OPENCLASSROOMS

FORMATION DATA SCIENTIST

NOTE MÉTHODOLOGIQUE

Implémentez un modèle de scoring
-Projet 7-

1 Introduction

La société « **Prêt à dépenser** » est une société financière qui accorde des prêts sous plusieurs formes à leurs clients. Pour éviter des pertes financières liées à un défaut de remboursement, elle souhaite mettre en œuvre un outil permettant d'évaluer la situation de leur client avant tout octroi de crédit. Il s'agit donc d'élaborer un modèle permettant de résoudre un problème de classification supervisée binaire. Aussi, pour remplir à bien cette tâche, nous avons choisi dans un premier temps, de concevoir un modèle de « **scoring crédit** » pour calculer la probabilité qu'un client rembourse son crédit, d'autre part de développer un « **dashboard** » pour aider à la prise de décision des gestionnaires de crédit.
La note méthodologie présentée ici en lien avec cette mission s'articule autour de quatre point principaux :

- La description des données et méthodologie;
- La fonction coût métier;
- L' interprétabilité du modèle;
- Les limites et améliorations possibles

2. Description des données et méthodologie

a) Données

Les données utilisées pour ce travail proviennent de la base de données Kaggle (https://www.kaggle.com/c/home-credit-default-risk/data). Elles sont constituées d'un ensemble de dix (10) fichiers de données contenant l'historique des demandes de prêt de la société financière «P**rêt à dépenser**». On peut citer en autres les informations générales des clients (âge, sexe, situation familiale, revenus, etc…) et des informations relatives au crédit des clients (montant du crédit, type de crédit, date de demande du crédit, etc…).
Parmis ces 10 fichiers, il y a le fichier de données d'entraînement appelés « **application_train.csv** » contient 307 511 clients et 122 variables dont la variable cible « TARGET » avec la valeur 0 pour client en règle et 1 pour client non solvable.

b) Méthodologie

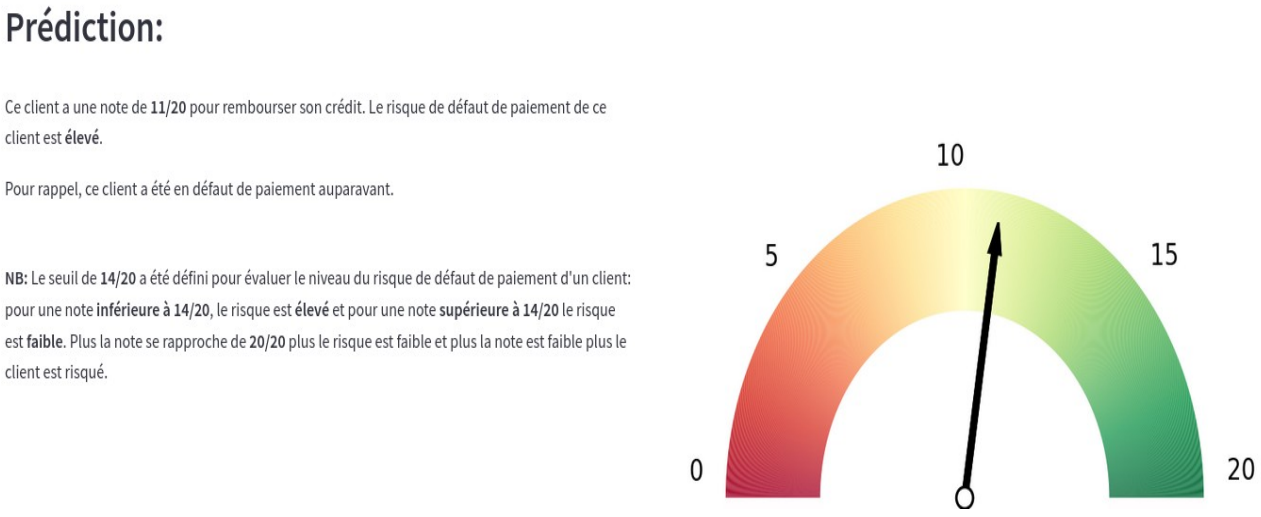
Dans cette section, j'aborde mon approche méthodologique.
Tout d'abord, pour comprendre les données que j'avais en ma possession, j'ai commencé par faire une analyse succincte des données (voir notebook_analyse_exploratoire.ipynb).
Puis, comme conseillé, j'ai sélectionné un kernel Kaggle (LightGBM with Simple Features) pour le prétraitement des données. Tout au long du kernel, des traitements sont faits pour transformer les variables catégorielles et pour créer de nouvelles variables.
À la fin, il y a une unification des jeux de données. Ensuite j'ai effectué un traitement des données manquantes en supprimant tout d'abord les colonnes avec 60 % de valeurs manquantes. Puis, en

