Implémentez un modèle de scoring

1. Introduction

L'entreprise "**Prêt à dépenser**", qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt, souhaite mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Pour ce but, l'entreprise souhaite développer un algorithme de classification en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de transparence vis-à-vis des décisions d'octroi de crédit. Cette demande de transparence des clients va tout à fait dans le sens des valeurs que l'entreprise veut incarner.

Prêt à dépenser décide donc de développer un dashboard interactif pour que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

Ce travail a pour objectif principal:

A. Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique.
B. Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client des chargés de relation client.

Pour répondre aux objectifs de cette mission, données de 307511 clients, disponibles sur https://www.kaggle.com/c/home-credit-default-risk/, ont été utilisées.

2. Nettoyage et ingénierie des caractéristiques des données

Le premier aperçu sur le jeu de données a révélé un déséquilibre entre les deux classes représentées (Figure 1). Seulement 8% des clients ont des défauts de paiement contre 92% des clients qui présentent des crédits remboursés (Figure 1).

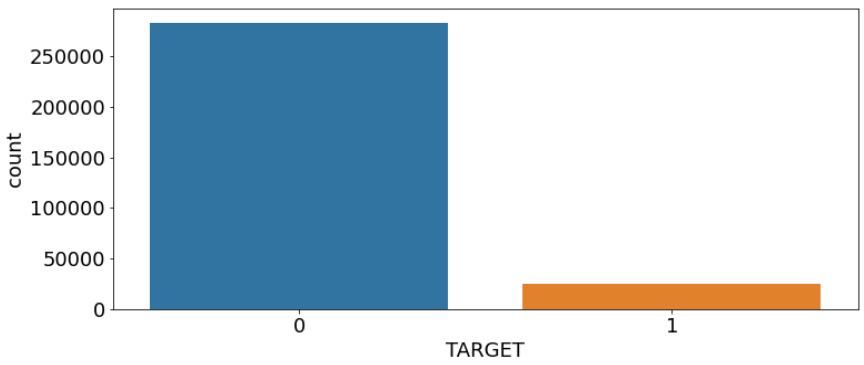


Figure 1. Représentation des différentes classes sur le jeu de données

Ensuite, nous avons procédé au nettoyage et ingénierie des caractéristiques. Les données ont été nettoyées, en supprimant les variables avec plus de 50% des données manquantes. Le restant de valeurs manquantes ont été remplacées par la moyenne. Pour traiter les valeurs aberrantes dans l'ensemble de données, nous avons utilisé l'écart interquartile. Pour cela, nous avons fixé des limites de 5% et 95% pour retirer les valeurs aberrantes. Les variables catégorielles ont été transformées en variables numériques, en utilisant la méthode OneHotEncoder**.**

3. Méthodologie d'entraînement du modèle

La stratégie de modélisation utilisée dans ce travail comprend le test de différents schémas de préparation des données, de différents algorithmes d'apprentissage et de différents hyperparamètres pour les algorithmes d'apprentissage. La procédure de construction de modèle ayant le meilleur score sera sélectionnée et utilisée.

- **Test de différents schémas de préparation des données**

Pour surmonter le problème des données non équilibrées, nous avons effectué la modélisation sur 3 ensembles de données clients:

A. Non-équilibrées
B. Équilibrées à l'aide de la méthode de sous-échantillonnage NearMiss
C. Équilibrées à l'aide de la méthode de sur-échantillonnage SMOTE

Les données clients sont ensuite séparées en trois jeux de données: les jeux d'apprentissage (70%), de validation (15%) et de test (15%).

- **Différents algorithmes d'apprentissage**

On a testé 3 modèles différents, notamment: *LGBMClassifier, CatBoostClassifier* et *XGBClassifier*.

- **Différents hyperparamètres pour les algorithmes d'apprentissage**

On a utilisé le *StratifiedKFold* pour la validation croisée (5 folds) et le *GridSearchCV* pour optimiser les paramètres des modèles.

3.1 Fonction coût métier

La fonction de coût métier utilisé dans cette exercice est la suivante:

$$Fc = -2 * (tn + tp) + fp + 10 * fn$$

où: ***tn*** est vrai négatif, ***tp*** est vrai positive, ***fp*** est faux positive et ***fn*** est faux négatif. Cette fonction permet de minimiser le ***fn*** et ***fp*** (qu'est coûteux pour la banque) tout et en récompensant à chaque fois que le modèle prédit correctement le ***tn*** et ***tp***. Dans cette étape, le jeu de données de validation a été utilisé pour la prédiction après un entraînement du jeu de données d'apprentissage.

