OPENCLASSROOMS

Note méthodologique

Mai 2022

Implémentez un modèle de scoring

Réalisé par :

Ibtissam RADI

Table des matières

1 Contexte

Cette note constitue un des livrables demandés du projet **"Implémentez un modèle de scoring"** du parcours Data Scientist d'Openclassrooms.

Le projet consiste à développer pour la société **"Prêt à dépenser"**, une société qui propose des crédits à la consommation pour les personnes ayant peu ou pas d'historique de prêt, un modèle de scoring qui permet de calculer la probabilité qu'un client rembourse son crédit et ainsi lui accorder ou pas le crédit .

Les données utilisées pour ce projet sont une base de données de clients comportant différents informations financières, comportementale ...etc (âge, sexe, emploi, situation familiale, revenus, ..etc)
Un projet de machine learning se construit en plusieurs étapes successives. La mise en oeuvre a été opérée selon une démarche fréquente qui se construit de la manière suivante :

— Pré traitement des données : avant l'utilisation du jeu de données récupéré sur Kaggle à des fins de prédiction, le traitement des données est une tache importante. En effet les données brutes sont bruitées et incomplètes. Leur utilisation peut générer des résultats trompeurs d'où l'intérêt de procéder à certaines modifications.

— Construction du modèle de prédiction à partir des données pré traitées, en paramétrant les algorithmes et ainsi les implémenter de manière optimale

— Évaluation de l'efficacité des algorithmes entraînés en calculant l'erreur de l'apprentissage en utilisant une métrique d'évaluation, à savoir l'aire sous la courbe ROC(AUC).

2 Méthodologie d'entraînement du modèle

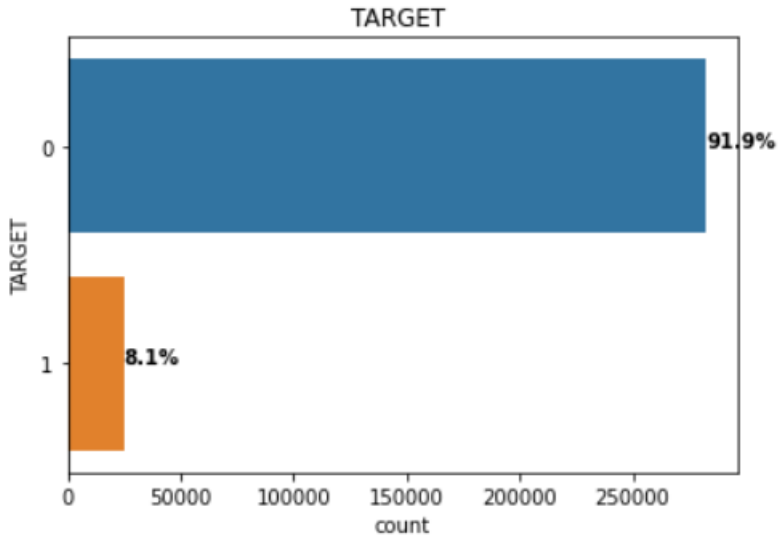
L'entrainement d'un modèle nécessite une étape nécessite une étape intermédiaire entre le preprocessing et la modélisation. Ce modèle de prédiction peut être représenté par une fonction qui prend des données en entrée et une décision en sortie. Dans le cas d'apprentissage supervisé, l'échantillon est classiquement subdivisé en 2 parties :

— Échantillon d'apprentissage il sert à ajuster le modèle

— Échantillon de test utilisé pour évaluer le modèle optimal (au sens du résultat de la validation croisée)

