Fondation ubs: Université Bretagne Sud

CHAIRE D-CC "DÉCISIONNEL-CONNAISSANCE CLIENT" - IUT de Vannes - 8 rue Montaigne - 56017 Vannes Cx / 02 97 62 64 64 / chaire.d-cc@univ-ubs.fr

Predicsis Model Producer, la productivité au service des scores comportementaux

Matthieu GOUSSEFF

February 3, 2017

Contents

Abstract

Le logiciel Model Producer propose un compromis entre la rapidité de traitement, la simplification du data management et la qualité prédictive de scores de comportement. La scalabilité permise par le moyennage optimisé de prédicteurs bayésiens naïfs produit des scores de bonne qualité dans des délais raisonnables. La construction d'agrégats à partir de tables contenant plusieurs mesures d'une variable par individu est une fonctionnalité susceptible d'améliorer la productivité des data-miners lors de la production de scores.

1 Présentation du Model Producer de Predicsis

1.1 Productions de scores comportementaux

La relation entre l'entreprise et le client repose en grande partie sur la capacité de proposer le bon produit au bon client au bon moment. Un des enjeux quotidiens des data-miners est donc d'estimer la probabilité qu'a chaque client d'une entreprise de répondre positivement à une sollicitation.

Le Model Producer (MP) est l'implémentation commerciale d'une méthode originale de combinaisons de classifieurs bayésiens naïfs.

1.2 L'entreprise

Predicsis a été fondée en 2013 afin de valoriser la technologie Khiops développée au sein d'Orange Labs sur une durée de 15 ans. Soutenue par le fonds d'investissement Innovacom VC, Predicsis a vu ses revenus tripler entre 2014 et 2015. Elle compte aujourd'hui 25 salariés et revendique une approche disruptive dans la production de scores comportementaux.

1.3 Model Producer

Les classifieurs bayésiens naïfs sont une méthode classique de classification supervisée. L'approche proposée par Predicsis est de combiner des classifieurs bayésiens naïfs en optimisant un critère original, le taux de compression. Les atouts revendiqués par Predicsis sont :

- Une simplification du data management par
  - la discrétisation optimisée des variables prédictrices continues
  - la création d'indicateurs (agrégats) pour les variables mesurées plusieurs fois
- Une scalabilité permettant de traiter à la fois un grand nombre d'individus et un grand nombre de variables prédictrices
- Un score de très bonne qualité au vu des objectifs métiers visés

L'outil Model Producer de Predicsis repose sur une **combinaison de classifieurs bayésiens naïfs** (bagging) qui permet de trouver **un compromis entre la scalabilité (temps de calcul, mémoire nécessaire), la facilité de construction d'un modèle et la qualité de prédiction**. Les mesures répétées sur un individu peuvent être prises en compte à partir de tables périphériques sans data management supplémentaire.

2 Test Utilisateur

Le test présenté a été effectué avec des versions allant jusqu'à la version 3.2. Les échanges avec les équipes de développeurs ont été pris en compte au fil des évolutions de version. Nous n'avons pas encore reconduit notre test avec la version actuelle, soit la version 3.4, qui est réputée encore plus stable.

2.1 Installation

L'outil est encapsulé dans une machine virtuelle vagrant/virtualbox. L'installation se fait à à partir de fichiers fournis par Predicsis et qui permettent le téléchargement sécurisé de la machine virtuelle, ainsi que l'activation de la licence.

Predicsis annonce que la prochaine version prévue (3.6) sera disponible sur la Market Place d'Amazon, et qu'elle serait provisionnable en un clic.

Avantage

- L'installation est aisée si l'on dispose des droits nécessaires et d'une connexion suffisante.
- Le lancement et la fermeture du Model Producer se font par une simple commande vagrant. Un système de clé permet le téléchargement sécurisé du contenu de la machine virtuelle.
- Le Model Producer (MP) est acccessible depuis tous les postes qui ont accès au serveur qui l'héberge, par le biais d'un simple navigateur web.
- Si le port 8080 est indisponible, MP propose d'autre ports qui sont affichés dans le terminal à partir duquel est fait le lancement.

Inconvénients

- Des problèmes de stabilité ont été rencontrés. Nécessité de fermer et relancer la machine virtuelle.
- Il faut être vigilant sur les paramètres systèmes relatifs à virtual box et s'assurer que le répertoire par défaut contient l'espace suffisant à la duplication des données.

- La mise en place de la machine virtuelle demande un certain temps. Ainsi, il est parfois nécessaire de rafraîchir plusieurs fois avant d'accéder au MP (à vérifier sur la V3.4).
- Pas de message d'erreur en cas de fermeture inopinée du Model Producer, nécessité de consulter les logs (via l'interface, donc peu simple si la machine ne se lance pas).

