Analyse et application de modèles de régression pour optimiser le retour sur investissement d'opérations commerciales

Thomas Piton*,**, Julien Blanchard**, Henri Briand**, Laurent Tessier***, Gaëtan Blain*,

* Groupe VM Matériaux, Route de la Roche sur Yon, 85 260 L'Herbergement
{tpiton, gblain}@vm-materiaux.fr, http://www.vm-materiaux.fr/
** LINA équipe COD - UMR 6241 CNRS, 2 rue de la Houssinière, 44322 Nantes
{julien.blanchard, henri.briand}@univ-nantes.fr, http://www.polytech.univ-nantes.fr/COD
*** KXEN, 25 quai Galliéni, 92158 Suresnes
laurent.tessier@kxen.com, http://www.kxen.com/

Résumé. Les activités de négoce de matériaux sont un marché extrêmement compétitif. Pour les acteurs de ce marché, les méthodes de fouille de données peuvent s'avérer intéressantes en permettant de dégager des gains de rentabilité importants. Dans cet article, nous présenterons le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement d'opérations commerciales. La synergie des informaticiens, du marketing et des experts métier a permis d'améliorer l'extraction des connaissances à partir des données de manière à aboutir à la connaissance actionnable la plus pertinente possible et ainsi aider les experts métier à prendre des décisions.

1 Introduction

À l'aube de la société de l'information, la maîtrise des données dans l'entreprise devient un enjeu majeur dans la compétition pour acquérir et conserver des parts de marché. Maîtriser l'information pour bien décider, c'est avoir les bonnes données, exploitées par de bons outils, au bon moment (Tufféry, 2005). Au premier rang des technologies actuelles de l'information, la fouille de données offre une réelle possibilité d'exploiter finement et rapidement les données afin de permettre aux utilisateurs de mieux orienter leurs actions. Afin que la communication puisse s'appuyer sur les modèles de fouille de données, une telle approche nécessite d'accorder un soin tout particulier à la qualité des modèles produits, par exemple par des méthodes d'évaluation intelligibles et des techniques de représentation (Guillet et Hamilton, 2007) (Briand et al., 2004).

VM Matériaux, entreprise de Négoce de matériaux, de menuiserie industrielle et de béton prêt à l'emploi réalise de nombreuses opérations commerciales, ciblant principalement ses clients professionnels. Pour une grande partie des campagnes, une invitation à participer est envoyée à chaque client « routé ». Le routage est réalisé manuellement par l'équipe marketing quelques semaines avant l'opération et se base principalement sur les clients ayant réalisé

un certain seuil de chiffre d'affaire (CA) l'année précédente. Ces opérations commerciales maîtrisées depuis une dizaine d'années mettent en jeu des dépenses et des recettes importantes.

Dès lors, le retour sur investissement (*Return On Investment* ou ROI) des opérations commerciales de VM Matériaux peut être amélioré par des techniques de fouille de données. La connaissance extraite des différents modèles doit permettre aux experts de comprendre le comportement de leurs clients et ainsi prendre des décisions en utilisant le savoir extrait à bon escient (paradigme de l'actionable knowledge (Cao, 2007; Graco et al., 2007)). Dans cet article, nous présentons le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le ROI des opérations commerciales. Nous développons plus particulièrement l'évaluation et la mise en œuvre des modèles de régression *ridge* (Dodge, 2004) pour perfectionner le routage d'une campagne marketing. Ces modèles ont été construits avec le logiciel KXEN qui se fonde sur la théorie de l'apprentissage statistique (Vapnik, 1998).

En premier lieu, nous présentons les mesures et les processus d'évaluation pour les experts métier dans le cadre de la régression *ridge*. Ensuite, nous présentons l'application réalisée avec KXEN pour améliorer le ROI d'une opération commerciale, et expliquons l'apport des modèles pour les experts du métier.

