Note méthodologique

La méthodologie d'entraînement du modèle (2 pages maximum)

Etude du jeu de données:

- Pour chacun des dix dataframes fournis réalisation d'une **analyse pré-exploratoire** avec l'aide du kernel Kaggle: *Home Credit Default Risk - WILL KOEHRSEN · 4Y AGO - Start Here: A Gentle Introduction*), d'un **nettoyage des données** et du **calcul des agrégats** pertinents.
- Réalisation d'un **merge** de l'ensemble des dataframes et des agrégats calculés en un dataframe unique avec l'aide du kernel Kaggle: *Home Credit Default Risk - AGUIAR 3Y AGO - LightGBM with Simple Features*).

Création d'un pipeline pour l'estimateur:

- **Encodage** des données catégorielles (avec scklearn OneHotEncoder).
- **Redimensionnement** des données (avec scklearn StandardScaler/ MinMaxScaler/RobustScaler).
- **Rééquilibrage des données** (avec la librairie Imblearn: oversampling: SMOTE, undersampling: RandomUnderSampler).
- Ajout d'un **algorithme de classification**.

Rééquilibrage des données:

- Les labels binaires de la cible du jeu de données se répartissent comme suit: 92% de 0, 8% de 1.
- Les labels du jeu d'entrainement sont rééquilibrés à égales proportions afin de favoriser l'apprentissage de l'algorithme pour la détection les cas positifs.
- Le jeu de test n'est pas rééquilibré pour correspondre à la réalité.

Choix de quatre classifieurs:

- DummyClassifier (approche naïve/baseline).
- LinearRegression (algorithme usuel).
- XGBoostClassifier (algorithme rapide et performant).
- XGBoostRandomForestClassifier (algorithme rapide et performant).

La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation (1 page maximum)

Utilisation de la libraire **hyperopt** avec minimisation de la fonction de perte **fbeta_score** de scklearn.

Une matrice de confusion avec **β>1** suppose une part accordée au recall plus importante que celle accordée à la précision.
Or nous cherchons ici à minimiser la part des faux négatifs i.e la part des clients qui ne peuvent pas rembourser leur prêt mais qui se sont vus malgré tout accorder un prêt.
Une précision nous permet de mesurer la robustesse du modèle aux faux positifs i.e les clients qui se sont vus refuser un prêt mais qui étaient capables de le rembourser. C'est une perte de clients et une perte de bénéfices à moyen terme (intérêts composés < à 10% du montant du prêt) pour l'entreprise mais une perte d'argent négligeable devant la perte nette d'un prêt non remboursé qui peut aller jusqu'à correspondre à l'intégralité du montant.

Choix de **β=10**.

L'interprétabilité globale et locale du modèle (1 page maximum)

Analyse des feature Importance:

- Utilisation des librairies **LIME** et **SHAP** d'analyse de la contribution des features dans la prédiction du modèle.
    - LIME: Analyse par échantillon et rang.
    - SHAP: Analyse par échantillon et rang. Les visualisation suivantes ont été effectuées: **force, summary, embedding, dependence, decision, bar, waterfall.**

Les limites et les améliorations possibles (1 page maximum)

- Etudier d'autres algorithmes de classification bien que ce ne soit pas l'objectif du projet de tous les étudier.
- Données déséquilibrées. Faible % de crédits refusés.
- Custom metric à revoir avec une équipe métier pour affiner le coefficient de pondération.
- Avoir plus de données et mieux réparties.
- Choix plus fin des variables. Calcul de nouveaux aggrégats.