard.

Description des modèles

Lorsque vous examinez les résultats d'un modèle, prenez des notes sur cette expérience de modélisation. Vous pouvez enregistrer ces notes avec le modèle lui-même à l'aide de la boîte de dialogue des annotations de noeud ou de l'outil de projet.

Liste des tâches

Pour chaque modèle, notez les informations qui répondent aux questions suivantes :

- Pouvez-vous tirer des conclusions significatives de ce modèle ?
- Le modèle révèle-t-il de nouvelles perspectives ou des motifs inhabituels ?
- Avez-vous rencontré des problèmes lors de l'exécution du modèle ? La durée du traitement était-elle satisfaisante ?
- Le modèle a-t-il rencontré des problèmes de qualité des données, tels qu'un nombre élevé de valeurs manquantes ?
- Y a-t-il eu des incohérences de calcul à mettre en évidence ?

Evaluation du modèle

A présent que vous disposez d'un ensemble de modèles initiaux, analysez-les en détail pour déterminer ceux qui sont suffisamment précis ou efficaces pour être dits finaux. Un modèle final peut désigner un modèle « prêt pour le déploiement » ou un modèle « illustrant des motifs intéressants ». La consultation du plan de test créé précédemment peut vous aider à effectuer cette évaluation en respectant le point de vue de l'entreprise.

Evaluation complète d'un modèle

Pour chaque modèle étudié, il est judicieux d'effectuer une évaluation méthodique basée sur les critères du plan de test. C'est ici que vous pouvez ajouter le modèle généré au flux, et utiliser des graphiques Evaluation ou des noeuds Analyse pour analyser l'efficacité des résultats. Vous devez également déterminer si les résultats sont logiques ou s'ils sont trop simplistes pour vos objectifs commerciaux (par exemple, une séquence révélant des achats tels que vin > vin > vin).

Une fois l'évaluation terminée, classez les modèles en fonction de critères objectifs (exactitude du modèle) et subjectifs (facilité d'utilisation ou interprétation des résultats).

Liste des tâches

- A l'aide des outils d'exploration de données de IBM SPSS Modeler, tels que les graphiques Evaluation, les noeuds Analyse ou les graphiques de validation croisée, évaluez les résultats de votre modèle.
- Passez les résultats en revue en fonction de votre compréhension du problème métier. Consultez des analystes de données ou d'autres experts pouvant approfondir la pertinence de résultats particuliers.
- Déterminez si les résultats d'un modèle sont facilement déployables. L'entreprise exige-t-elle que les résultats soient déployés sur le Web ou renvoyés à l'entrepôt de données ?
- Analysez l'impact des résultats sur vos critères de réussite. Répondent-ils aux objectifs établis lors de la phase de compréhension de l'entreprise ?

Si vous êtes parvenu à traiter les problèmes susmentionnés et pensez que les modèles actuels répondent aux objectifs, vous pouvez passer à une évaluation plus approfondie des modèles et au déploiement final. Si ce n'est pas le cas, utilisez ce que vous avez appris et réexécutez les modèles en ajustant les valeurs des paramètres.

Exemple d'eCommerce : évaluation des modèles

Scénario de Web-Mining à l'aide de CRISP-DM

Recommandations améliorées. L'un des réseaux Kohonen et une classification TwoStep fournissent des résultats satisfaisants. La société d'eCommerce hésite entre les deux. A la longue, l'entreprise espère utiliser les deux, en respectant les recommandations communes aux deux techniques et étudiant plus en détail les situations dans lesquelles elles diffèrent. Si la société d'eCommerce dispose d'une connaissance métier appliquée et qu'elle poursuit ses efforts, elle peut développer d'autres règles pour résoudre les différences entre les deux techniques.

La société d'eCommerce constate également avec étonnement que les résultats incluant les informations de session sont très satisfaisants. Il semble que des recommandations puissent être liées à la navigation sur le site. Un ensemble de règles définissant les prochaines pages que le client visitera pourrait être utilisé en temps réel pour modifier le contenu du site directement pendant la navigation du client.

Navigation sur le site améliorée. Le modèle Séquence offre à la société d'eCommerce un niveau de confiance élevé concernant la prédiction de certains chemins d'accès client, produisant des résultats suggérant un nombre raisonnable de changements à apporter à la conception du site.

Suivi des paramètres révisés

En vous basant sur ce que vous avez appris lors de l'évaluation du modèle, examinez de nouveau les modèles. Deux possibilités s'offrent à vous :

- Ajuster les paramètres des modèles existants.
- Choisir un autre modèle pour résoudre votre problème d'exploration de données.

