1- La méthodologie d'entraînement du modèle

2- La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation

3 Les limites et les améliorations possibles

1- La méthodologie d'entraînement du modèle

1- 1 Preprocessing des données

La première phase a été, comme dans la plupart des cas, la récupération des données, suivie du nettoyage.
Toutes les variables, pour lesquelles on observe plus de 40 voire 50% de valeurs nulles, ont été tout bonnement supprimées.
En ce qui concerne l' historique passée des crédits immobiliers, prêts auprès d'autres institutions financières, soldes des comptes etc …), pour chaque emprunteurs:

- les valeurs les plus fréquentes des variables catégorielles ont été établies
- la moyenne, le min, max, variance, et écart-type ont été établies pour les variables numériques

Certaines variables, comme l'âge de l'emprunteur,son niveau d'expérience professionnelle, étaient exprimées en jours, ont été converties en années, et les valeurs aberrantes corrigées.

1- 2 Feature engeneering

Plusieurs variables ont été créés, dont à titre d'exemple:

- pourcentage de l'annuité par rapport au crédit: rapport entre le montant de l'annuité et celui du crédit dû

- rapport entre le montant de l'annuité et le revenu annuel du client

- la durée du prêt: rapport entre le montant du crédit et l'annuité

- l'âge de l'emprunteur lorsque le crédit arriverait à terme

- pourcentage du prêt alloué à l'achat de biens et ou services

- revenu net du client après paiement de l'annuité: différence entre le revenu annuel et le montant de l'annuité

- revenu net par tête du foyer familial après paiement de l'annuité: rapport le nombre de personnes composant la famille entre le revenu net

1- 3 Ajustement des variables catégorielles

Toutes les variables catégorielles, dont les modalités excèdent 10, ont été supprimées afin de ne pas altérer la convergence du modèle d'apprentissage

1- 4 Sélection des variables

A titre arbitraire, on a sélectionné les 50 variables les plus corrélées à la variable cible(TARGET). Il s'est avéré, après vérification, que certaines de ses variables étaient très corrélées entre elles (exemple PRO_SENIORITY qui est le niveau d'expérience professionnelle et DAYS_WORKING_PERCENT). Afin de parer à cette multicolinéarité, plusieurs variables ont été sélectionnées au dépens d'autres. On se retrouve avec 29 variables

1- 5 PCA réduction dimensionnelle

En gardant 95% de la variance, on passe de 29 à 24 variables

1- 6 Cluster Kmeans

Grâce à la méthode de la silhouette, on identifie 5 classes.

2- La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation

Etant donnée la nature limitée de nos ressources (CPU, ram) et la taille des données, nous avons opté pour l'algorithme LightGBM, une alternative de xgboost beaucoup plus légère, dans le cadre des modèles d'apprentissage d'ensemble séquentiels.
Bien entendu, la matrice de confusion est primordiale.