3.2 Algorithme d'optimisation et métrique d'évaluation

La meilleure combinaison d'hyperparamètres de chaque algorithme a été identifiée et utilisée. En utilisant la meilleure combinaison d'hyperparamètres et la fonction de coût métier, on a construit un modèle et entraîné à nouveau le jeu d'apprentissage. La prédiction a été faite avec le jeu de données test. Le meilleur modèle est donc CatBoostClassifier avec méthode de sur-échantillonnage (Figure 2 et 3). Ce choix a été effectué en retenant le modèle avec le:

- Meilleur score **AUC** sur le jeu de données test.
- ***fp et fn*** le plus faible sur le jeu de données test.

- ***tp et tn*** le plus élevé sur le jeu de données test.

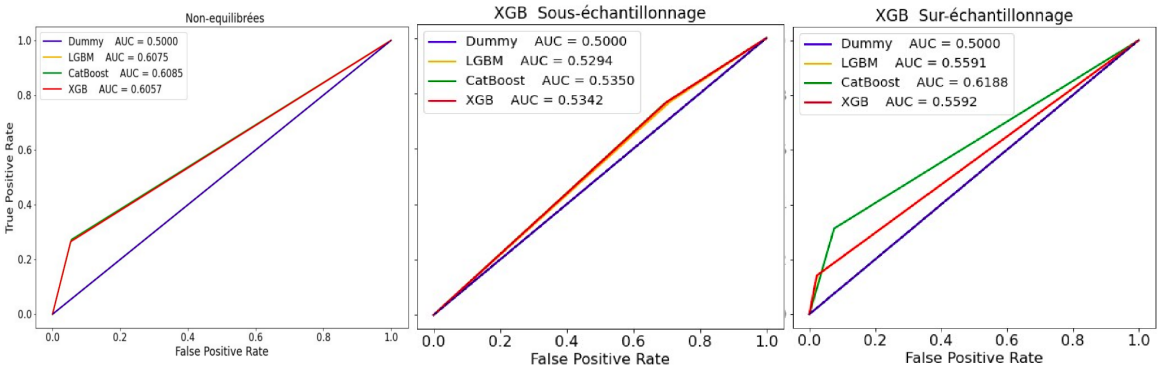


Figure 2. Courbe ROC pour les trois différents modèles et schémas de préparation des données utilisés

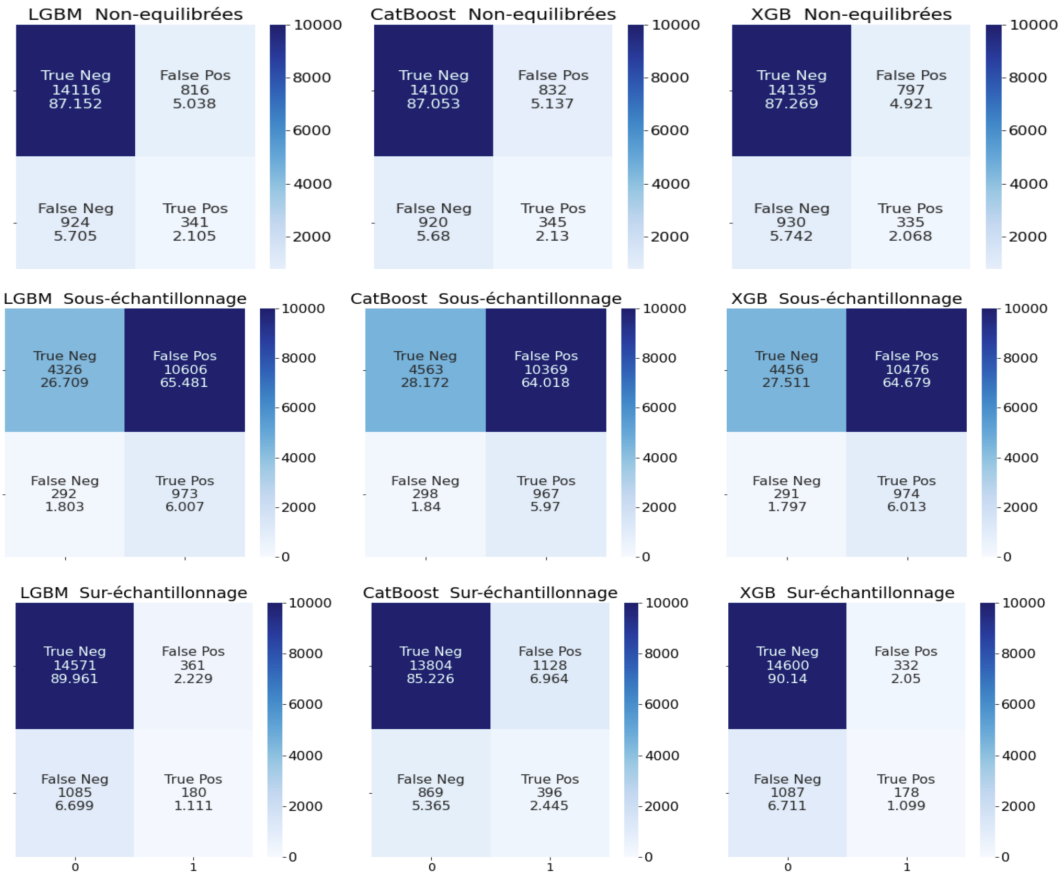
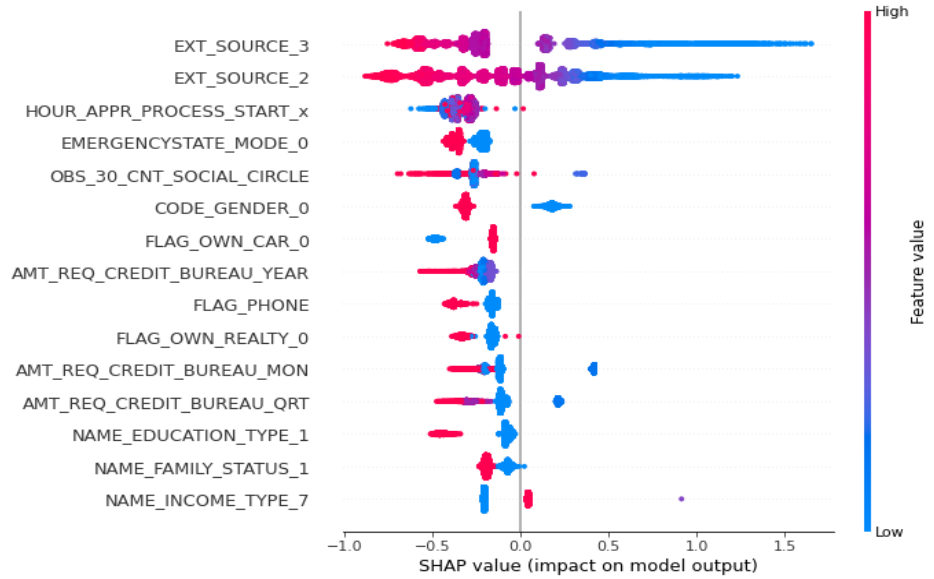