Dans les deux cas, vous reviendrez à la tâche de création de modèles et la répéterez jusqu'à ce que les résultats soient satisfaisants. N'hésitez pas à répéter cette étape. Il est très fréquent que les data miners évaluent et réexécutent les modèles plusieurs fois avant de trouver celui qui répond à leurs exigences. Voilà une bonne raison de créer plusieurs modèles à la fois et de comparer leurs résultats avant d'ajuster les paramètres de chacun d'entre eux.

Prêt pour la prochaine étape ?

Avant de passer à l'évaluation finale des modèles, examinez votre évaluation initiale afin de savoir si elle était assez complète.

Liste des tâches

- Comprenez-vous les résultats des modèles ?
- Les résultats du modèle sont-ils compréhensibles d'un point de vue purement logique ? Existe-t-il des incohérences nécessitant une exploration plus approfondie ?
- Par rapport à votre première analyse, les résultats semblent-ils répondre à la problématique commerciale de l'entreprise ?
- Avez-vous utilisé des noeuds Analyse, et des graphiques de lift ou des graphiques de gains pour comparer et évaluer l'exactitude des modèles ?
- Avez-vous exploré plusieurs types de modèle et comparé les résultats ?
- Les résultats de votre modèle sont-ils déployables ?

Si les résultats de votre modélisation des données paraissent précis et pertinents, effectuez une évaluation plus approfondie avant de passer au déploiement final.

Chapitre 6. Evaluation

Présentation de l'évaluation

A ce stade, vous avez réalisé la plus grande partie de votre projet d'exploration de données. Vous avez également déterminé, lors de l'étape de modélisation, que les modèles créés sont techniquement corrects et efficaces en fonction des **critères de réussite de l'exploration de données** définis précédemment.

Néanmoins, avant de poursuivre, vous devez évaluer les résultats de vos efforts en utilisant les **critères de réussite commerciale** établis au début du projet. Cette étape est primordiale car elle permet de vous assurer que l'entreprise peut utiliser les résultats que vous avez obtenus. L'exploration de données produit deux types de résultat :

- Les **modèles** finaux sélectionnés au cours de la phase précédente de CRISP-DM.
- Les conclusions ou déductions tirées des modèles eux-mêmes, ainsi que du processus d'exploration de données. Elles sont appelées **constatations**.

Evaluation des résultats

A ce stade, vous formalisez votre évaluation concernant la conformité des résultats du projet avec les critères de réussite commerciale. Cette étape exige une compréhension claire des objectifs commerciaux énoncés. Par conséquent, faites participer les principaux décideurs à l'évaluation du projet.

Liste des tâches

Vous devez d'abord fournir une évaluation complète concernant la conformité des résultats de l'exploration de données avec les critères de réussite commerciale. Considérez les questions suivantes dans votre rapport :

- Vos résultats sont-ils énoncés de façon claire et sous une forme facilement présentable ?
- Existe-t-il des constatations originales ou uniques à mettre en évidence ?
- Pouvez-vous classer les modèles et les constatations par ordre d'applicabilité aux objectifs commerciaux ?
- De façon générale, comment ces résultats répondent-ils aux objectifs commerciaux de votre entreprise ?
- Quelles nouvelles questions vos résultats ont-ils soulevées ? Comment formuleriez-vous ces questions en termes commerciaux ?

Une fois que vous avez évalué les résultats, dressez une liste des modèles approuvés à inclure dans le rapport final. Cette liste doit contenir les modèles satisfaisant à la fois les objectifs d'exploration de données et les objectifs commerciaux de l'entreprise.

Exemple d'eCommerce : évaluation des résultats

Scénario de Web-Mining à l'aide de CRISP-DM

Les résultats généraux de la première expérience d'exploration des données de la société d'eCommerce sont assez faciles à communiquer d'un point de vue commercial : l'étude a abouti à ce qui apparaît comme de meilleures recommandations pour les produits et une meilleure conception du site. La conception améliorée du site est basée sur les séquences de navigation du client, qui indiquent les caractéristiques du site recherchées par le client, mais requérant de nombreuses étapes. Etant donné que les règles de décision peuvent être très compliquées, il est plus difficile de prouver que les recommandations sur les produits sont meilleures. Pour produire le rapport final, les analystes essaieront d'identifier quelques tendances générales dans les ensembles de règles, pouvant être expliquées plus facilement.

