**NOTE METHODOLOGIQUE**

**Méthodologie d’entrainement du modèle**

***Etape1****: Téléchargement et analyse des bases de données*

Pour notre étude nous disposons de 10 bases de données.

**Application\_train** et test contiennent principalement les informations personnelles des individus demandeurs de prêts.

**Bureau et bureau\_balance** concernent les informations sur d’autres prêts contractés auprès d’autres établissements de crédits

**Previous\_application** contient des information sur les précédents chez Home credit

Nous ne décrirons pas toutes les bases car le constat fait est qu’il y a certaines mêmes informations qui reviennent dans les bases de données.

Pendant l’analyse exploratoire la première information importante que nous avons est la distribution fortement asymétrique entre les crédits remboursés plus de 90% par rapport aux non remboursé.

***Etape2****: Entrainement du modèle*

1. Séparation de notre donnée en deux d’une part la variable cible d’autre part les variables explicatives.

Equilibrage de la variable Target avec le suréchantillonnage aléatoire naïf.

1. Utilisation de notre base en train et test (20% des données) avec le modèle train test split de sklearn.
2. Normalisation de nos matrice X (X\_train fit\_tansform(), et X\_test transform) avec StandardScaler() de sklearn.
3. Déclaration de notre modèle RandomForestClassifier(). Définition de nos paramètres, optimisation de nos paramètres pour notre modèle avec GridSearchCV().
4. Création de notre pipeline qui contiendra notre standardisation et notre modèle avec les meilleurs paramètres.

**Fonction de coût métier, algorithme d’optimisation et métrique d’évaluation**

***1-Fonction de coût***

Y ici est notre variable cible, celle que cherche à prédire

Avec y=0 crédit accordé et y=1 crédit non accordé

X est la matrice de nos variables d’entrée, celles qui, en fonction de leur dépendance par rapport à la variable cible, permettront de bien la prédire.

***2- Algorithme d’optimisation et métrique d’évaluation***

RandomForestClassifier() fait partie des méthodes ensemblistes de la famille des méthodes de moyennage dont le principe est la construction de plusieurs estimateurs indépendamment puis de moyenner leurs prédictions. Elle sera combinée avec la Methode du GridSearchCV de la bibliothèque sklearn

Le coefficient de détermination R² est la première métrique utilisée pour voir la qualité de précision du modèle. Ensuite l’accuracy qui calcule la précision des prédictions correctes

**Interprétabilité globale et locale du modèle**

1. La précision globale d’un modèle consiste à la pertinence des résultats qu’il obtient en fonction des objectifs de départ. Dans notre cas en fonction des données dont nous disposons, du prétraitement que nous en avions fait, répond-il aux attentes ? Pour cela nous verrons principalement les résultats des métriques d’évaluation du modèle choisis : le coefficient de détermination, accuracy.
2. Pour ce qui est de la précision locale il s’agira de voir qu’elle a été l’influence des variables d’entrées sur la variable cible. Il a été question dans notre cas de faire une liste des variables les plus importantes (leurs coefficients) ayant un impact sur les variables cibles.

Limites et les améliorations possibles

La qualité de la prédiction n’est pas très bonne. Plusieurs améliorations peuvent être apportées : une nouvelle exploration de données afin de pouvoir déjà détecter des variables qui nous sembleraient pertinentes pour l’étude, regarder d’autres méthodes d’échantillonnages pour l’équilibre des données, utiliser d’autres algorithmes de prédiction pour voir s’ils obtiennent de meilleures performances. Et enfin faire une exploration métier pour voir s’il y a des informations latentes qu’on peut ressortir pour aider le modèle dans sa performance.