2 Évaluation des modèles

Cette partie présente les mesures et processus d'évaluation intelligibles pour les experts métier dans le cadre de la régression *ridge*. Cette dernière a l'avantage de pénaliser les paramètres lorsque la variable est fortement bruitée et a le privilège d'être peu sensible aux corrélations (Dodge, 2004). Afin d'illustrer les sections ci-dessous, nous considérons que la variable cible binaire désigne la participation à une opération commerciale, c'est-à-dire à l'achat d'un produit (1 pour acheter, 0 sinon). La finalité de la régression est donc de prévoir les acheteurs d'une opération commerciale dans une population de clients (Berry et Gordon, 1997). Étant donné un seuil s, on prédit qu'un client est un acheteur si le score calculé par le modèle est supérieur à s. On note $u(s)$, la proportion de clients x dont le score calculé par le modèle est supérieur à s :

$$u(s) = P(score(x) \geqslant s)$$

On note $v(s)$ la proportion d'acheteurs réels détectés par le modèle :

$$v(s) = P(score(x) \geqslant s \mid x = acheteur)$$

2.1 Précision et robustesse

Afin que les experts métier puissent visualiser la précision et la robustesse de nos modèles, nous utilisons des courbes lift. Une courbe lift (variante de la courbe ROC) est une courbe paramétrique qui représente la proportion d'acheteurs détectés $v(s)$ en fonction de la proportion de clients sélectionnés $u(s)$ (Tufféry, 2005). Elle est construite en triant les clients par ordre de score décroissant. Par ailleurs, il peut être profitable à l'expert métier de visualiser la « représentation du lift » présentant cette fois-ci le taux d'augmentation du lift en ordonnée, la courbe étant par conséquent décroissante.

La précision et la robustesse d'un modèle peuvent être mesurées en comparant la courbe lift à une courbe aléatoire et une courbe idéale (cf. figure 1). La courbe aléatoire est la courbe

$y = x$ (on détecte α % des acheteurs en sélectionnant α % des clients). La courbe idéale est celle dans laquelle tous les acheteurs sont sélectionnés en premier.

À partir de la courbe lift, deux indicateurs peuvent être calculés (cf. figure 1). Le premier indicateur est l'indice de GINI (Tufféry, 2005), nommé KI dans KXEN. Il correspond à l'aire entre la courbe de *lift* et la courbe aléatoire. Le KI de validation est égal au rapport des aires $C/(A+B+C)$ et le KI d'apprentissage est égale à $(B+C)/(A+B+C)$. Il mesure la précision du modèle, c'est-à-dire la capacité des variables d'entrée à expliquer la cible. L'indicateur compris entre 0 (modèle purement aléatoire) et 1 (modèle parfait) permet de classer les modèles en fonction de leur pouvoir explicatif face à la variable à expliquer. KI est lié à l'aire sous la courbe ROC par la formule suivante : $KI = 2AUC - 1$. Le deuxième indicateur, nommé KR dans l'outil KXEN, correspond à la différence d'aire entre les deux courbes de *lift*, soit $(1 - B)/(A + B + C)$. Il mesure la robustesse du modèle, c'est-à-dire sa capacité à fournir le même niveau de qualité sur un nouveau jeu de données, typiquement le jeu de données de validation. Il est également compris entre 0 et 1 et il est préférable qu'il soit supérieur à 0,95 pour que le modèle soit robuste. Par exemple, sur la figure 1, le point M montre que sur l'ensemble d'apprentissage, en ciblant 50 % des clients (les meilleurs selon le modèle), on détecte 80 % des acheteurs.

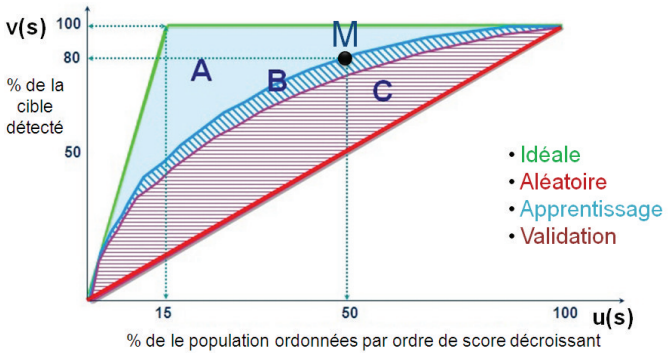

FIG. 1 – *Courbes de lift.*

2.2 Retour sur investissement

De manière à ce que les experts métier puissent estimer le retour sur investissement engendré par un modèle sur une opération commerciale, nous utilisons des courbes de profit sur l'ensemble d'apprentissage ou de validation. Une courbe de profit est la transformation d'une courbe de *lift* à l'aide d'une matrice des coûts définie par l'utilisateur. Le profit peut être défini de la manière suivante : il s'agit de la marge réalisée en contactant une proportion $u(s)$ de clients, c'est-à-dire

$$profit(s) = N*[P(x = acheteur|score(x) \geqslant s)*C - P(x = non\,acheteur|score(x) \geqslant s)*D]$$

avec N le nombre de clients dans l'échantillon étudié, C la marge moyenne nette par acheteur et D la dépense moyenne de communication par client. Le profit maximal théorique,