remplaçant les valeurs manquantes par la moyenne de la distribution de chaque variable (voir preprocessing_1_S.ipynb).

Même si dans le Kernel choisi, une fonction à été codé afin d'entraîner les données avec l'algorithme LightGBM et la métrique d'évaluation ROC_AUC, il était important au vu des nombreux modèles de machine learning existantes de faire une comparaison des modèles pour choisir le modèle le plus adapté. Cependant, l'analyse exploratoire du jeu de données à montrer un déséquilibre des classes avec une surreprésentation de la classe '**0**' (91.9%). Aussi pour mieux comparer nos modèles et surtout choisir le bon modèle pour de meilleur raison il était important d'utiliser une méthode qui traite des données déséquilibrées telles que **SMOTE, class weight,..**
Pour cette étude j'ai d'abord choisi d'utiliser l'algorithme SMOTE. Ainsi, le jeu de données qui était de 91.9% de clients solvables et 8.1% de clients non solvables est devenue un jeu de donnée avec 50% de chaque classe. Mais, cette méthode de réechantillonnage avant l'entraînement créer un overfiting. Aussi pour ces données déséquilibrés j'ai finalement choisi d'utiliser des modèles d'ensembles qui sont entraînés à chaque étape de l'algorithme sur un échantillon rééquilibré automatiquement entre les différentes classes. Enfin les 3 métriques **ROC_AUC**, **f1_score** et **recall** qui limite les **faux négatifs** (la prédiction de client non solvables prédits comme solvables) et maximiser les **vrai positifs** (la prédiction des clients solvables prédits comme solvables) m'ont permis d'évaluer les modèles et de trouver le meilleur modèles et de l'optimiser en adaptant les hyperparamètres pour éviter des pertes financières à la société (voir notebook_autre_modelisation.ipynb).