© Copyright IBM Corp. 1994, 2015

Classement des modèles. Etant donné que plusieurs modèles initiaux semblaient être utiles d'un point de vue commercial, le classement au sein de ce groupe a été basé sur des critères statistiques, la facilité d'interprétation et la diversité. Par conséquent, le modèle fournissait différentes recommandations en fonction des situations.

Nouvelles questions. La question la plus importante mise en évidence par l'étude est : « Comment la société d'eCommerce peut-elle en savoir plus sur ses clients ? » Les informations de la base de données sur les clients jouent un rôle important dans la création des clusters de recommandations. Si des règles spéciales sont disponibles pour les recommandations faites aux clients dont les informations sont manquantes, ces recommandations sont d'ordre plus général que celles émises pour les clients enregistrés.

Processus de révision

Les méthodologies efficaces prévoient généralement du temps pour réfléchir sur les points positifs et négatifs du processus qui vient de se terminer. L'exploration de données fonctionne de la même manière. Une partie du processus CRISP-DM consiste à tirer des leçons de votre expérience de façon à ce que les futurs projets d'exploration de données soient plus efficaces.

Liste des tâches

Vous devez d'abord récapituler les activités et les décisions pour chaque phase, en incluant les étapes de préparation des données, la création des modèles, etc. Ensuite, pour chaque phase, vous devez tenir compte des questions suivantes et émettre des propositions d'amélioration :

- Cette étape a-t-elle contribué à la valeur des résultats finaux ?
- Existe-t-il des moyens de simplifier ou d'améliorer cette étape ou opération particulière ?
- Quelles ont été les erreurs ou les échecs rencontrés au cours de cette phase ? Comment peuvent-ils être évités la prochaine fois ?
- Avez-vous constaté que des modèles particuliers ne présentaient aucune perspective d'avenir ? Existe-t-il des moyens de prévoir ces impasses et, par conséquent, de mieux concentrer les efforts ?
- Avez-vous eu des surprises (bonnes ou mauvaises) pendant cette phase ? Avec du recul, existe-t-il un moyen de prédire ces événements ?
- Des décisions ou des stratégies alternatives auraient-elles pu être utilisées lors d'une phase donnée ? Notez ces alternatives en vue de futurs projets d'exploration de données.

Exemple d'eCommerce : rapport de révision

Scénario de Web-Mining à l'aide de CRISP-DM

Après avoir révisé le processus du projet d'exploration de données initial, la société d'eCommerce est plus à même de comprendre les relations qui unissent les étapes du processus. D'abord hésitant à utiliser la stratégie du « retour en arrière » dans le processus CRISP-DM, la société d'eCommerce est désormais consciente que la nature cyclique du processus augmente son efficacité. La révision du processus a également permis à la société d'eCommerce de comprendre les points suivants :

- Un retour au processus d'exploration est toujours justifié si quelque chose d'inhabituel apparaît dans une autre phase du processus CRISP-DM.
- La préparation des données, notamment celle des logs Web, requiert de la patience car elle demande beaucoup de temps.
- Il est primordial de rester concentré sur le problème métier car une fois les données prêtes pour l'analyse, il est très facile de commencer à créer des modèles en perdant de vue le problème de base.
- Une fois la phase de modélisation terminée, la compréhension de l'entreprise est encore plus importante lorsqu'il s'agit de décider comment mettre en oeuvre les résultats et de déterminer quelles nouvelles études sont justifiées.

Détermination des étapes suivantes

A ce stade, vous avez produit des résultats et évalué votre expérience d'exploration de données. Vous vous demandez peut-être **quelles sont les étapes suivantes ?** Cette phase vous aide à répondre à cette question en fonction de vos objectifs commerciaux en matière d'exploration de données. Deux possibilités s'offrent à vous :

- **Passer à la phase de déploiement.** La phase suivante vous aidera à incorporer les résultats du modèle dans votre processus métier et à produire un rapport final. Même si vos travaux d'exploration de données ont été vains, vous devez utiliser la phase de déploiement de CRISP-DM pour créer un rapport final à remettre au commanditaire du projet.
- **Revenir en arrière, et affiner ou remplacer vos modèles.** Si vous pensez que vos résultats ne sont pas tout à fait optimaux, envisagez de procéder à une nouvelle modélisation. Vous pouvez utiliser ce que vous avez appris dans cette phase pour affiner les modèles et produire de meilleurs résultats.