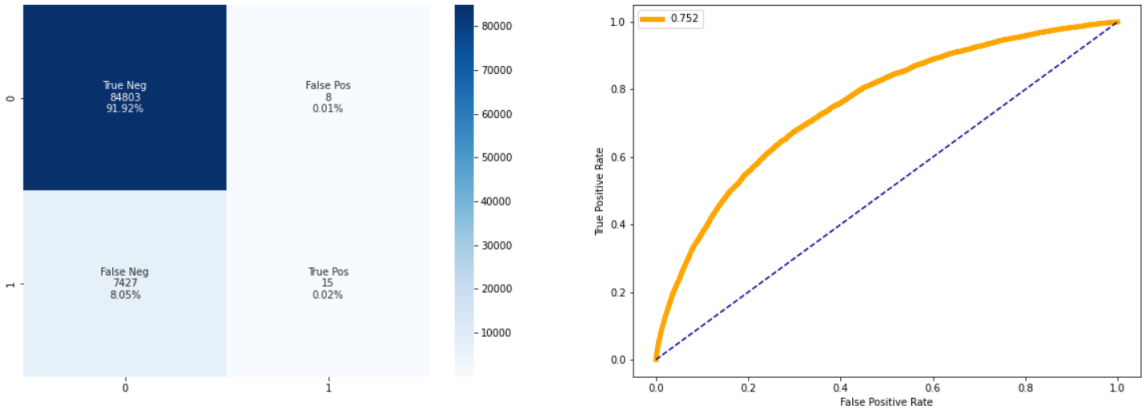
Ce découpage de données dans un projet de Machine learning est une étape très importante qu'il ne faut pas négliger car il existe un risque de surévaluer le modèle ou le sous évaluer.
Ce projet de machine learning traite une une classification binaire, donc il faut s'assurer que notre échantillon d'entraînement contient une proportion raisonnable des deux classes 0 et 1. L'analyse exploratoire a permis d'identifier un déséquilibre entre les deux classes 0 et 1, en effet l'échantillon contient 92% en classe 0 et seulement 8% en classe 1. Dans ce cas un traitement Oversampling(sur échantillonnage) permet d'ajuster la distribution de classe de manière à avoir des proportions égales des 2 classes .

Déséquilibre de la variable cible

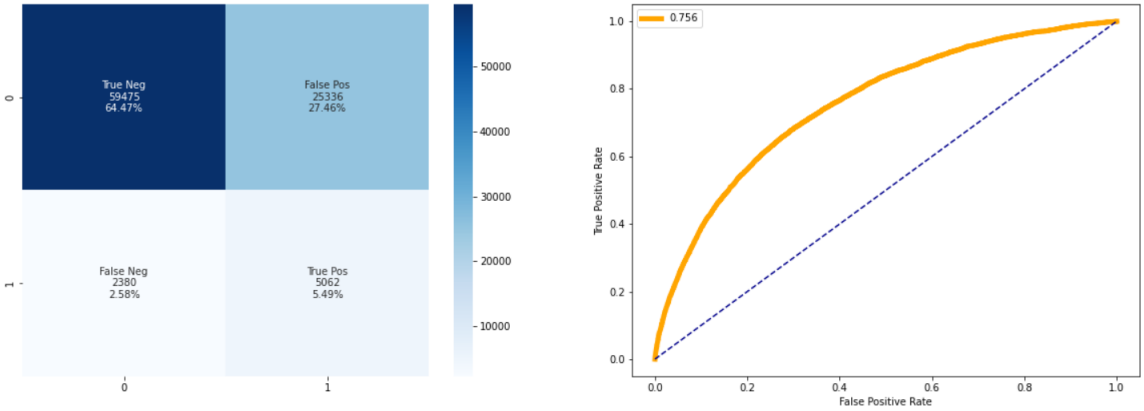


Pour le sur échantillonnage, **SMOTE** est considéré comme l'un des algorithmes les plus populaires dans le Machine learning, avec cet algorithme la classe minoritaire est sur échantillonnée en créant des exemples "synthétiques".