$profitMAX$ est le modèle où tous les acheteurs sont sélectionnés en premier. Ainsi, une courbe de profit est une courbe paramétrique qui représente le taux de profit ($profit(s)/profitMAX$) en fonction de la proportion de clients sélectionnés $u(s)$. Cette courbe présente une ordonnée différente de la courbe de *lift* avec non plus le pourcentage d'acheteurs détectés mais le pourcentage de ROI maximal de manière à mesurer graphiquement le retour financier de l'opération commerciale.

3 Application : optimisation des opérations commerciales

Nous nous concentrons ci-dessous sur le *scoring* d'une opération commerciale de VM Matériaux à travers un modèle de régression *ridge*.

3.1 Contexte

L'activité Négoce du groupe VM Matériaux organise deux journées commerciales destinées à promouvoir l'ensemble des produits. Cette opération est réservée aux professionnels du bâtiment. Une opération promotionnelle liée aux achats HT permet d'obtenir différents cadeaux ou gains de points VM sous réserve de la réalisation d'un chiffre d'affaires HT passé en commande pendant ces deux jours et facturé en fin de mois. Actuellement, tous les clients professionnels ayant réalisé un CA supérieur à un certain seuil HT sur l'année 2007 sont démarchés par courrier et par leurs commerciaux respectifs.

3.2 Modélisation

À l'aide de l'entrepôt de données existant, créons notre modèle de données basé sur les clients routés l'année précédente. Ajoutons leurs les caractéristiques de la table des clients. Ensuite, enrichissons le modèle avec le résultat d'une opération commerciale similaire mais printanière. Par la suite, créons des agrégats temporels basés sur le chiffre d'affaire, la marge nette et le nombre de lignes de commandes sur six périodes de un mois. Enfin, ajoutons une cible binaire relative à la détection des acheteurs (égale à 1 si le client a acheté, à 0 sinon). Cette phase de pré-traitement des données génère un modèle de 66 variables et de 10 378 lignes.

Les indicateurs collectés illustrent la robustesse (KR à 0,991) et la précision du modèle à expliquer la cible (KI à 0,815). De plus, 25,48 % de la cible possède la valeur 1, signifiant que 25,48 % des routés ont acheté lors de l'opération commerciale 2007. Ensuite, 50 des 66 variables ont été incluses dans le modèle en réalisant une sélection des variables préservant la précision et la robustesse du modèle et facilitant l'interprétation des experts métier.

3.3 Interprétation

Le modèle de régression *ridge* a été créé dans la partie précédente. À présent, créons un jeu d'application de même structure mais composé des clients professionnels ayant réalisés du chiffre d'affaire sur les douze derniers mois : 16 365 observations répondent à nos exigences. Visualisons à présent la contribution des variables (cf. figure 2) permettant d'indiquer les variables contribuant à l'achat durant les journées de l'opération commerciale.

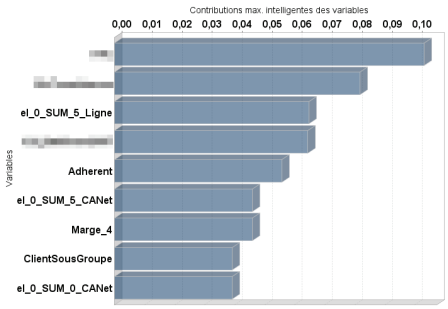

FIG. 2 – *Contribution maximale intelligente des variables.*

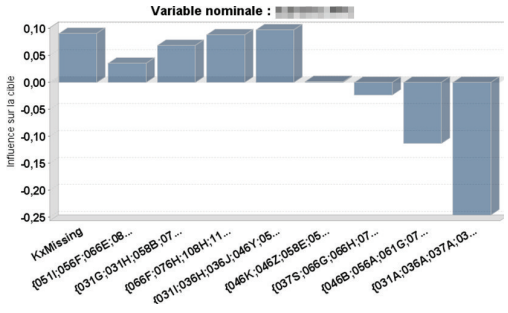

FIG. 3 – *Contribution variable nominale.*

La visualisation des catégories de chacune des variables nous amène à des interprétations pour les experts du métier. La figure 3 illustre que certains intervenants contribuent fortement à la participation à l'opération commerciale. Suite à l'extraction de cette connaissance, la table décisionnelle correspondante a été enrichi pour étudier le comportement des intervenants lors des campagnes marketing. En analysant les fréquences des catégories, nous avons remarqué que 34,1 % des intervenants contribuent de manière significative à l'opération.