A ce stade, votre décision implique l'exactitude et la pertinence des résultats de la modélisation. Si les résultats satisfont vos objectifs d'exploration de données et vos objectifs commerciaux, vous êtes prêt à passer à la phase de déploiement. Quelle que soit la décision que vous prenez, fournissez un processus d'évaluation complet.

Exemple d'eCommerce : étapes suivantes

Scénario de Web-Mining à l'aide de CRISP-DM

La société d'eCommerce est relativement confiante quant à l'exactitude et à la pertinence des résultats du projet, et passe par conséquent à la phase de déploiement.

De son côté, l'équipe de projet est également prête à revenir en arrière et à développer certains modèles pour y inclure des techniques prévisionnelles. Ils attendent simplement de recevoir les rapports finaux, ainsi que le feu vert des décideurs.

Chapitre 7. Déploiement

Présentation du déploiement

Le déploiement est le processus consistant à utiliser vos nouvelles connaissances pour apporter des améliorations au sein de l'entreprise. Ceci peut se traduire par une intégration formelle telle que la mise en oeuvre d'un modèle IBM SPSS Modeler produisant des scores d'attrition qui sont ensuite lus dans un entrepôt de données. Le déploiement peut également signifier que vous utilisez les connaissances obtenues suite à l'exploration de données pour provoquer un changement dans votre entreprise. Par exemple, vous avez peut-être découvert des motifs alarmants dans vos données indiquant un changement de comportement des clients âgés de plus de 30 ans. Ces résultats peuvent ne pas être intégrés formellement dans vos systèmes d'informations, mais ils seront sans aucun doute utiles pour la planification et la prise de décisions marketing.

De façon générale, la phase de déploiement de CRISP-DM comprend deux types d'activité :

- Planification et surveillance du déploiement des résultats
- Exécution de tâches de synthèse, telles que la production d'un rapport final et la révision du projet

En fonction des impératifs de votre entreprise, vous pouvez être amené à effectuer l'une de ces étapes, voire les deux.

Planification pour le déploiement

Même si vous êtes impatient de partager le fruit de votre travail d'exploration de données, prenez le temps de planifier les éléments pour obtenir un déploiement régulier et complet des résultats.

Liste des tâches

- La première étape consiste à récapituler vos résultats (modèles et constatations). Vous déterminez ainsi les modèles à intégrer dans vos systèmes de bases de données, ainsi que les constatations à présenter à vos collègues.
- Pour chaque modèle déployable, créez un plan détaillé pour le déploiement et l'intégration dans vos systèmes. Notez tout détail technique tel que les exigences de base de données pour la sortie du modèle. Par exemple, votre système peut exiger que la sortie de modèle soit déployée dans un format délimité par des tabulations.
- Pour chaque constatation probante, créez un plan afin de fournir ces informations aux personnes décidant des stratégies.
- Existe-t-il des plans de déploiement alternatifs à mentionner pour les deux types de résultat ?
- Prenez en considération le contrôle du déploiement. Par exemple, comment un modèle déployé en utilisant IBM SPSS Modeler Solution Publisher sera-t-il mis à jour ? Sur quels critères déciderez-vous qu'un modèle n'est plus applicable ?
- Identifiez tout problème de déploiement et prévoyez des plans de secours. Par exemple, les décideurs peuvent souhaiter obtenir davantage d'informations sur les résultats de la modélisation et plus de détails techniques.

Exemple d'eCommerce : planification du déploiement

Scénario de Web-Mining à l'aide de CRISP-DM

Un déploiement réussi des résultats de l'exploration des données de la société d'eCommerce suppose que les informations correctes soient fournies aux personnes concernées.

© Copyright IBM Corp. 1994, 2015

Décideurs. Les décideurs doivent être informés des recommandations et des suggestions de changements à apporter au site, et recevoir de brèves explications sur l'utilité de ces changements. En supposant qu'ils acceptent les résultats de l'étude, les personnes qui mettront en oeuvre les changements doivent être averties.

Développeurs Web. Les personnes responsables de la maintenance du site Web devront y incorporer les nouvelles recommandations et la nouvelle organisation du contenu du site. Informez-les des changements *éventuels* qui pourraient survenir lors d'études ultérieures afin de leur permettre de préparer le travail dès à présent. Préparer l'équipe en vue de la construction immédiate du site en fonction de l'analyse séquentielle en temps réel peut s'avérer utile pour la suite.

Experts de base de données. Les personnes responsables de la maintenance des bases de données des clients, des achats et des produits doivent être tenues informées de la façon dont les informations provenant des bases de données sont utilisées et des attributs pouvant être ajoutés aux bases de données dans les projets à venir.