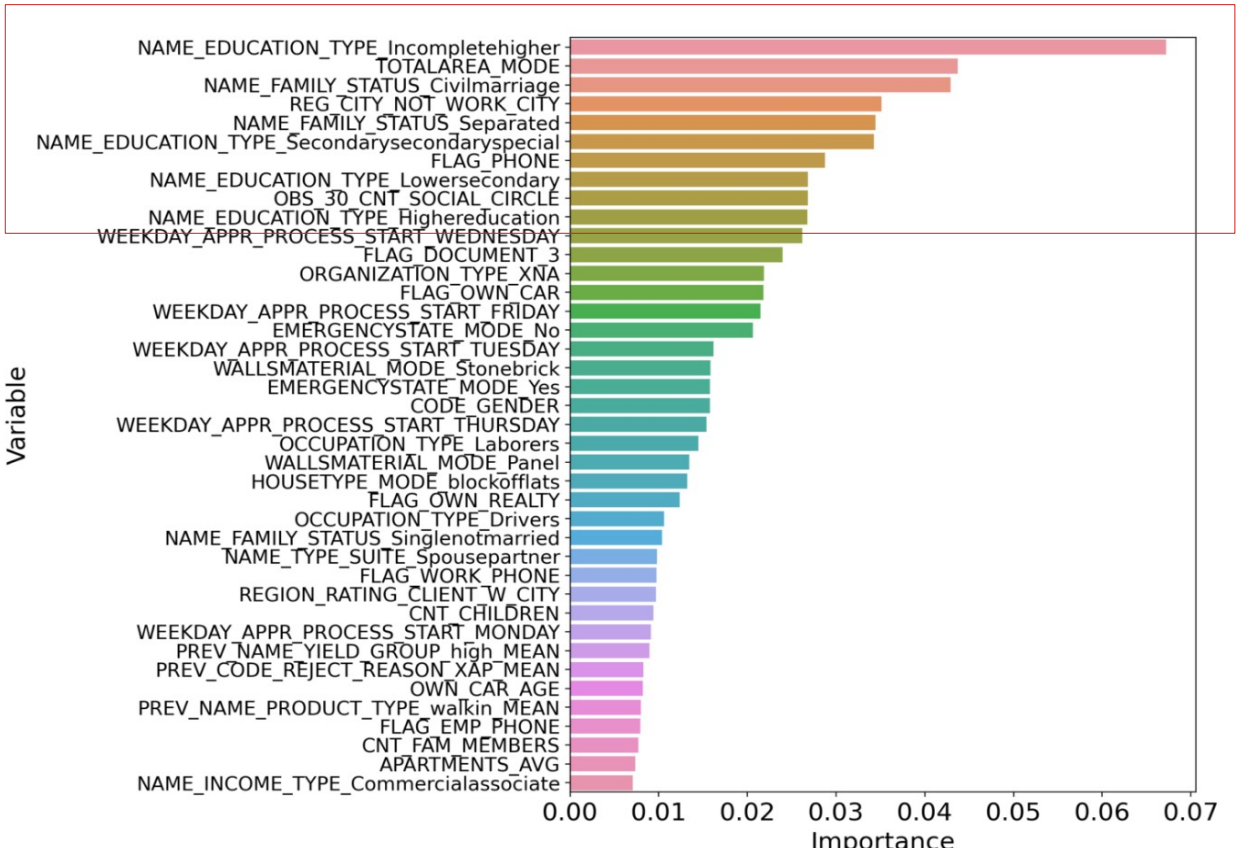
3. Fonction coût métier

La fonction coût métier est une fonction qui prend en compte les exigences métiers. Elle permet aux gestionnaires de savoir si avec le modèle choisi il n'y aura pas de perte d'argent. En effet, si on accorde un prêt à un client non solvable cela va coûter beaucoup plus cher à la société que de ne pas accorder de prêt à un client solvable.
Donc pour favoriser le faite qu'il y ai moins de perte d'argent, il est important de fixer un seuil très élevé à partir du quel on peut considérer que ce qui client est fiable. J'ai donc choisi que le seuil, c'est à dire la valeur de probabilité à partir duquel le client est considéré comme un bon client à **0.7**. Ce seuil correspond à un score de 14/20 (voir figure ci-dessous)