Conclusion sur l'installation L'installation par machine vagrant est en principe simple et sécurisée, et elle sera encore plus simple dès la version 3.6 provisionnable en un clic sur la Market Place d'Amazon. L'accès au Model Producer par un navigateur est un avantage pour un usage multi-utilisateurs. L'interrogation du modèle par une API à l'aide d'un SDK n'a pas pu être testée.

2.2 Préparation des données

Les données, tables centrales et tables périphériques pour les données répétées, doivent être déposées dans un répertoire créé au sein du répertoire mp-data d'où est lancé le Model Producer (MP). Le nom du répertoire sera le nom du projet associé. Une nomenclature est également exigée pour le nommage des variables, bien que les dernières versions aient apporté un peu plus de souplesse.

MP ne propose pas d'échantillonner pour l'utilisateur un sous-échantillon d'apprentissage et un sous-échantillon de test.

Cette structuration contraignante peut cependant être aisément automatisée dans la phase d'extraction des données.

A noter que le séparateur de champs est spécifié au niveau de l'outil et non au niveau du projet.

Enfin, un répertoire tobedeployed doit être créé pour le jeu de données sur lequel l'équation de prédiction sera appliquée.

Avantages

- Une convention de nommage explicite.
- Toutes les données traitées par le MP sont dans le même répertoire.
- Une plus grande souplesse depuis les dernières versions.
- La possibilité d'utiliser des tables périphériques.

Inconvénients

- Pas d'import convivial de données.
- Création des sous-échantillons test et apprentissage en dehors de l'outil.
- Ergonomie du choix du séparateur, des variables d'identification et cible au niveau outil et non au niveau projet.
- Les données doivent obligatoirement être encodées en UTF-8.

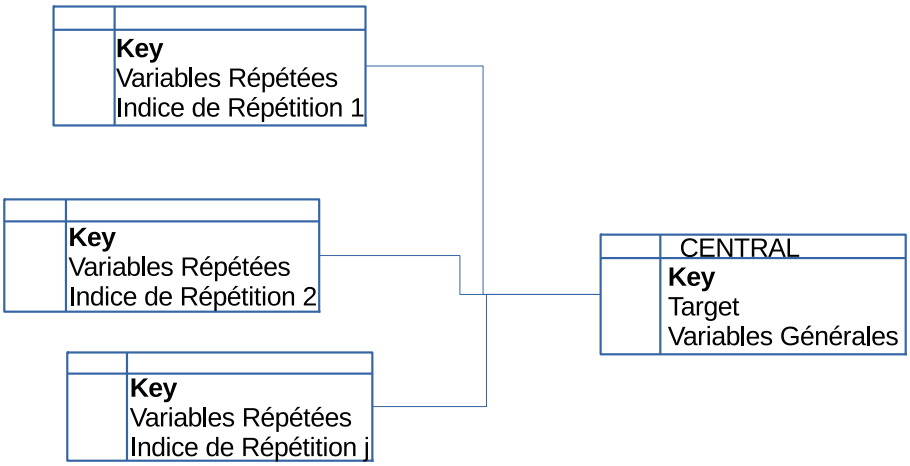

Figure 1: Structure en étoile pour la construction automatique d'agrégats

Conclusion sur la préparation des données Pour le moment, l'outil exige une certaine rigueur dans le stockage et le nommage des jeux de données à scorer, ce qui peut sembler contraignant, mais favorise des bonnes pratiques de data management. Predicsis prévoit d'intégrer à une version 3.6 une gestion plus fluide des imports et exports de fichiers dans l'outil.

2.3 Production d'un score

La page d'accueil de MP présente autant de tuiles qu'il y a de répertoires contenant des données à scorer dans le répertoire mp-data (voir la figure 2, page 7). On accède à la construction de modèles en cliquant sur la tuile correspondant aux données qu'on veut traiter.

2.3.1 Premier score

Lors de la création du premier score, MP effectue un certain de nombre de tâches préliminaires. Ces tâches ne sont effectuées qu'une fois, mais ce sont elles qui demandent le plus de temps de calcul.