Exemple de résultats obtenus avant utilisation de SMOTE



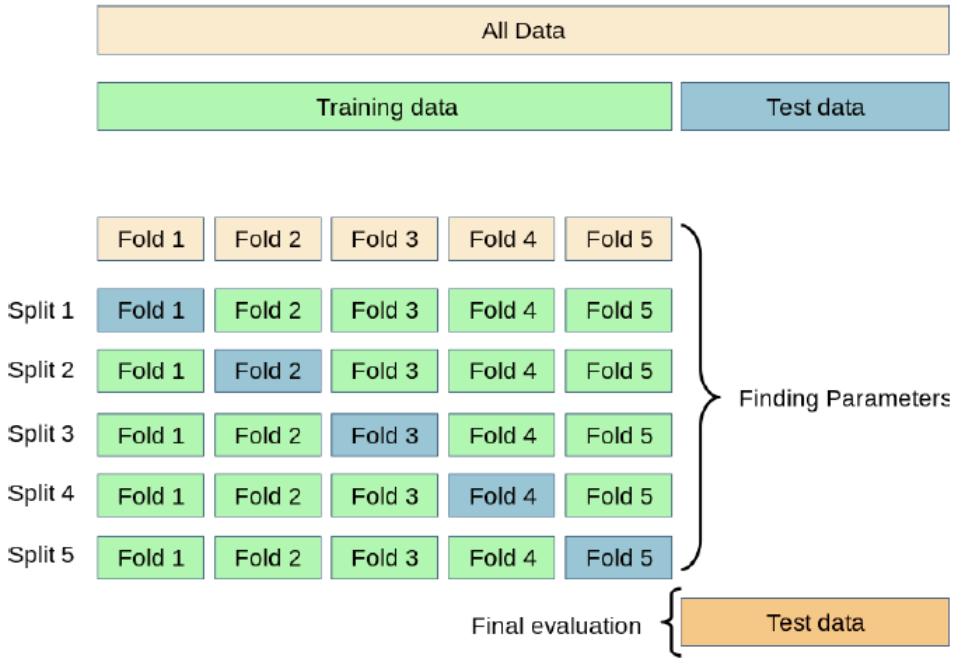
Exemple de résultats obtenus après utilisation de SMOTE



Pour la modélisation, différents modèle ont été testées avec recherche d'hyperparamètres et cross validation (5folds) :

1. **Logistic Regressor**
2. **XGboost**
3. **Random Forest Classifier**

Démarche d'entraînement d'un modèle



Le choix du modèle a été effectué en retenant le modèle avec le meilleur score sur le jeu de validation.

3 Fonction coût et métrique d'évaluation

L'entreprise nommée "Prêt à dépenser" lutte contre les défauts de paiement des clients, les pertes financières ne sont en effet pas souhaitables (frais de recouvrement, pertes,...).
Le modèle ne permettra pas d'éviter totalement ce risque, à titre d'exemple une erreur de prédiction aura pour conséquence soit un défaut de paiement du client, soit un refus de crédit à un client qui pourrait rembourser sa dette sans aucune défaillance. Les erreurs de prédiction doivent être minimisées, dans cette logique une fonction coût ayant pour objectif de pénaliser les Faux Positifs et les Faux Négatifs a été implémentée.

Terminologie :

- Faux positifs (FP) les cas où la prédiction est positive, mais où la valeur réelle est négative. Perte d'opportunité si le crédit client est refusé à tort, alors qu'il aurait été en mesure d'être remboursé.
- Faux négatifs (FN) les cas où la prédiction est négative, mais où la valeur réelle est positive. Perte réelle si le crédit client accepté se transforme en défaut de paiement.
- Vrais positifs (TP) les cas d'acceptation, le crédit client sera remboursé.
- Vrais négatifs (TN) les cas de refus, le crédit client ne pourra pas être remboursé.

Ainsi, les pertes d'un crédit en raison d'une mauvaise classification dépendent des probabilités Faux Positifs et Faux Négatifs. L'idée est d'éviter les clients avec un fort risque de défaut. Il est donc nécessaire de pénaliser les FP et FN cités précédemment.Pour réduire ce risque de perte financière, il faut maximiser deux critères Recall et Precision.

