Louis Maillet IR4

Projet Final : Data Engineering & Credit Scoring

1. Contexte du Projet et Objectifs

Ce projet a pour but de prédire la capacité d'un demandeur de prêt à rembourser son emprunt (cible binaire TARGET) en utilisant l'algorithme LightGBM. L'objectif principal était de mettre en œuvre un pipeline MLOps complet intégrant le suivi d'expérimentation via MLflow et l'optimisation des hyperparamètres via Optuna.

La métrique de performance utilisée est l'AUC (Area Under the ROC Curve).

2. Phase 1 : Modélisation Baseline (Partie 2)

La première étape a consisté à entraîner un modèle LightGBM avec des hyperparamètres par défaut sur les données brutes (application_train.csv) après un nettoyage minimal (gestion des types et des valeurs catégorielles).

- Stratégie : Entraînement sur Train/Validation Split (80/20) avec *early stopping* (50 rounds).
- Plateforme : La run a été tracée dans MLflow sous le nom lightgbm_baseline.

Résultat Baseline

Métrique	Score
Validation AUC (Local)	0.7590

Ce score a servi de référence à battre lors de la phase d'optimisation.

3. Phase 2 : Optimisation des Hyperparamètres (Partie 3 / Étape 4.2)

Pour améliorer la performance, une optimisation a été menée en utilisant le framework Optuna pour explorer l'espace des hyperparamètres de LightGBM.

- Stratégie : 50 essais (trials) d'Optuna, chacun utilisant le même split Train/Validation que la baseline.
- MLOps : Chaque essai Optuna a été enregistré comme une nested run (run imbriquée) sous un run principal MLflow (LGBM_Optimization_Optuna), assurant un traçage complet de l'espace de recherche.

Résultat de l'Optimisation

Le meilleur score a été obtenu lors du Trial 2.

Métrique	Score	Amélioration
Meilleure Validation AUC (Local)	0.7614	+0.0024 points d'AUC

Meilleurs Hyperparamètres (Trial 2)

Le jeu de paramètres optimal, enregistré par MLflow, est le suivant :

Paramètre	Valeur Optimale
learning_rate	0.0957
num_leaves	33
max_depth	3 (Indiquant une préférence pour un modèle peu profond et régularisé)
min_child_samples	82
reg_alpha (L1)	$4.53 \times 10^{-5}
reg_lambda (L2)	$1.35 \times 10^{-7}

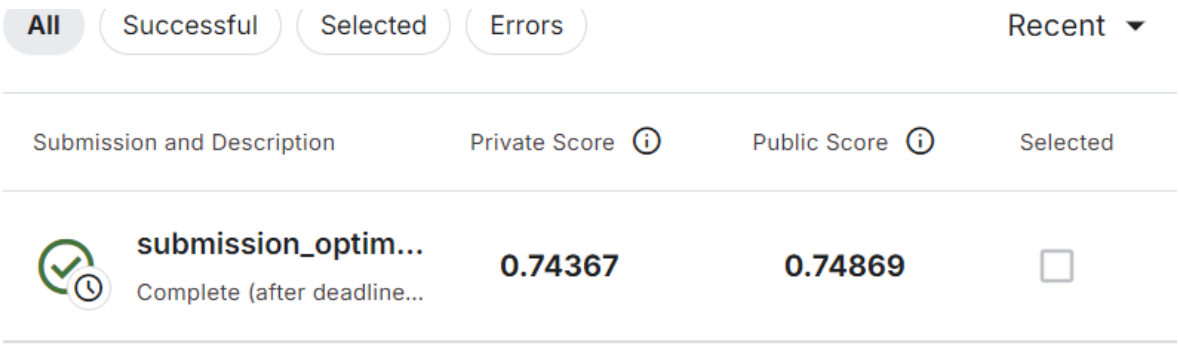
4. Phase 3 : Entraînement Final et Soumission (Partie 4 / Étape 4.3)

Le modèle final a été entraîné en utilisant l'ensemble des données d'entraînement (application_train.csv) et les meilleurs hyperparamètres identifiés par Optuna.

- Entraînement : Effectué sur la totalité de X_train_full (sans split de validation).
- Traçage : Enregistré dans MLflow sous la run final_optimized_model.
- Artefacts : Le modèle a été enregistré dans le MLflow Model Registry et le fichier de soumission CSV a été loggué comme artefact.

Score Officiel Kaggle

Le fichier de soumission (submission_optimized_auc_0.7614_...csv) a été téléversé sur la plateforme Kaggle.



Analyse de l'Écart de Score

L'écart entre l'AUC locale (0.7614) et l'AUC officielle de Kaggle (0.74869) est attendu pour un modèle entraîné sur un simple *holdout split* :

- Le score local est souvent optimiste car le modèle est évalué sur des données issues de la même distribution.
- Le score Kaggle révèle une légère sur-spécialisation (overfitting) au jeu de validation local.
- Conclusion : Le score officiel Kaggle de 0.74869 est le véritable indicateur de performance.

5. Conclusion et Perspectives

Ce projet a permis de construire et de tracer un pipeline de modélisation complet, de la baseline à l'optimisation, en utilisant des outils MLOps standards. L'intégration de MLflow a garanti la transparence et la reproductibilité des résultats (Baseline vs. Optimisation).

Perspectives d'Amélioration (pour atteindre un score > 0.80) :

1. Feature Engineering Avancé : Intégrer les informations des autres tables de données fournies (BUREAU, POS_CASH_BALANCE, CREDIT_CARD_BALANCE) via des agrégations.
2. Stratégie de Validation : Remplacer le simple *split* par une Validation Croisée Stratifiée (K-Fold) lors des phases de modélisation et d'optimisation pour obtenir un score local plus robuste et plus proche du score Kaggle.