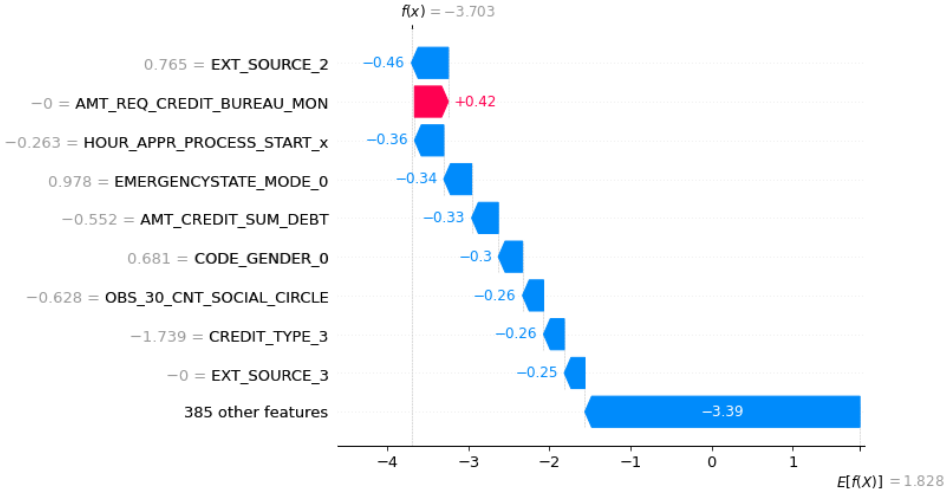
Figure 3. Matrice de confusion dérivés des 9 simulations effectués

4. Interprétabilité globale et locale du modèle

Figure 4, qui présente la tendance moyenne de l'importance des différentes caractéristiques dans décisions d'octroi de crédit, montre que *ext_source_2* et *ext_source_3* sont les caractéristiques les plus influentes dans la décision d'octroi de crédit. Ces caractéristiques ont tendance à pousser la prédiction vers la mobilité descendante, réduisant de cette façon les chances d'un refus de prêt. Au contraire, le Hour_APPR_PROCESS_START_X poussent la prédiction vers la mobilité ascendante, augmentant la chance d'un refus de prêt.



Cet ordre d'importance des caractéristiques peut varier d'un client à l'autre. On peut le constater dans l'exemple ci-dessous (Figure 5). Dans cette figure, où la couleur représente la direction de l'influence des caractéristiques, il est visible que seule la caractéristique xxx tend à contribuer au rejet du prêt.



5. Limites et les améliorations possibles

- Chercher une fonction coût optimale
- Utiliser la fonction métier pendant l'entraînement du modèle
- Tourner en directe le modèle dans l'API