Il est primordial que l'équipe de projet soit en contact avec chacun de ces groupes afin de coordonner le déploiement des résultats et la planification des projets à venir.

Planification de la surveillance et de la maintenance

Dans le cadre d'une intégration et d'un déploiement complets des résultats de la modélisation, votre travail d'exploration de données peut se poursuivre. Par exemple, si un modèle est déployé pour prévoir des séquences d'achats réalisés avec un panier d'achat virtuel, il aura probablement besoin d'être évalué périodiquement pour vérifier son efficacité et apporter des améliorations continues. De même, un modèle déployé pour développer la fidélisation de la clientèle parmi les clients les plus importants aura probablement besoin d'être réajusté une fois qu'un niveau particulier de fidélisation aura été atteint. Le modèle peut être modifié et réutilisé pour conserver les clients appartenant à un niveau inférieur mais néanmoins rentable, sur la pyramide des valeurs.

Liste des tâches

Prenez des notes sur les problèmes suivants et incluez-les dans le rapport final.

- Pour chaque modèle ou constatation, quels sont les facteurs ou influences (valeur marchande ou variation saisonnière) à prendre en compte ?
- Comment la validité et l'exactitude de chaque modèle peuvent-elles être mesurées et surveillées ?
- Comment déterminerez-vous le moment où un modèle a « expiré » ? Donnez des détails sur les seuils d'exactitude ou sur les changements attendus dans les données, etc.
- Que se passera-t-il lorsqu'un modèle aura expiré ? Pouvez-vous tout simplement recréer le modèle avec des données plus récentes ou apporter de légères rectifications ? Ou les changements seront-ils assez importants pour nécessiter un nouveau projet d'exploration de données ?
- Ce modèle peut-il être utilisé pour des problèmes métier similaires une fois qu'il aura expiré ? xUne documentation de bonne qualité se montre ici essentielle pour l'évaluation de l'objectif commercial de chaque projet d'exploration de données.

Exemple d'eCommerce : surveillance et maintenance

Scénario de Web-Mining à l'aide de CRISP-DM

La première tâche à effectuer dans le cadre de la surveillance est de déterminer si la nouvelle organisation du site et les recommandations améliorées fonctionnent véritablement. En d'autres termes, les utilisateurs peuvent-ils accéder de manière plus directe aux pages qu'ils recherchent ? Les ventes de produits associés ont-elles augmenté pour les articles recommandés ? Après quelques semaines de surveillance, la société d'eCommerce pourra déterminer la réussite de l'étude.

L'inclusion de nouveaux utilisateurs enregistrés peut être gérée automatiquement. Lorsque des clients s'enregistrent sur le site, les ensembles de règles actuels peuvent être appliqués à leurs informations pour déterminer les recommandations qui doivent être faites à ces clients.

Il est plus difficile de décider du moment où la mise à jour des ensembles de règles doit être effectuée pour déterminer les recommandations. La mise à jour des ensembles de règles n'est pas un processus automatique car la création de cluster requiert une intervention humaine en ce qui concerne la pertinence des solutions de classification.

Etant donné que les futurs projets génèreront des modèles plus complexes, le besoin de surveillance et l'importance qu'elle représente augmenteront de façon presque certaine. Dans la mesure du possible, la surveillance doit être automatique et des rapports planifiés produits régulièrement à des fins de vérification. L'entreprise peut également opter pour la création de modèles fournissant des prévisions immédiates. L'équipe doit alors mener des recherches plus approfondies que lors du premier projet d'exploration de données.

Production d'un rapport final

L'élaboration d'un rapport final permet non seulement de compléter les points manquants de la documentation antérieure, mais également de communiquer les résultats. Même si cette tâche peut paraître simple, il est important de présenter vos résultats aux différentes personnes ayant un intérêt à les connaître. Il peut s'agir non seulement des administrateurs techniques responsables de la mise en oeuvre des résultats de la modélisation, mais aussi des commanditaires (marketing et gestion) qui prendront des décisions en fonction de vos résultats.