4. Interprétabilité du modèle

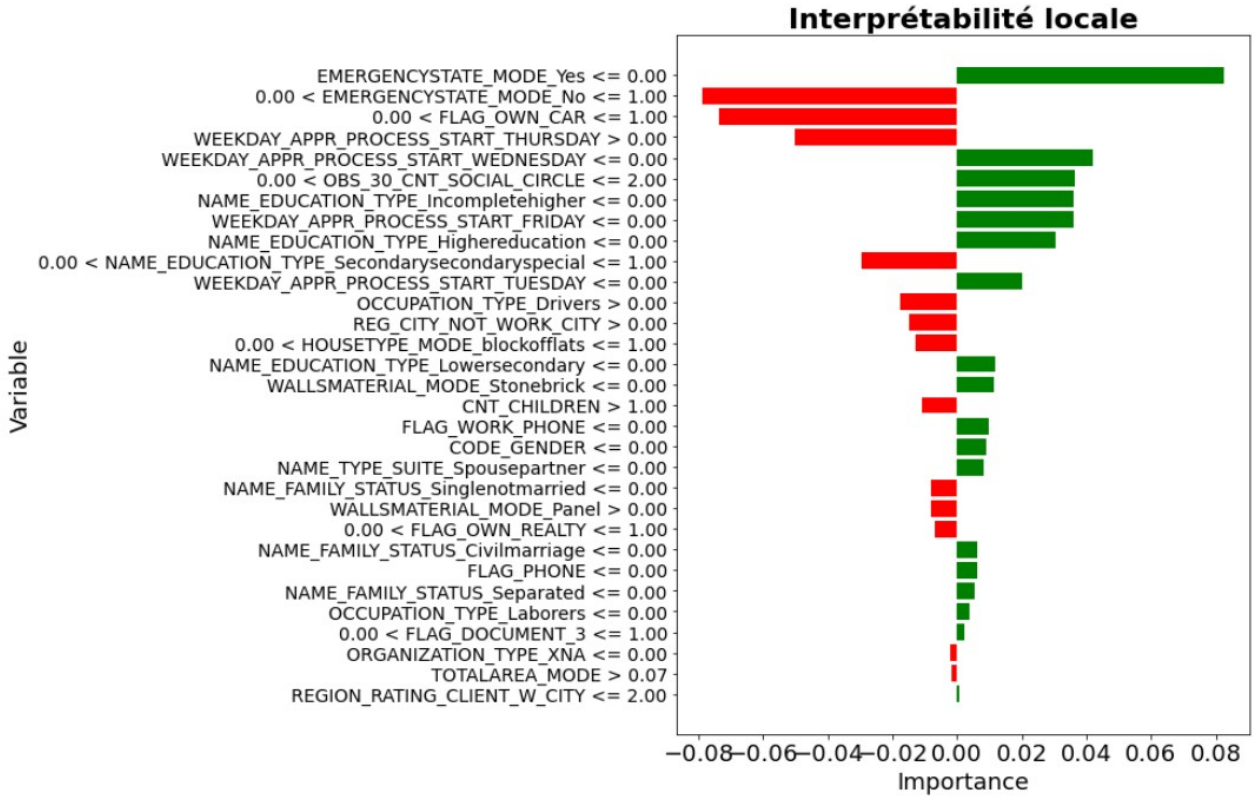
b) Interprétabilité globale



En analysant les features importances je note que les variables les plus importantes sont liés aux niveaux d'éducation le plus élevé atteint par le client, ses défauts de paiements, sa situation familiale, aux renseignements données de son lieux d'habitation et si il possède un fixe.

b) Interprétabilité locale

Pour ce travail, j'ai utilisé la méthode LIME pour interpréter localement les modèles (voir figure ci-dessous). Son implémentation se fait en deux étapes : une première étape où l'algorithme génère des nouvelles données en se basant sur le voisinage proche de l'individu à expliquer. Dans la deuxième phase, le modèle LIME entraîne un modèle plus simple sur les prédictions du modèle complexe qu'on cherche à interpréter. En vert ce sont les facteurs favorables et en rouge défavorables à la prédiction. Par exemple on voit que les informations sur le logement du client, lorsqu'elles sont véridiques, favorisent le client tandis que dans le cas contraire, elles le défavorisent. Faire des hautes études



5. Limites et améliorations possibles

Le fait d'avoir des données très déséquilibrées dès le départ (91.9 % contre 8.1%) à induit l'utilisation d'algorithme de rééquilibrage des données. Il aurai été plus intéressant d'avoir accès à plus de données qui réduisent ce déséquilibre et permettent à l'algorithme de mieux apprendre les caractéristiques des clients.
Ensuite au vu de la masse de variables, j'aurai pu envisager une réduction de données en utilisant des techniques tels que PCA.
Je peux également rajouter dans les axes d'améliorations, l'augmentation du champ de hyperparamètres lors de l'optimisation des modèles.