UCE Application Business Intelligence
Projet :
Démissions d'un organisme bancaire

Résumé

L'objectif de ce projet est de mettre en place différentes approches issues de la fouille de données afin de pouvoir prédire les démissions de sociétaires d'un organisme de prêt bancaire. Il s'agira donc de pouvoir appliquer différents algorithmes de classification, avec objectif principal de pouvoir expliquer les choix de la décision de ces algorithmes. Cela permettra de mettre en valeur les caractéristiques conduisant à la démission, ou au contraire, au fait de rester dans l'organisme bancaire. Dans un second temps, vous explorerez des approches non-supervisées de clustering permettant de faire apparaitre des groupes d'utilisateurs ensemble.

1. Description du projet

Un certain nombre de variables ont été relevées sur les sociétaires démissionnaires d'un organisme bancaire (*i.e.* clients ayant quitté la banque), ainsi que sur un échantillon de sociétaires actuels non-démissionnaires. Une méthode d'analyse prédictive permet de modéliser le fait de démissionner ou pas à partir des valeurs des variables pour chaque sociétaire. Un score est produit, qui sera ici égal à la probabilité d'être démissionnaire.

Bien évidemment, si le modèle est pertinent, cette probabilité sera forte pour les démissionnaires, faible pour les sociétaires actuels. Mais il peut arriver que ce score soit élevé pour certains sociétaires (non-démissionnaires). On peut alors en conclure que le profil de ce sociétaire est proche de celui des démissionnaires, et donc que le risque de démission de ce sociétaire est élevé.

On a ainsi constitué un score d'attrition, ou risque de démission, calculable pour chaque sociétaire de la banque (et actualisable en fonction de l'évolution de sa situation). Ce score peut servir de base à des campagnes de gestion de la relation-client (prévention du départ des sociétaires).

2. Préparation des données

Deux jeux de données sont fournis.

data_mining_DB_clients_tbl.csv qui contient les 30 332 démissionnaires d'un organisme bancaire, entre 1999 et 2006.

Id	auto increment
CDSEXE	catégorielle. Ne pas s'inquiéter du nombre de modalités (1 à 4 !) qui permettait de différencier des classes ...
MTREV	numérique
NBENF	nombre d'enfants
CDSITFAM	situation familiale, catégorielle
DTADH	date adhésion à l'organisme

CDTMT	statut (siège ou bien tsmt) = catégorielle
CDDEM	code démission
DTDEM	date démission
ANNEE_DEM	année démission
CDMOTDEM	motif démission Rien si non-démissionnaire ou si motif inconnu.
CDCATCL	catégorie (sociétaire / adhérent)
AGEAD	âge du client à l'adhésion
rangagead	idem par tranches
agedem	âge du client à la démission
rangagedem	idem par tranches
rangdem	date démission sous format mois année
adh	durée en années de l'adhésion
rangadh	idem par tranches

data_mining_DB_clients_tbl_bis.csv contient un échantillon aléatoire de 15 022 sociétaires de la banque. Cible : démissionnaire (0 non / 1 oui).

Id	auto increment
CDSEXE	catégorielle
DTNAIS	data naissance adhérent
MTREV	montant revenu
NBENF	nombre d'enfants
CDSITFAM	situation familiale, catégorielle
DTADH	date adhésion à l'organisme
CDTMT	statut (siège ou bien tsmt) = catégorielle
CDMOTDEM	motif démission. Rien si non-démissionnaire
CDCATCL	catégorie (sociétaire / adhérent)
Bpadh	variable inconnue
DTDEM	date démission (31/12/1900 si non démissionnaire)

Remarque : Les variables suivantes permettent de repérer les démissionnaires :
- CDMOTDEM : motif démission
- ou DTDEM : date démission (31/12/1900 si non-démissionnaire)

Remarques :
- l'ancienneté dans l'organisme est obtenue :
  - pour les sociétaires actuels en retranchant 2007 (l'année de l'application) à l'année d'adhésion,
  - pour les démissionnaires en retranchant l'année de démission à l'année d'adhésion.

- l'âge est obtenu à partir de l'âge à l'adhésion et de l'ancienneté pour le premier fichier, de l'année de naissance pour le second fichier. Même remarque que pour l'ancienneté (c'est l'âge à la démission si démissionnaire et l'application est datée en 2007).

3. Préparation des données

Avant de réaliser la classification et clustering proprement dits, il est nécessaire de préparer les données, en suivant un certain nombre d'étapes standard.