Liste des tâches

Commencez par tenir compte des personnes qui liront votre rapport. S'agit-il de développeurs techniques ou de responsables intéressés par le marché ? Vous devrez peut-être créer des rapports distincts en fonction de chaque type de personne si leurs exigences diffèrent. Dans les deux cas, votre rapport doit inclure la majorité des points suivants :

- Une description complète du problème métier initial
- Le processus utilisé pour effectuer l'exploration de données
- Les coûts du projet
- Des remarques sur tout écart par rapport au plan de projet initial
- Un récapitulatif des résultats de l'exploration de données (modèles et constatations)
- Une présentation du plan proposé pour le déploiement
- Des recommandations pour tout travail d'exploration de données ultérieur, incluant des pistes intéressantes issues de l'exploration et de la modélisation

Préparation d'une présentation finale

Outre le rapport du projet, il est possible que vous deviez présenter les constatations du projet à une équipe de commanditaires ou aux services concernés. Si c'est le cas, vous pouvez utiliser une grande partie des informations contenues dans votre rapport en élargissant la perspective. Les graphiques de IBM SPSS Modeler peuvent facilement être exportés pour ce type de présentation.

Exemple d'eCommerce : rapport final

Scénario de Web-Mining à l'aide de CRISP-DM

Le plus grand écart par rapport au plan de projet initial constitue également une piste intéressante pour tout travail d'exploration de données ultérieur. Le plan initial était destiné à rechercher un moyen pour que les clients passent plus de temps sur le site et consultent davantage de pages par visite.

Il se trouve qu'essayer de contenter un client ne signifie pas simplement faire en sorte qu'il reste connecté au site. Les proportions d'effectifs concernant le temps passé par session (distinguant les sessions ayant débouché sur un achat des autres) ont montré que les durées de connexion de la plupart des sessions se concluant par un achat se situent entre les durées de connexion de deux clusters de sessions pendant lesquelles le client n'achète rien.

Le problème est maintenant de savoir si les clients passant beaucoup de temps sur le site sans acheter ne font que naviguer ou s'ils ne parviennent pas à trouver ce qu'ils recherchent. L'étape suivante consiste à savoir comment leur fournir ce qu'ils recherchent de façon à encourager les achats.

Exécution d'une révision de projet finale

Il s'agit de la dernière étape de la méthodologie CRISP-DM. Elle vous offre la possibilité de formuler vos impressions finales et de regrouper les leçons apprises lors du processus d'exploration de données.

Liste des tâches

Vous devez questionner brièvement les personnes impliquées dans le processus d'exploration de données. Questions à prendre en compte lors des entretiens :

- Quelles sont vos impressions générales sur le projet ?
- Qu'avez-vous appris lors du processus (à la fois sur l'exploration de données en général et sur les données disponibles) ?
- Quelles sont les parties du projet qui ont fonctionné correctement ? A quel moment des difficultés ont-elles été rencontrées ? Existait-il des informations qui auraient pu éviter la confusion ?

Une fois les résultats de l'exploration de données déployés, vous pouvez également interroger les personnes concernées par les résultats, telles que les clients ou les partenaires commerciaux. Votre objectif est ici de déterminer si le projet était utile et a offert les avantages escomptés.

Les résultats de ces entretiens, ainsi que vos propres impressions sur le projet, peuvent être récapitulés dans un rapport final ciblé sur les leçons tirées de cette expérience d'exploration des magasins de données.

Exemple d'eCommerce : révision finale

Scénario de Web-Mining à l'aide de CRISP-DM

Entretiens avec les membres du projet. La société d'eCommerce constate que les membres du projet les plus impliqués dans l'étude du début à la fin sont pour la plupart satisfaits des résultats et attendent avec impatience d'autres projets. Le groupe responsable des bases de données est optimiste, mais émet des réserves ; ses membres apprécient l'utilité de l'étude, mais soulignent la charge supplémentaire imposée aux ressources de base de données. Un consultant était disponible durant l'étude mais la nécessité de faire appel à un employé responsable de la maintenance des bases de données se fera sentir au fur et à mesure que le projet avancera et qu'il prendra de l'ampleur.

Entretiens avec les clients. Les réactions des clients ont jusqu'à présent été positives. Cependant, le problème de l'impact du changement de la conception du site sur les clients fidèles n'a pas été pris en compte. Au bout de quelques années, les clients enregistrés développent certaines attentes concernant l'organisation du site. Les réactions des utilisateurs enregistrés ne sont donc pas aussi positives que celles des clients non enregistrés, et certains ont largement critiqué les changements. La société d'eCommerce ne doit pas négliger ce problème. Elle doit chercher à savoir si un changement apportera suffisamment de nouveaux clients pour compenser ceux qu'elle risque de perdre.

