MÉMOIRE MÉTHODOLOGIQUE
Mai 2022
Projet 7 - Data Science : «Implémentez un modèle de scoring »

1. Introduction

Ce mémoire constitue l'un des livrables du projet «Implémentez un modèle de scoring » du parcours Data Science d'Openclassrooms. Il présente le processus de modélisation et d'interprétabilité du modèle mis en place dans le cadre du projet.

Ce dernier consiste à développer pour la société « Prêt à Dépenser », une société de crédit de consommation, un modèle de scoring de la probabilité de défaut de paiement de nouveaux clients.

Pour ce faire, une base de données de 307 000 clients comportants 121 features (âge, sexe, emploi, logement, revenus, ancienneté dans la banque, etc.) est à disposition.

2. Méthodologie d'entraînement du modèle

Après avoir effectué une analyse exploratoire et du feature engineering, un modèle LGBM a été choisi pour prédire la capacité de remboursement des clients. Ces étapes ont été réalisés en s'inspirant du kernel Kaggle suivant :

https://www.kaggle.com/code/willkoehrsen/start-here-a-gentle-introduction

Le jeu de données initial a été séparé en plusieurs parties de façon à disposer :

- d'un jeu de training (80% des individus) qui a été séparé en plusieurs folds pour entraîner les différents modèles et optimiser les paramètres (cross validation) sans overfitting;

- d'un jeu de test (20 % des individus) pour l'évaluation finale du modèle.

Le cas d'étude est un problème de classification binaire (mauvais payeur/bon payeur) avec une classe sous représentée (9 % de clients en défaut contre 91 % de clients sans défaut). Ce déséquilibre des classes doit être pris en compte dans l'entraînement des modèles puisqu'un modèle « naïf » prédisant systématiquement que les clients sont sans défaut aurait une accuracy (précision) de 92% et pourrait être considéré à tort comme un modèle performant alors qu'il ne permettrait pas de détecter les clients à risque. En effet, le modèle serait capable de classifier correctement la majorité des clients bons payeurs mais n'identifierait que très peu de mauvais payeurs.

2.1. Techniques d'échantillonnage

Afin de palier à ce problème, plusieurs approches ont été testées et évaluées :

- Un under-sampler dans le but de réduire le nombre de bons payeurs de notre base de données et ainsi faire en sorte qu'il soit égal au nombre de mauvais payeurs;

- Un over-sampler (SMOTE) dans le but d'augmenter artificiellement le nombre de mauvais payeurs de notre base de données pour qu'il soit égal au nombre de bons payeurs;

- Un class-weight dans l'élaboration des modèles pour spécifier à ces derniers de prendre en compte la répartition des payeurs au sein de notre base.

2.2. Métrique d'apprentissage

En gardant à l'esprit que le destinataire du modèle est une banque cherchant à minimiser son taux de mauvais payeurs et à maximiser son taux de bons payeurs, il semble nécessaire d'utiliser une métrique différente de l'accuracy. En termes de Machine Learning, ce problème revient donc à maximiser le recall et la précision, respectivement définis comme :

- $Recall = vrais\ positifs\ /\ (vrais\ positifs + faux\ négatifs)$
- $Precision = vrais\ positifs\ /\ (vrais\ positifs + faux\ positifs)$

En supposant que la banque préférera limiter un risque de perte financière plutôt qu'un risque de perte de client potentiel, nous utiliserons une fonction permettant de prendre en compte ces 2 critères tout en donnant plus d'importance au recall.

La fonction Fβ-score est définie par :

$$F_\beta = \frac{(1+\beta^2)\cdot(\text{précision}\cdot\text{rappel})}{(\beta^2\cdot\text{précision}+\text{rappel})}$$

où β est le coefficient d'importance relative accordée au recall par rapport à la précision.

Pour répondre aux attentes de la banque, une valeur de β = 3 est choisie. Faute de temps, un passage en revue de différentes valeurs de β a été effectué et cette valeur est celle qui permet de maximiser les différentes métriques et est optimale pour notre problématique.

3. Entraînement des modèles

Considérant les contraintes établies préalablement, trois modèles de classification sont sélectionnés :

- Light Gradient Boosting Machine (LGBM)
- Random Forest
- Ridge

3.1. Résultats obtenus

Pour chacun des modèles sélectionnés, un entraînement est réalisé en faisant varier les méthodes d'échantillonnage et en réalisant des validations croisées. Les résultats suivants sont ainsi obtenus pour le β-score:

Modèle	Under-sampling	SMOTE	Class-Weight
LGBM	0.464	0.115	0.447
Random Forest	0.462	0.084	0.002
Ridge	0.477	0.482	0.482

De manière générale, la technique d'under-sampling permet d'obtenir les meilleurs résultats. Ridge est par ailleurs le meilleur modèle pour l'ensemble des techniques d'échantillonnage.

Cependant, le modèle LGBM obtient des résultats équivalents à Ridge sous Under-sampling, ce premier sera donc préféré dans la mesure où il offre une meilleure interprétabilité de ses prédictions, ce qui permettra à la banque de savoir précisément les raisons qui ont conduit le modèle à effectuer une prédiction de remboursement.

Après entraînement du modèle, le score obtenu est finalement de 0.464.

3.2. Interprétabilité du modèle

Interprétabilité locale

Localement, le modèle peut être expliqué aux équipes commerciales de la banque à l'aide de l'interpréteur LIME qui permet d'expliquer, pour chaque réponse donnée, les variables d'entrée ayant motivé ce choix.

Globalement, l'interpréteur SHAP permet de comprendre quelles sont les variables d'entrée qui ont la plus grande influence sur l'ensemble des réponses générées.

Ces deux interpréteurs sont implémentés dans le Notebook et le Dashboard afin d'offrir une compréhension plus poussée du jeu de données.

4. Conclusion et axes d'amélioration

La mise en place du modèle a été effectuée sous l'hypothèse selon laquelle β=3 était le meilleur choix possible pour nos prédictions. Cette dernière nécessite d'être confirmée par un expert métier qui sera plus à même de définir explicitement ses besoins.

De plus, certaines caractéristiques plus précises des anciens clients sont également disponibles et n'ont pas été utilisées dans le cadre de ce projet par manque de temps. Ces dernières pourraient être utilisées afin de perfectionner le modèle de prédiction.

Enfin, la volonté de fournir une interprétation du modèle limite également les résultats obtenus, comme cela a pu être vu avec le classifieur Ridge. Une amélioration des modèles pourrait donc également se faire au profit de l'inteprétabilité fournie.