Exploration. Il est recommandé, dans un premier temps, de naviguer manuellement dans les données afin de mieux les appréhender, et d'extraire quelques statistiques descriptives. Comment les attributs sont-ils distribués ? Quelles sont les valeurs moyennes, modales, les quantiles, etc. N'hésitez pas à produire des graphiques pour visualiser ces résultats, et ils permettront aussi d'illustrer votre analyse dans votre rapport.

Nettoyage. Puis se pose la question de la préparation des données proprement dite. Ces données nécessitent-elles un nettoyage ? Faut-il écarter certaines instances qui ne sont pas liées au problème ? Y a-t-il des valeurs manquantes ? Des valeurs aberrantes ? Des attributs redondants ? Des attributs superflus ? Les valeurs numériques correspondent-elles vraiment à des attributs de nature numérique, ordinale, ou catégorielle ? Comment traiter ces différents problèmes ?

Fusion. Les données fournies sont éclatées sur plusieurs fichiers. La question se pose donc de savoir s'il faut les fusionner, et si oui : comment ? Pensez à bien justifier toutes vos décisions. Là encore, plusieurs approches sont possibles, qu'il vous est possible de comparer empiriquement en les mettant en oeuvre et en les évaluant.

Recodage. En fonction des algorithmes de fouille que vous allez appliquer, il peut être nécessaire de recoder certains champs : discrétisation d'attributs réels, catégorisation d'attributs numériques, normalisation d'attributs numériques, numérisation d'attributs catégoriels... Certains outils ne peuvent pas du tout être appliqués sur des données dont le codage n'est pas approprié. D'autres fonctionneront mieux pour certains codages. Il est recommandé de tester l'effet du codage sur les différents outils considérés.

Prétraitement. Des méthodes de prétraitement peuvent être appliquées avant de réaliser le traitement proprement dit. Par exemple, effectuer une réduction de la dimension des données, peut permettre de rendre le problème traitable, computationnellement parlant (i.e. faire que l'outil de fouille s'exécute en un temps raisonnable), ou bien d'améliorer la qualité et/ou la lisibilité des résultats. Mais le prétraitement peut aussi les rendre difficile à interpréter. Là encore, il est possible de tester différentes méthodes avec différents paramétrages.

4. Analyse des données

L'analyse de données se décompose en deux parties relativement indépendantes. Tout d'abord, on veut étudier le lien entre la démission d'un sociétaire et les autres variables : il s'agit d'une tâche de classification supervisée binaire, dans laquelle la classe correspond à une démission ou non. Puis, on veut identifier des profils-types de démissionnaires : il s'agit d'une tâche de clustering (classification non-supervisée) dans laquelle chaque cluster correspondra potentiellement à un profil donné de démissionnaire. Cela peut bien entendu impliquer les non-démissionnaires.

Prédiction de la démission. Peut-on prédire la démission à partir des autres caractéristiques des sociétaires ? Pour répondre à cette question, la méthode proposée dans ce projet est d'effectuer une

classification supervisée. Entraînez au moins deux classifieurs de votre choix, afin de pouvoir comparer leurs performances, et analysez les résultats obtenus.
Le classifieur sélectionné doit être le plus interprétable possible, car on désire comprendre pourquoi la démission (ou non) est prédite. Il faut donc éviter à tout prix les outils de type boîte noire. Cette analyse doit permettre de déduire quels facteurs semblent décider de la démission d'un sociétaire, ou au contraire de conclure que les caractéristiques présentes dans les données ne sont pas suffisamment informatives.

Typologie des démissionnaires. L'étape suivante consiste à identifier des classes de démissionnaires. Pour cela, on se propose d'effectuer un regroupement non-supervisé (ou clustering). Sélectionnez au moins deux méthodes différentes, afin de pouvoir comparer leurs résultats. En fonction de votre choix, il est possible que vous deviez définir une mesure de dissimilarité permettant de comparer deux démissionnaires.
Une fois les classes identifiées, caractérisez et étudiez le profil de démissionnaires correspondant à chacune d'entre elles. Considérez en particulier l'homogénéité/hétérogénéité des classes relativement aux attributs décrivant les démissionnaires. Cette étape peut nécessiter d'appliquer un autre outil pour faciliter l'interprétation. Par exemple, on peut envisager de rechercher des règles d'association prédisant l'appartenance aux différents clusters, pour peu que ceux-ci soient suffisamment grands.

5. Implémentation

Vous devez fournir un script (ou un ensemble de scripts) en Python (le langage est imposé) qui, une fois lancé, effectuera l'intégralité du traitement à partir des fichiers originaux : préparation des données, application des algorithmes de fouille, calcul des performances, comparaison des algorithmes, etc. Aucune étape ne doit faire l'objet d'une intervention manuelle, de manière à pouvoir être facilement reproduit par la suite.

