Rapport Final

Credit Scoring

Contexte et objectifs métier

Dans le cadre du développement de l'activité crédit de l'entreprise « Home Credit », notre mission consistait à élaborer un algorithme capable de prédire la probabilité de faillite d'un client au remboursement de son prêt.

L'enjeu n'est pas seulement technique mais économique. Le modèle doit répondre à une contrainte métier forte : l'asymétrie du coût de l'erreur.

- Un Faux Négatif (FN) entraîne une perte de capital estimée à 10 fois le coût d'un simple manque à gagner.
- En Faux Positif (FP) représente un coût d'opportunité standard.

L'objectif final est de livrer une API docker fonctionnelle, exposant le modèle minimisant ce coût métier total

Démarche de préparation des données

La performance du modèle repose sur une préparation rigoureuse des données brutes issues de multiples sources relationnelles.

Nettoyage et Feature Engineering

Nous avons traité le jeu de données principal (application_train.csv) en corrigeant les anomalies (trou dans le dataset et valeurs aberrante. Pour enrichir l'information, nous avons créé des variables « métier » :

- Ratios Financiers : Taux d'endettement et Ratio Crédit/Revenu, souvent plus prédictifs que les montants bruts.
- Synthèse de Scores : Création d'une moyenne des scores externes

Agrégation des données relationnelles

Afin de capter l'historique bancaire complet, nous avons fusionné les tables périphériques:

- Historique externe (bureau.csv) : agrégation des crédits contractés dans d'autres banques

- Historique interne : Synthèse des demandes passées chez Home Credit

Gestion des biais et data leakage

Nous avons particulièrement fait attention à la séparation des données. L'imputation des valeurs manquantes avec la médiane et l'encodage ont été calibrés uniquement sur le jeu d'entraînement pour éviter toute fuite d'information vers le jeu de test.

Stratégie de modélisation et comparaison

Nous avons mis en place une approche comparative rigoureuse basée sur une validation croisée stratifiée pour garantir la robustesse des résultats face au déséquilibre des classes (seulement 8% de mauvais client).

Algorithmes Testé

Trois familles d'algorithmes ont été évaluées :

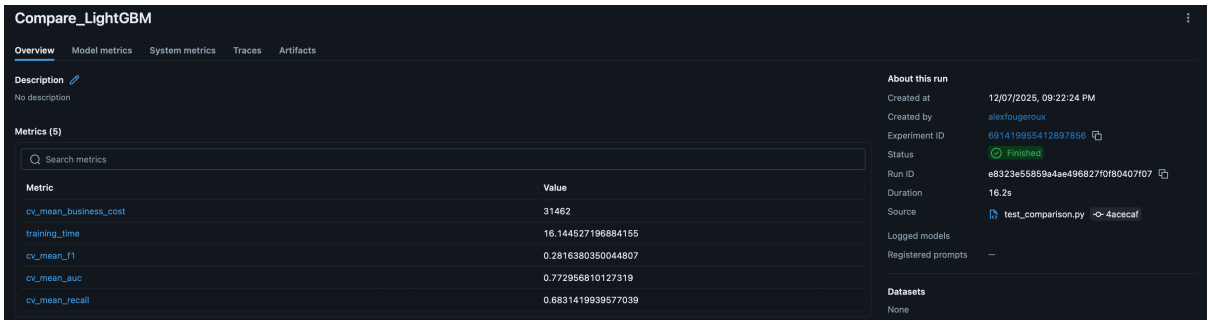
- Linéaire : Régression logistique

- Random Forest

- Boosting : LightGM et XGBoost

Tracking MLOps avec MLflow

Chaque expérimentation a été tracée via MLflow, enregistrant systématiquement les hyperparamètres, l'AUC, et le modèle sérialisé.



Résultats principaux et optimisation

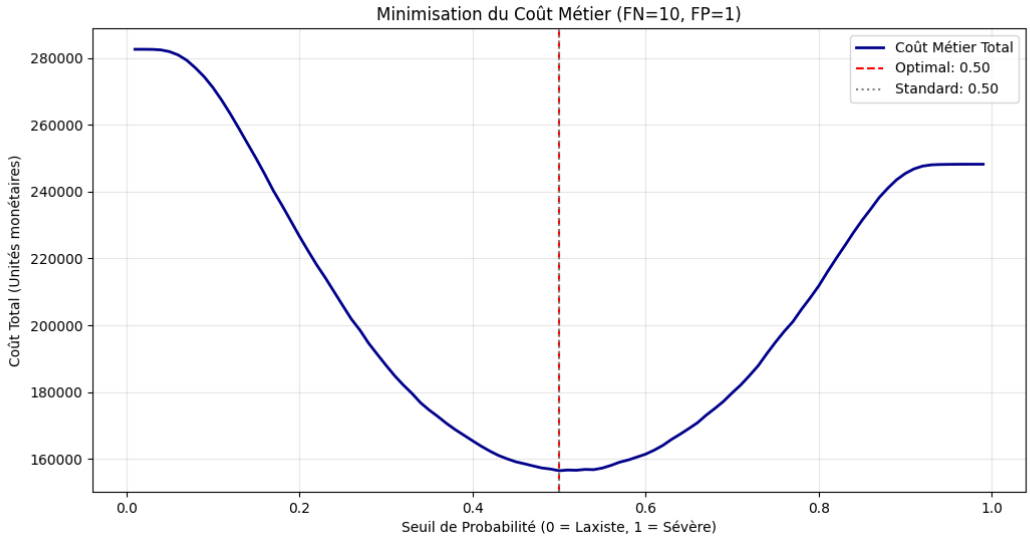
Sélection du modèle « gagnant »