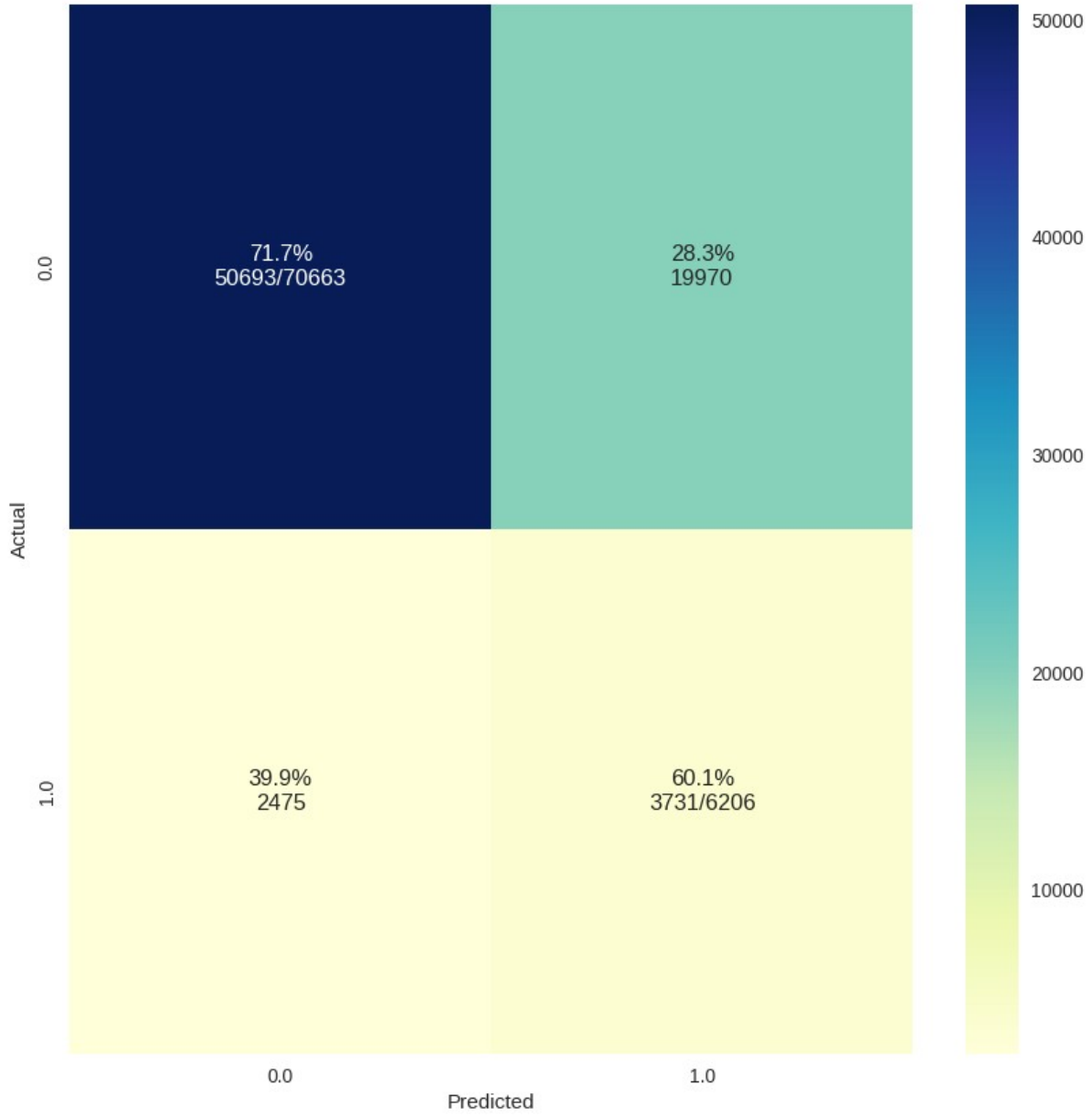
Dans notre cas , il y a un déséquilibre de classe assez significatif. En effet, dans notre échantillon, il y a 90% de clients solvables contre 10% insolvables, et par conséquent l'accuracy ou la précision globale (proportion des prévisions correctes) n'est pas pertinente dans l'évaluation de la performance.
La classe positive reste le client en défaut de paiement.
F1 score est la métrique choisie à optimiser dans le gridsearchcv de la classification, pour améliorer le rappel(recall) ou la capacité du modèle à identifier les vraies classes positives et la précision globale (accuracy).

En effet, le problème est le suivant:

- prédire un client défectueux comme solvable fait perdre de l'argent à la banque
- prédire un client solvable comme insolvable fait perdre une opportunité à la banque

Nous avons procédé à deux phases distinctes.La première a consisté à ne pas tenir compte du déséquilibre de classes, et la seconde en tentant de corriger ce déséquilibre de classes.
Pour résumer, sans prise en compte du déséquilibre de classe, le rappel est très faible, seuls 0.2% des clients défectueux sont bien classés contre 99% de clients solvables bien classés.En essayant de corriger le déséquilibre de classe grâce à l'algorithme SMOTE, en générant de nouveaux clients défectueux , le modèle arrive à bien classer 60% des clients défectueux, et près de 70% des clients solvables

Matrice de confusion



Les variables les plus influentes découlant du modèle de classification, reste le terme du crédit, les scores dans les autres institutions financières, la ville de résidence,

CREDIT_TERM	208
EXT_SOURCE_MEAN	194
BUREAU_DAYS_CREDIT_mean	185
REGION_RATING_CLIENT_W_CITY	181
NAME_INCOME_TYPE_Pensioner	156
REGION_RATING_CLIENT_W_CITY	155
PRO_SENIORITY	154
PREVIOUS_APPLICATION_CNT_PAYMENT_sum	149
ACCOUNT_SENIORITY	128
frequent-BUREAU-CREDIT ACTIVE unknown	119

3 Les limites et les améliorations possibles

- connaissance métier du risque de crédit, des KPI reconnus par les banques
- au niveau technique, optimisation des hyperparamètres que ce soit pour l'algorithme SMOTE, lightgbmc classifier
- plus de feature engineering

site web du dashboard: https://scoring.senrv.biz