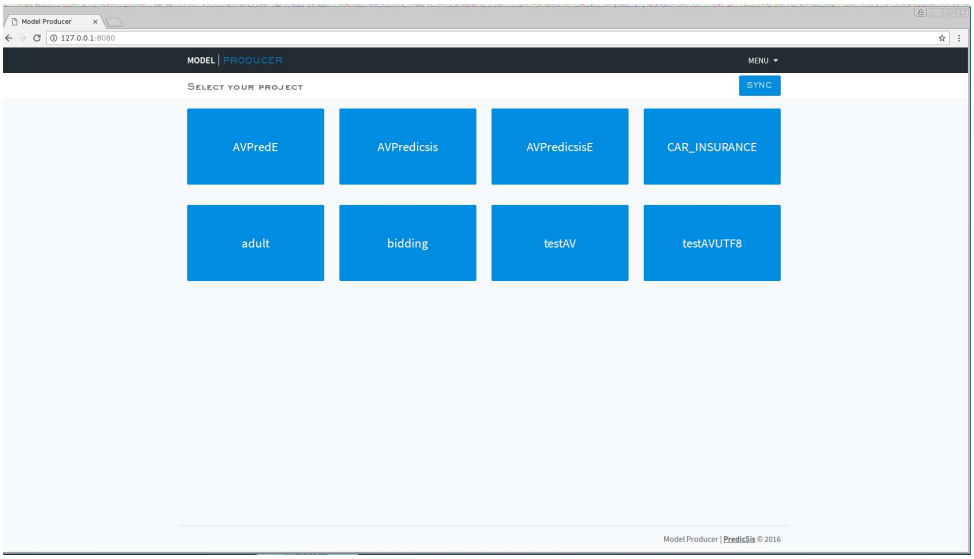

Figure 2: Page d'accueil du Model Producer

- Analyse des noms de fichier
- Recherche du séparateur
- Identification des variables d'identification et à prédire
- Tri et discrétisation des données

L'étape de tri/discrétisation devient longue pour les très gros jeux de données, mais elle n'est effectuée qu'une seule fois et n'est plus nécessaire pour les raffinements ultérieurs de modèles.

Le premier score est ensuite produit. Par défaut il ne prend pas en compte les tables périphériques. Ce premier modèle intègre la liste des modèles disponibles. Il est ensuite possible de raffiner le modèle, de produire un rapport, ou de déployer le score sur un autre jeu de données.

2.4 Rapport d'analyse

Le rapport d'analyse d'un score est présenté à l'utilisateur comme une page web interactive, qui peut ensuite être exportée et distribuée au format pdf. On peut voir un exemple de la page principale sur la figure 3, page 8.

L'indicateur synthétique par défaut est l'indicateur de Gini (aire sous la courbe ROC). Il permet de comparer la qualité des différents modèles créés.