Le modèle LightGBM s'est imposé comme le plus performant, offrant le meilleur compromis entre l'AUC et Recall.

- AUC Moyen (Cross-validation) : 0.77 (sur échantillon) / 0.71 (dataset complet)
- Gestion du déséquilibre : L'utilisation du paramètre class_weith='balanced' a permis de forcer le modèle à pénaliser les erreurs sur la classe minoritaire.

Optimisation du seuil métier

Nous avons défini une fonction de coût personnalisée: Cost = 10xFN + 1xFP. En analysant la courbe du coût en fonction du seuil de décision, nous avons observé que grâce à l'équilibrage interne du modèle (balanced), le seuil optimal se situe autour de 0.48.

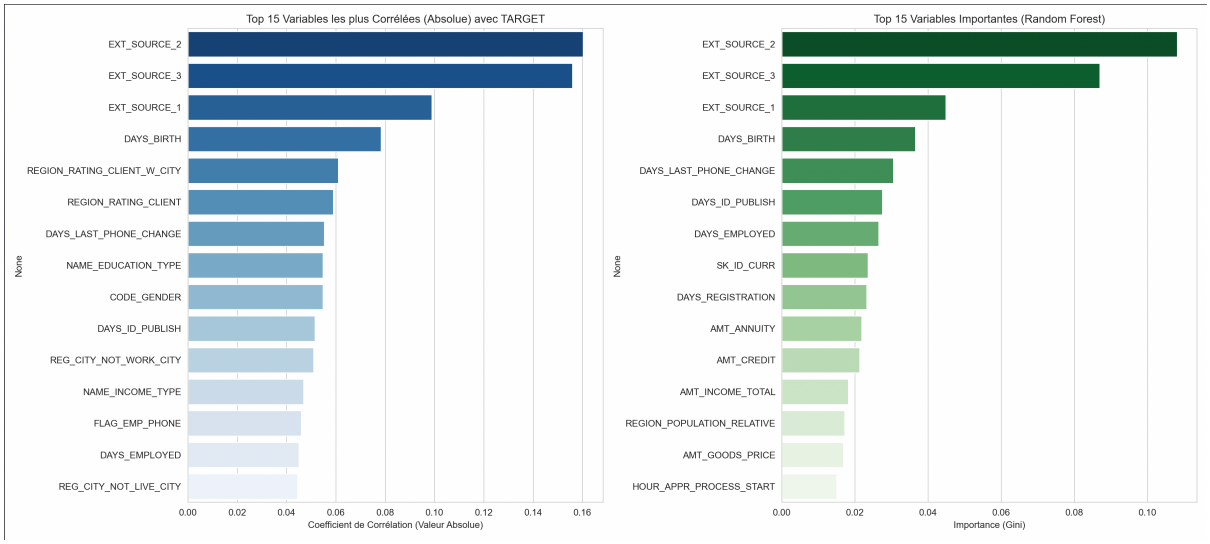


Minimisation du coût métier. Le seuil optimal (rouge) permet de réduire le risque financier par rapport à une approche standard.

- Seuil retenu : 0.48
- Interprétation : Si la probabilité de défaut dépasse 48%, le crédit est refusé. Ce seuil permet de sécuriser la banque contre les défauts coûteux tout en maintenant un volume d'affaires acceptable.

Interprétabilité du modèle

Pour garantir la transparence des décisions, nous avons analysé les variables les plus influentes du modèle.



Les facteurs déterminants pour le scoring sont:

1. EXT_SOURCE(1,2,3) : Les scores normalisés provenant d'agences de crédit externes sont les prédicteurs les plus fiables.
2. PAYMENT_RATE = Le poids de l'annuité par rapport au crédit total. Un taux élevé corrèle souvent avec un risque accru.
3. DAYS_BIRTH : Les clients plus jeunes présentent statistiquement un risque de défaut plus élevé que les clients seniors.
4. DAYS_EMPLOYED : La stabilité de l'emploi joue un rôle clé dans la capacité de remboursement.

Conclusion:

Le projet a permis de délivrer un modèle LightGBM, capable de discriminer efficacement les bons des mauvais payeurs. L'approche MLOps mise en oeuvre (Tracking MLflow, Pipeline automatisé, Docker) assure que la solution est non seulement performante, mais aussi reproductible et prête pour un déploiement industriel via l'API développée.