3.4 Liste de routage

L'application du jeu de données précédent a permis de générer un score et une probabilité pour chaque client, soulignant le potentiel que chacun achète ou non durant l'opération commerciale. L'ensemble des clients a été trié par probabilité de participation.

Ensuite, nous avons renseigné la matrice des coûts pour générer une courbe de profit personnalisée, ayant calculé préalablement par client la marge nette moyenne s'il achète durant l'opération et le coût de la démarche. La courbe de profit admet une tangente horizontale au point d'abscisse 54,6 %. Par conséquent, les 54,6 % des premiers clients de notre liste permettent d'optimiser la différence entre les dépenses et les recettes de l'opération commerciale. Pour ne pas altérer le résultat de l'équipe marketing, les experts métier ont décidé de concaténer à la liste marketing initiale les clients non routés parmi les acheteurs les plus probables. Quantitativement, l'ajout a été de 9,65 % clients. De cette manière, 12 170 clients ont été ré-appliqués au modèle pour être triés puis routés. Les listes définitives de routage ont été préparées par agence et par commercial. Chaque client n'ayant pas participé l'année dernière et ne figurant pas dans la liste marketing a été coloré. De cette manière, l'étude a permis d'aboutir à un pré-mâchage automatisable du travail des commerciaux sur le terrain.

3.5 Mesure du ROI

Lors de la dernière opération, le taux d'acheteurs à la campagne était de 25,48 %. La probabilité calculée pour chaque client routé est une vraie probabilité de participer à la campagne marketing. De ce fait, la somme des probabilités est un estimateur du nombre de répondants.

Nous estimons ainsi que nous devrions augmenter le nombre de participants d'environ 10 %, engendrant ainsi une augmentation de 11 % du chiffre d'affaire de l'opération commerciale.

4 Conclusion

Nous avons présenté dans cet article le retour d'expérience du projet de fouille de données mené chez VM Matériaux pour améliorer le retour sur investissement des opérations commerciales. L'application des modèles de régression *ridge* construits avec l'outil KXEN a permis de valoriser la richesse de l'entrepôt de données de VM Matériaux pour prévoir le comportement de ses clients professionnels. En évaluant la qualité des modèles à l'aide d'indicateurs intelligibles et de représentations graphiques, nous avons pu obtenir le soutien des intervenants sur le terrain et un retour sur investissement mesurable pour les experts métier. Ainsi, une intégration de la connaissance métier dans le processus d'extraction permettrait d'améliorer la justesse et la pertinence des modèles, et par conséquent d'interagir en meilleure adéquation avec les experts métier.

Références

Berry, J. et S. Gordon (1997). *Data Mining : techniques appliquées au marketing, à la vente et aux services clients*. Masson.

Briand, H., M. Sebag, R. Gras, et F. Guillet (2004). *Mesures de qualité pour la fouille de données*. Cepadues.

Cao, L. (2007). Domain-driven, actionable knowledge discovery. *IEEE Intelligent Systems 22(4)*, 78–88.

Dodge, Y. (2004). *Analyse de régression appliquée*. Dunod.

Graco, W., T. Semenova, et E. Dussobarsky (2007). Toward knowledge-driven data mining. *ACM SIGKDD Workshop on Domain Driven Data Mining*, 49–54.

Guillet, F. et H. Hamilton (2007). *Quality Measures in Data Mining*. Springer.

Tufféry, S. (2005). *Data Mining et statistique décisionnelle, l'intelligence dans les bases de données*. Technip.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Summary

Trading activities of materials are an extremely competitive market. Data mining methods may be interesting to generate substantial profits for business people. In this paper, we propose a feedback on a data-mining project carried out at the VM Matériaux company to improve the return on investment of marketing campaigns and commercial operations. The synergy of computer sciences, marketing experts and business people has improved extracting knowledge in order to achieve actionable knowledge discovery as useful as possible and help retails experts to make business decisions.