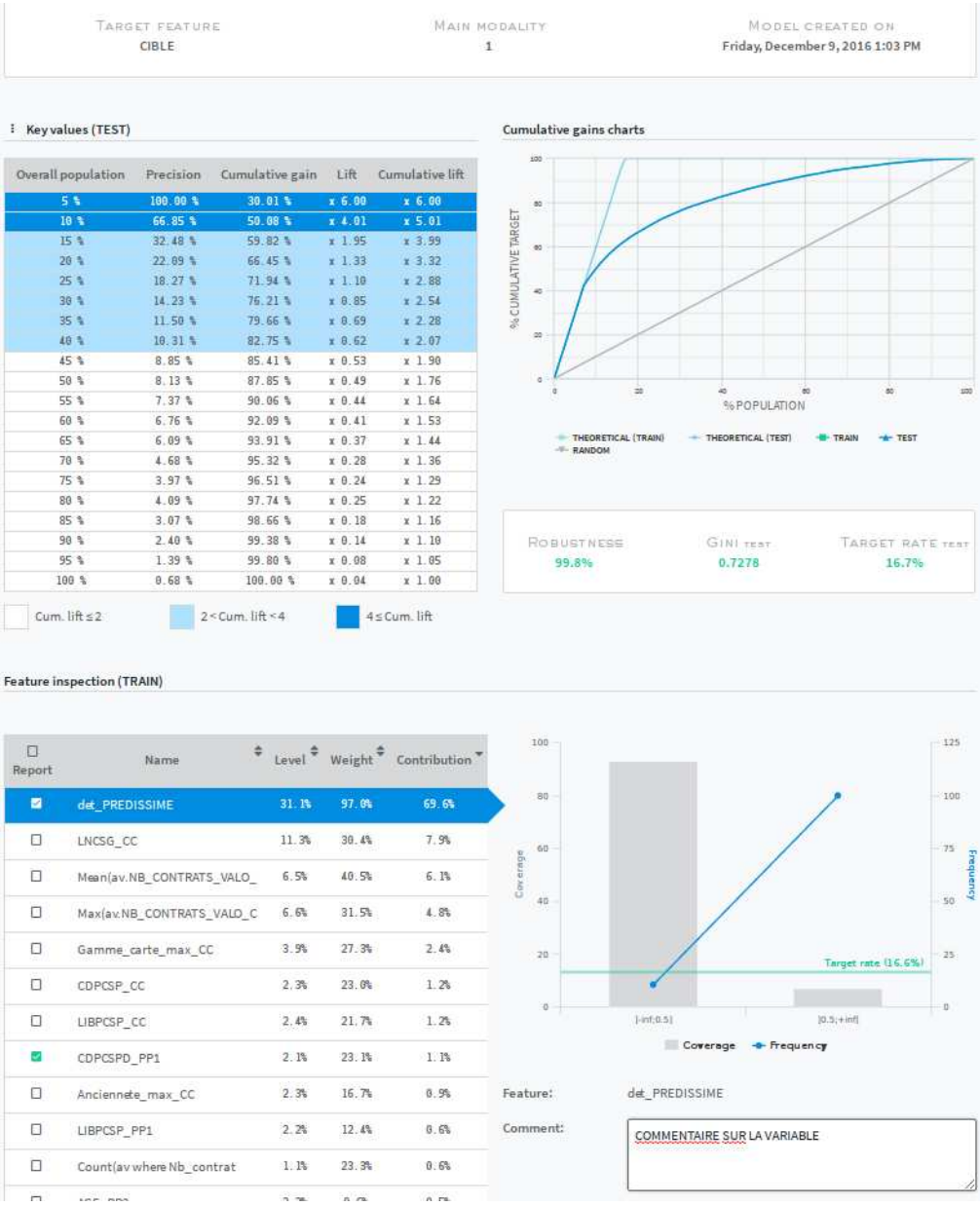

Figure 3: Exemple de rapport produit par Model Producer.

Le gain et et le levier sont affichés par tranche de 5% de l'échantillon. Les tranches pour lesquelles le levier cumulé est supérieur à 2 sont indiquées par un surlignage de couleur. Cette représentation aide à la décision quant à la taille de la cible à retenir, en fonction des objectifs de concentration qu'on cherche à atteindre ou de ressources disponibles.

La liste des prédictrices retenues est affichée. Cliquer sur une variable permet d'accéder à un graphique représentant l'évolution du pourcentage de répondants par modalité, ainsi que la fréquence de chaque modalité.

Il est possible de choisir les prédictrices qu'on veut voir apparaître dans le rapport exporté en pdf, et on peut associer à chacune un commentaire.

La production de ce rapport est rapide et aisée. Elle est un avantage certain de l'outil et permettra de communiquer avec les experts métiers sur la base de documents intelligibles et appropriables.

2.4.1 Raffinage de modèle

La force de l'outil est indéniablement la facilité avec laquelle il est possible de raffiner le modèle initial. L'utilisateur peut de manière très intuitive forcer la prise en compte ou le rejet de variables prédictrices, fixer un nombre maximal de variables et un nombre maximal de variables issues des tables périphériques,**Ce qui est primordial pour l'appropriation du modèle**.

Ces raffinements sont pris en compte très rapidement pour la construction d'un nouveau modèle, car ils ont lieu après le tri et la discrétisation des variables. Le travail de construction d'un modèle prédictif est donc rapide et intuitif **pour qui connaît bien les variables d'entrée.**

2.5 Un exemple concret : score d'appétence assurance vie

Nous avons testé le Model Producer sur un jeu de données ayant servi à la construction d'un score d'appétence à un produit d'assurance-vie.

2.6 Déroulement de la production du score

Les données ont été mises en forme et exportées dans le répertoire mp-data. Les premières versions du Model Producer ont exhibé des erreurs internes qui ont été corrigées par la suite. **Le support technique a été réactif et efficace dans la correction des ces bugs.**

Le jeu de données comportait 120 000 lignes et 30 variables dans la table centrale. La table périphérique contenait 42 variables mesurées à 4 périodes différentes sur les mêmes 120 000 individus.

Ce jeu de données a ensuite été scindé en un échantillon d'apprentissage et un échantillon de test (tirage aléatoire sans remise). Il a été nécessaire de ré-encoder le jeu de données en UTF-8 pour que le MP accepte de le traiter.

