UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

M2 ILSEN – 2018/19
UE Business intelligence – Systèmes décisionnels
UCE Application Business Intelligence

Vincent Labatut

Projet | Campagne de prospect d'une compagnie d'assurances

1 Présentation

Une compagnie d'assurance désire augmenter son nombre de clients. Elle a constitué une base de données décrivant 370 000 entreprises françaises, et a sélectionné environ 108 000 prospects parmi elles, en se basant notamment sur leur situation financière. Les prospects ont été distribués aux caisses régionales de la compagnie, chacune traitant les entreprises présentes sur son territoire (cf. Figure 1).

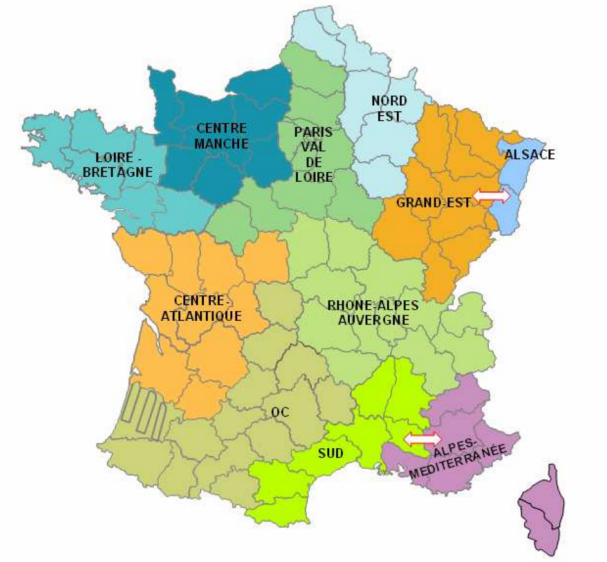

Figure 1. Caisses régionales de la compagnie d'assurance considérée.

Chaque entreprise a été sollicitée par la caisse correspondante, pour obtenir un premier rendez-vous téléphonique avec un commercial de la compagnie. Environ 12 000 d'entre-elles ont répondu favorablement. Le but de ce rendez-vous était de présenter les offres de la compagnie d'assurance, d'identifier les besoins de l'entreprise, et surtout d'obtenir un second rendez-vous. À long terme, la compagnie vise une augmentation de sa clientèle de 5%.

La compagnie désire maintenant effectuer une étude des données résultant de cette prise de contact. Il s'agit de présenter les résultats de la campagne, d'identifier les différents profils d'entreprises, et de décrire leur comportement vis-à-vis de la prise de rendez-vous.

2 Données

Les données fournies prennent la forme d'une table au format CSV, dont les attributs sont décrits dans la Table 1.

Tous les attributs sont issus d'une enquête préalable de la compagnie d'assurance, à l'exception de `rdv`. Celui-ci a été obtenu lors de la prise de contact avec les entreprises, et indique si elles ont accepté un rendez-vous avec un commercial.

L'attribut `risque` est une note attribuée à l'entreprise, et reflétant sa santé financière. Ainsi, 7 représente un risque élevé dû à une mauvaise situation financière, tandis que 14 correspond

Attribut	Signification
`code_cr`	Code de la caisse régionale
`dept`	Numéro du département
`effectif`	Effectif de l'entreprise (nombre d'employés)
`ca_total_FL`	Chiffre d'affaires (CA) total, en k€
`ca_export_FK`	Part du chiffre d'affaires à l'exportation
`Risque`	Score de santé financière de l'entreprise
`endettement`	Ratio d'endettement (capitaux propres / total bilan)
`evo_benefice`	Taux d'évolution du bénéfice
`ratio_benef`	Ratio sur bénéfice (= bénéfice / CA total)
`evo_effectif`	Taux d'évolution de l'effectif
`evo_risque`	Évolution du risque
`age`	Âge de l'entreprise
`type_com`	Type commune de résidence de l'entreprise
`activite`	Secteur d'activité de l'entreprise
`actionnaire`	Type d'actionnariat
`forme_jur_simpl`	Forme juridique (simplifiée) de l'entreprise
`chgt_dir`	Changement de direction récent (`0`=non, `1`=oui)
`rdv`	Rendez-vous pris (`0`=non, `1`=oui)

Table 1. Attributs de `table1.csv`.

à un risque faible, en raison de la très bonne situation financière de l'entreprise. Les prospects sélectionnés sont supposés tous avoir une note d'*au moins* 7.