Remarques

Ces informations ont été développées pour les produits et services offerts dans le monde.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, programme ou service IBM n'implique pas que seul ce produit, programme ou service IBM puisse être utilisé. Tout produit, programme ou service fonctionnellement équivalent peut être utilisé s'il n'enfreint aucun droit de propriété intellectuelle d'IBM. Cependant l'utilisateur doit évaluer et vérifier l'utilisation d'un produit, programme ou service non IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. L'octroi de ce document n'équivaut aucunement à celui d'une licence pour ces brevets. Vous pouvez envoyer par écrit des questions concernant la licence à :

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Pour le Canada, veuillez adresser votre courrier à :

IBM Director of Commercial Relations
IBM Canada Ltd
3600 Steeles Avenue East
Markham, Ontario
L3R 9Z7
Canada

Pour toute demande au sujet des licences concernant les jeux de caractères codés sur deux octets (DBCS), contactez le service Propriété intellectuelle IBM de votre pays ou adressez vos questions par écrit à :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japon

Le paragraphe suivant ne s'applique ni au Royaume-Uni, ni dans aucun pays dans lequel il serait contraire aux lois locales. 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. Certains états n'autorisent pas l'exclusion de garanties explicites ou implicites lors de certaines transactions, par conséquent, il est possible que cet énoncé ne vous concerne pas.

Ces informations peuvent contenir des erreurs techniques ou des erreurs typographiques. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Toute référence dans ces informations à des sites Web autres qu'IBM est fournie dans un but pratique uniquement et ne sert en aucun cas de recommandation pour ces sites Web. 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 à votre égard, 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 Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Ces informations peuvent être disponibles, soumises à des conditions générales, et dans certains cas payantes.

Le programme sous licence décrit dans le présent document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions internationales d'utilisation des Logiciels IBM ou de tout autre contrat équivalent.

Toutes les données sur les performances contenues dans le présent document ont été obtenues dans un environnement contrôlé. Par conséquent, les résultats obtenus dans d'autres environnements d'exploitation peuvent varier de manière significative. Certaines mesures peuvent avoir été effectuées sur des systèmes en cours de développement et il est impossible de garantir que ces mesures seront les mêmes sur les systèmes commercialisés. De plus, certaines mesures peuvent avoir été estimées par extrapolation. Les résultats réels peuvent être différents. Les utilisateurs de ce document doivent vérifier les données applicables à leur environnement spécifique.

les informations concernant les produits autres qu'IBM ont été obtenues auprès des fabricants de ces produits, leurs annonces publiques ou d'autres sources publiques disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Aucune réclamation relative à des produits non IBM ne pourra être reçue par IBM. Les questions sur les capacités de produits autres qu'IBM doivent être adressées aux fabricants de ces produits.

Toutes les déclarations concernant la direction ou les intentions futures d'IBM peuvent être modifiées ou retirées sans avertissement préalable et représentent uniquement des buts et des objectifs.

Ces informations contiennent des exemples de données et de rapports utilisés au cours d'opérations quotidiennes standard. Pour les illustrer le mieux possible, ces exemples contiennent des noms d'individus, d'entreprises, de marques et de produits. Tous ces noms sont fictifs et toute ressemblance avec des noms et des adresses utilisés par une entreprise réelle ne serait que pure coïncidence.

Si vous consultez la version papier de ces informations, il est possible que certaines photographies et illustrations en couleurs n'apparaissent pas.

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines dans de nombreux pays. Les autres noms de produits et de services 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 l'adresse www.ibm.com/legal/copytrade.shtml.

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 commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux Etats-Unis et dans d'autres pays.

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

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

UNIX est une marque déposée de The Open Group aux Etats-Unis et dans d'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.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

Index

A

agrégation 22
aide
  CRISP-DM 2
ajout de données 22
algorithmes 26
analyse coût-bénéfice 9
apprentissage/test 27
arrière-plan
  collecte d'informations 6
attributs
  calcul 21
  sélection 19

B

blancs
  collecte de données 13
  vérification de la qualité des données 16

C

compréhension
  besoins métier 5
  données 13
  objectifs de l'exploration de données 9
compréhension de l'entreprise 5
compréhension des données 13
conclusions 31
constatations 31
construction de données 21
contraintes
  établissement d'une liste 8
CRISP-DM
  aide 2
  présentation 1
  ressources supplémentaires 2
  utilisation dans IBM SPSS Modeler 1
critères
  réussite commerciale 7
  réussite de l'exploration de données 10
critères de réussite
  point de vue commercial 7
  point de vue de l'exploration de données 9
  point de vue technique 10