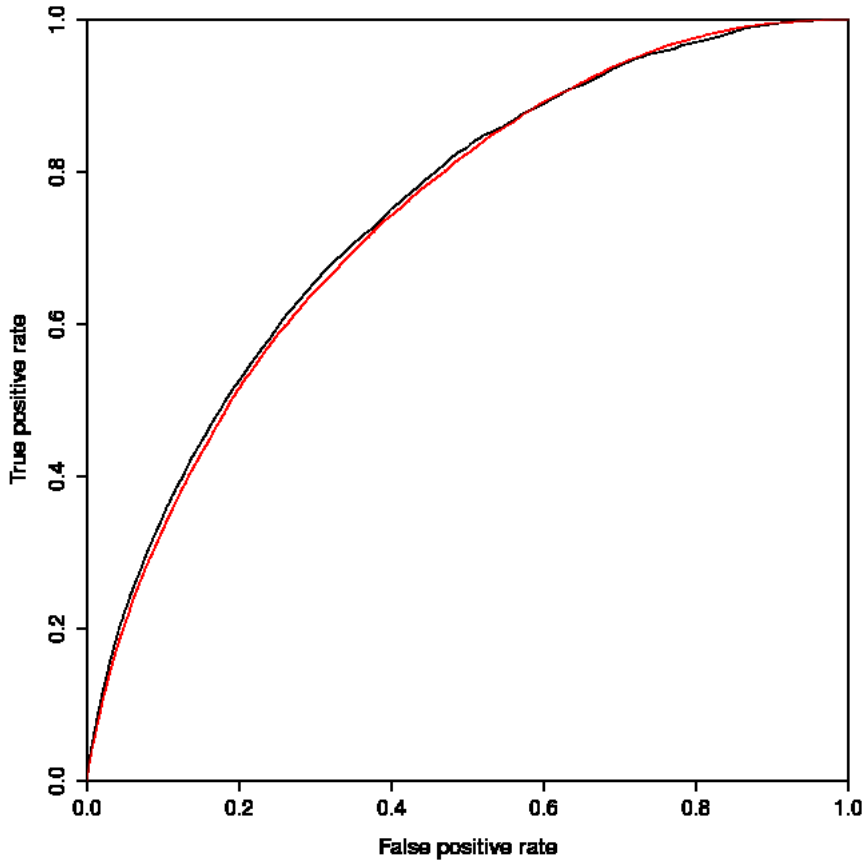

Figure 4: Courbes Roc des scores produits par régression logistique et par Model Producer.

2.6.1 Comparaison score régression logistique

Parallèlement à la construction du score avec le Model Producer de Predicsis, un score a été construit à l'aide d'une régression logistique.

On a ensuite construit les courbes ROC correspondant aux deux modèles.

On constate avec une certaine surprise que **sur notre jeu de données, la régression logistique et le score produit par le Model Producer présentent des performances tout à fait similaires,** avec des courbes ROC quasiment superposables.

Mais dans le cas de la régression logistique, la durée de construction du modèle de score peut être évaluée à environ une douzaine de jours homme.

Dans le cas de Model Producer, la mise en forme des données et la production du score final ont pu être effectuées dans la même journée (après la prise en main de l'outil et la préparation sommaire des données).

Conclusion sur la production de modèles Le Model Producer est donc un outil performant qui **améliore la productivité d'une équipe de data-miners**, même s'il ne garantit pas l'optimalité d'un modèle selon le critère de l'aire sous la courbe ROC. Dans le cadre de la production fréquente de scores servant à cibler les individus les plus disposés à présenter un comportement donné, Model Producer est donc une solution performante et ergonomique.

2.6.2 Scalabilité

Le jeu de données d'origine a été échantillonné avec remise pour produire des jeux de données de plus grande taille. Aucun problème n'a été rencontré jusqu'au million de lignes pour la table centrale. Cependant, nous rencontrons actuellement des soucis pour des jeux de données de 10 millions de lignes, soit un volume total d'environ 7 Go. La recherche des paramètres à régler est toujours en cours et un test de la version 3.4 devrait permettre de vérifier que ce problème est résolu.

Conclusion

Le Model Producer de Predicsis permet d'obtenir des scores de qualité tout à fait satisfaisante en termes d'indice de Gini et de robustesse, et dans des délais qui **améliorent la productivité d'un facteur parfois supérieur à 10 en présence de nombreuses prédictrices et/ou de mesures répétées de certaines variables prédictrices.**

Pour le détail de la procédure de discrétisation, voir [**?**], pour la construction des agrégats à partir des mesures répétées, voir [**?**].

Remerciements

Les travaux présentés dans ce document ont été financés, dans le cadre de la chaire Décisionnel-Connaissance Client de la Fondation UBS, par les entreprises suivantes : Crédit Agricole de Bretagne, Groupe Avril, Cofilmo et Business&Decision. La chaire D-CC remercie Predicsis pour la mise à disposition d'une licence ayant permis de tester le logiciel Model Producer.