3 Préparation des données

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 ?

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 trois parties. Tout d'abord, on veut comprendre comment le taux de pénétration (i.e. la proportion d'entreprise ayant accepté un rendez-vous) est affecté par leurs caractéristiques. Puis, on veut segmenter nos données afin d'identifier des classes d'entreprises similaires. Enfin, on veut déterminer les profils de ces classes, afin de pouvoir les utilister dans des campagnes futures.

Analyse descriptive des rendez-vous. Analysez de manière globale la répartition des rendez-vous obtenus auprès des prospects. Envisagez différents critères de recoupement possible, aussi bien financiers (ex. comment le chiffre d'affaire affecte-t-il le taux de pénétration ?), spatiaux (Quel est le taux de pénétration par département ? Par caisse ?), qu'autres (ex. comment l'effectif affecte-t-il le taux de pénétration ?). Notez que l'exploitation des données spatiales est l'occasion d'afficher les résultats obtenus sous forme de cartes.

Typologie des entreprises. L'étape suivante consiste à identifier des classes d'entreprises similaires. Pour cela, il faut effectuer une *classification non-supervisée* (ou *clustering*). Sélectionnez au moins deux méthodes différentes, afin de pouvoir comparer leurs résultats.

On veut recouper cette typologie avec trois attributs spécifiques, qu'il ne faut donc surtout pas utiliser lors de la classification : `code_cr` (code région), `dept` (code département) et `rdv` (rendez-vous).

Une fois les classes identifiées, caractérisez et étudiez le profil d'entreprise correspondant à chacune d'entre elles. Comment sont distribués les trois attributs mentionnés ci-dessus dans ces classes ? Plus précisément, certaines sont-elles homogènes pour ces attributs ?

Prédiction d'attributs. On se pose les deux questions suivantes : peut-on prédire l'attribut `rdv` à partir des autres champs ? Et la classe de l'entreprise identifiée à l'étape précédente ?

Pour répondre à ces questions, vous devez effectuer une classification supervisée. Entraînez un classificateur de votre choix en validation croisée, et analysez les résultats obtenus. Le classificateur sélectionné doit être le plus interprétable possible, car on désire comprendre *pourquoi* telle valeur de `rdv` ou telle classe est prédite. Il faut donc éviter à tout prix les outils de type boîte noire. Cette analyse doit permettre de déduire quels facteurs ont amenés à l'acceptation (ou au refus) de rendez-vous, et d'identier le profil d'entreprise associé à chaque classe.

L'intérêt de ce type d'analyse est que les résultats obtenus permettent d'optimiser les prochaines campagnes. En effet, sur la base des classes obtenues et de leur analyse, il est possible d'assigner de nouvelles entreprises [1] aux classes existantes. Sur cette base, on peut décider de démarcher en priorité certaines classes d'entreprises pour lesquelles on suppose avoir plus de chances, mais également personnaliser l'argumentaire de l'entretien téléphonique suivant la typologie de l'interlocuteur, et aussi avoir une meilleure idée de la difficulté d'obtenir un rendez-vous.

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.

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

1. C'est-à-dire : absentes de notre base au moment de cette analyse.

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. 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/dkkvhpznwfwq

> **Attention :** Veillez à bien visualiser la version du deuxième semestre, correspondant au fichier `planS2.tex`.

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) de ressources issues 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. Plus de détail sur la notion de plagiat dans le tutoriel LaTeX.

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, effectuez leur analyse descriptive (1er point de la Section 4), et rédigez cette partie du rapport.
2. Identifiez et étudiez les outils de fouille disponibles et susceptibles de résoudre le problème consistant à classifier les entreprises (2ème point de la Section 4). 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 ces outils (éventuellement plusieurs pour chaque outil). Écrivez les scripts implémentant la préparation, et rédigez la partie correspondante du rapport.
4. Développez les scripts permettant d'invoquer les outils. Complétez éventuellement la partie portant sur les outils de fouille dans le rapport.
5. Appliquez les outils, évaluez la qualité des classes obtenues. Étudiez la distribution des trois attributs ciblés. Complétez la partie du rapport portant sur la typologie des entreprises.

6. Repartez au point 2) et appliquez la même approche pour la prédiction d'attributs (3ème point de la Section 4). Complétez la partie du rapport portant sur la prédiction d'attributs.
7. Finalisez le rapport. S'il reste du temps, vous pouvez tester des prétraitements ou des outils supplémentaires.