La manière dont ce script doit être exécuté devra être clairement expliquée à la fois dans le rapport (cf. la Section 2.4 du modèle de rapport mentionné en Section 6) et dans un fichier *readme.txt* à placer dans le dossier contenant le(s) script(s).

Tout ce qui peut être réalisé avec les bibliothèques utilisées en cours et TP (prétraitement des données, apprentissage des outils de fouille, calcul et comparaison des performances...) doit l'être en priorité. Si vous avez besoin de fonctionnalités supplémentaires, vous pouvez utiliser d'autres bibliothèques que celles-ci, mais cela doit être justifié dans le rapport (et le mieux est d'en discuter oralement en séance avec l'encadrant). Tout le reste du traitement doit être implémenté dans le script lui-même.

6. Rapport

En plus de votre code source (script, fichiers de configuration...), vous devez rendre un rapport décrivant le traitement que vous avez mis en place pour résoudre le problème proposé.

Structure et forme du rapport. Le plan du rapport est disponible en ligne sur Overleaf, à l'adresse suivante :

https://www.overleaf.com/read/yphxhkwmcxxm

Ce plan de rapport n'est accessible qu'en lecture seule. Donc, si vous décidez d'utiliser LATEX pour écrire votre rapport, vous devez d'abord en créer une copie avant de pouvoir l'éditer. Le rapport rendu doit être conforme aux instructions contenues dans le tutoriel suivant :

https://www.overleaf.com/latex/templates/modele-rapport-uapv/pdbgdpzsgwrt

Notez que vous n'êtes pas tenus d'utiliser LATEX : n'importe quel autre outil fait l'affaire, tant que le rapport rendu prend la forme d'un PDF. En revanche, la structure du rapport est imposée, vous devez la suivre obligatoirement, en respectant les titres et la numérotation indiquée. De plus, la gestion de la bibliographie doit respecter les standards LATEX (cf. le tutoriel indiqué ci-dessus).

Utilisation de ressources. Vous avez le droit (et c'est même recommandé) d'utiliser n'importe quelle ressource qui pourra vous aider dans votre travail : rapports, articles, code source, pages Web, etc. La seule restriction est que vous ne pouvez pas utiliser des ressources produites par d'autres groupes de ce projet.
De plus, **toute ressource doit explicitement être indiquée dans le texte de votre rapport**, là où elle est pertinente. Le détail de la source doit apparaître dans la dernière section du rapport (bibliographie), comme expliqué dans le tutoriel LATEX.

Avertissement : L'utilisation (citée ou non) d'une ressource issue d'un autre groupe, et l'utilisation non-citée ou incorrectement citée d'une ressource extérieure constituent des plagiats. En cas de plagiat, les groupes concernés seront sanctionnés en conséquence. Vous trouverez plus de détails sur la notion de plagiat dans le tutoriel LATEX cité précédemment.

7. Organisation

Le projet est à réaliser en groupes de **deux personnes**. Les étapes très fortement recommandées pour ce travail sont les suivantes :

1. Explorez les données, détectez les erreurs et problèmes et corrigez-les, nettoyez les données et effectuez leur analyse descriptive. Rédigez la partie du rapport correspondante.
2. Identifiez et étudiez les outils de fouille disponibles et susceptibles de résoudre le problème consistant à classifier les démissionnaires (de façon aussi bien supervisée que non-supervisée). Considérez notamment les paramètres possibles et les types de données supportés. Rédigez la partie correspondante du rapport.
3. Identifiez la préparation des données à effectuer pour que celles-ci soient exploitables par les outils sélectionnés. Il est possible de prévoir plusieurs types de préparations pour le même outil, par exemple dans le but de déterminer laquelle de ces préparations est la meilleure. Écrivez les scripts implémentant la préparation, et rédigez la partie correspondante du rapport.
4. Développez les scripts permettant d'invoquer les outils de fouille sélectionnés. Le passage à la pratique peut vous révéler de nouvelles informations concernant ces outils, à intégrer dans la partie de votre rapport qui décrit les outils sélectionnés.
5. Appliquez les outils, évaluez la qualité des classes obtenues. Comparez les performances des différents outils utilisés. Interprétez les résultats obtenus. Complétez la partie du rapport portant sur la prédiction de la démission et celle concernant l'identification de clusters.
6. Finalisez le rapport. S'il reste du temps, vous pouvez tester des prétraitements ou des outils supplémentaires. Une bonne piste est notamment la recherche de séquences fréquentes : on veut considérer un ensemble de facteurs caractéristiques pouvant conduire à la démission.