D

définition
  terminologie du projet 9
délimiteurs 17
déploiement 35
données
  attributs 13
  collecte 13
données *(suite)*
  construction de nouvelles données 21
  description 14
  examen de la qualité 16
  exclusion 19
  exploration 15
  fichiers non hiérarchiques 17
  format 15
  formatage en vue de la modélisation 23
  fusion 22
  intégration 22
  nettoyage 20
  partition 27
  rapport sur la collecte 14
  rapport sur la qualité 17
  sélection 19
  sélection d'attributs 19
  statistiques relatives à la taille 14
  tri 23
  types 13
  valeurs manquantes 16
  vérification de la qualité 16
  visualisation 15

E

élaboration
  plan du projet 11
  rapport sur l'exploration des données 16
  rapport sur la collecte de données 14, 15
  rapport sur la qualité des données 17
  rapport sur le nettoyage des données 21
enregistrements
  génération 21
  sélection 19
erreurs 20
évaluation
  détermination des étapes suivantes 33
  modèles 29
  outils disponibles 11
  phase de CRISP-DM 31
  situation commerciale actuelle 7
exemples
  eCommerce 22
  phase d'évaluation 31, 32, 33
  phase de compréhension de l'entreprise 5, 7, 10, 11
  phase de compréhension des données 13, 14, 15, 17
  phase de modélisation 25, 27, 28, 29
  phase de préparation des données 19, 20, 21, 22
exploration de données
  détermination des étapes suivantes 33
  révision du processus 32
  utilisation de CRISP-DM 1

F

fichiers non hiérarchiques 17
fusion des données 13, 22

H

HTML
  génération de rapports 2
hypothèse
  formulation 16

I

impératifs
  établissement d'une liste 8
info-bulles 2

M

maintenance 36
métadonnées 16, 20
modèle
  évaluation des résultats 31
modèles
  création 27
  liste des modèles approuvés 31
  non supervisés 26
  paramètres 28
  supervisés 26
  types 28
modèles approuvés 31
modèles non supervisés 26
modèles supervisés 26
modélisation 25
  définition des options 27
  évaluation de la sortie 29
  impératifs des données 23
  préparation des données 19
  techniques 25, 26
  test des résultats 26

N

nettoyage des données 20
noeud Ajouter 22
noeud Binariser 21
noeud Calculer 21
noeud Fusionner 22
normalisation 21

O

objectifs
  ajustement 16
  définition d'objectifs commerciaux 5
  définition des objectifs commerciaux 5
  définition des objectifs de l'exploration de données 9
  tâches impliquées 6
options
  modélisation 28
organigrammes 6
outil de projet 2
outils
  évaluation 11
outils de visualisation 15
ouvrages
  CRISP-DM 2

P

paramètres
  modélisation 28, 30
parasite 17, 20
partition 27
phase
  compréhension de l'entreprise 5
  compréhension des données 13
  évaluation 31
  modélisation 25
  préparation des données 19
planification
  déploiement des résultats 35
  élaboration du plan du projet 11
  surveillance et maintenance 36
préparation des données 19
présentation des résultats 37
processus
  révision de l'exploration de données 32
projets
  élaboration du rapport final 37
  exécution d'une révision finale 38
  inventaire des ressources 8
  liste des impératifs, des hypothèses et des contraintes 8
  liste des risques et des plans de secours 9
  réalisation d'une analyse coût-bénéfice 9

Q

qualité
  examen des données 16
  rapport sur la qualité des données 17
qualités d'ajustement 26

R

rapports
  collecte de données 14
  description des données 15
  exploration des données 16
rapports *(suite)*
  génération à partir de l'outil de projet 2
  nettoyage des données 21
  plan du projet 11
  projet final 37
  qualité des données 17
ressources
  inventaire des ressources du projet 8
  ressources supplémentaires sur CRISP-DM 2
résultats
  évaluation 31
  présentation 37
réussite commerciale
  évaluation des résultats 31
révision
  processus d'exploration de données 32
risques 9

S

sélection de données 19
statistiques
  exploratoire 16
statistiques exploratoires 16
surveillance du déploiement 36

T

taille
  jeux de données 14
techniques
  modélisation 26
terminologie 9
tri 23

V

valeurs booléennes 14
valeurs manquantes 13, 16, 20, 21
valeurs numériques 14
valeurs symboliques 14

W

Web-mining
  eCommerce 5, 7, 10, 19, 20, 21, 22, 25, 27, 28, 29, 31, 32, 33