$$Recall = \frac{\text{TP}}{\text{TP+FN}}$$

$$Precision = \frac{\text{TP}}{\text{TP+FP}}$$

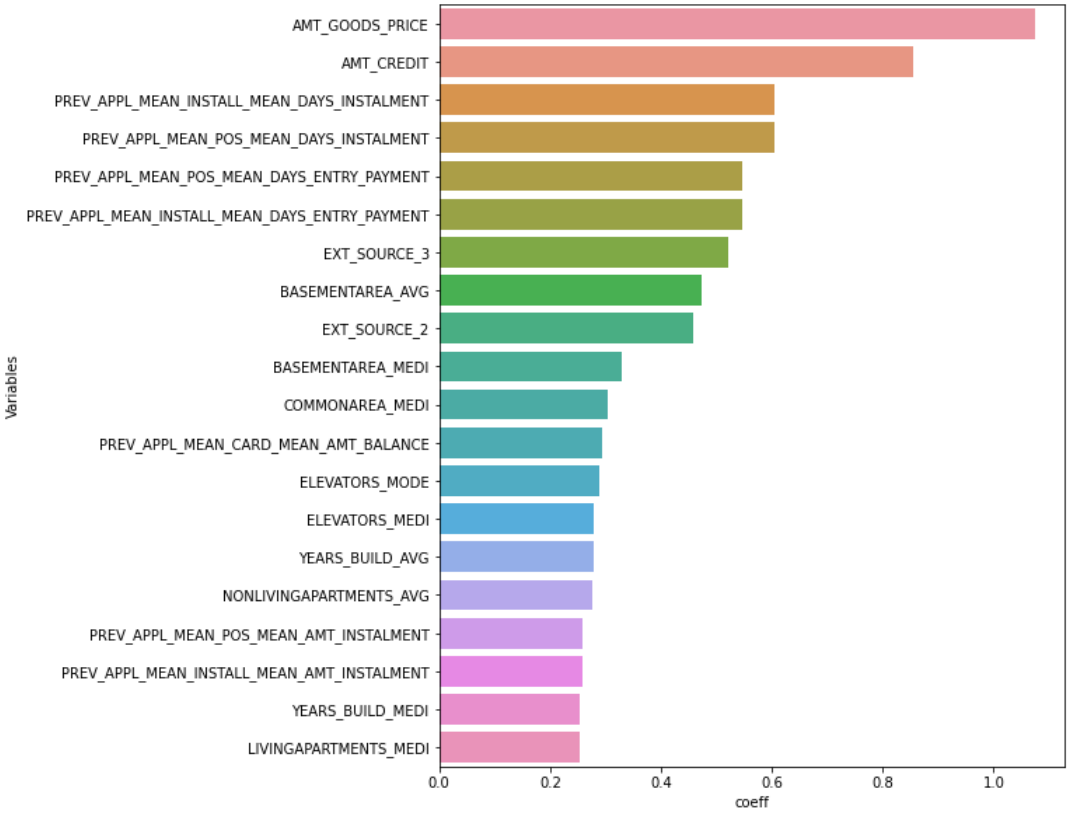
Pour notre problématique métier, le Recall est plus important que la Precision car on préférera vraisemblablement limiter un risque de perte financière plutôt qu'un risque de perte de client potentiel.
On cherche donc une fonction qui optimise les 2 critères en donnant plus d'importance au recall. Fonction permettant de faire cela : F Beta Score

(https ://en.wikipedia.org/wiki/F1_score) avec Beta le coefficient d'importance relative du recall par rapport à la précision.

$$F_\beta = (1 + \beta^2) * \frac{\text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}}$$

4 Interprétabilité du modèle

La réponse au besoin d'interprétabilité est prépondérante, le contexte de prédiction n'est pas uniquement appliqué à des experts de la data science mais au contraire à des experts du crédit. Un chargé de clientèle doit pouvoir utiliser le modèle via l'application mise à disposition, en face à face avec son client, dans le but de lui expliquer le plus simplement possible la décision envisagée dans l'étude de son dossier.
En d'autres termes, « l'interprétation » désigne l'évaluation globale du processus de prise de décision. Elle vise à représenter l'importance relative de chaque variable.



5 Limites et améliorations

La modélisation effectuée dans le cadre du projet a été effectuée sur la base d'une hypothèse forte : la définition d'une métrique d'évaluation : le F Beta Score avec Beta fixé suivant certaines hypothèses non confirmées par le métier. L'axe principal d'amélioration serait de définir plus finement la métrique d'évaluation en collaboration avec les équipes métier.
Par ailleurs, la partie de traitement préalable du jeu de données a été abordée de façon superficielle peut être un perfectionnement de traitement des données ainsi que la création des nouvelles variables en collaboration avec les équipes métier peuvent améliorer la modélisation
le choix des hyperparamètres peut aussi faire la différence, la question d'élargissement vers d'autres hyperparamètres peut également permettre d'augmenter les performances actuelles.