

ESAIP

# Rapport Data Engineering & Analysis

Majeur IA



École d'Ingénieurs La<sup>★</sup>Salle

Mathis Herbreteau - Mathias Le Pottier - Samuel Pasquier  
12/12/2025

## Table des matières

Contexte et objectifs .....	2
Préparation des données et méthodologie de construction du dataset.....	2
Modélisation et métrique métier .....	3
Résultats obtenus .....	3
Conclusion .....	4
Annexe .....	4

## Contexte et objectifs

Ce projet s'inscrit dans le cadre d'une mission pour une société financière spécialisée dans les crédits à la consommation pour des clients ayant peu ou pas d'historique de prêt. L'objectif principal est de développer un outil complet de Credit Scoring capable de prédire automatiquement la probabilité de défaut d'un client et de classifier chaque demande en crédit accordé ou refusé.

## Préparation des données et méthodologie de construction du dataset

Pour constituer notre jeu de données, nous avons exploité huit tables relationnelles (données applicatives, historiques des bureaux de crédit, soldes, etc.), fusionnées selon la stratégie suivante :

1. **Prétraitement et Nettoyage** : Nous avons traité les anomalies, telles que des durées d'emploi aberrantes, en les remplaçant par des valeurs nulles (*NaN*).
2. **Ingénierie des fonctionnalités (Feature Engineering)** : Afin de synthétiser l'information financière, quatre indicateurs clés ont été créés :
  - CREDIT\_INCOME\_PERCENT : ratio crédit/revenu.
  - ANNUITY\_INCOME\_PERCENT : ratio annuité/revenu.
  - CREDIT\_TERM : ratio annuité/credit.
  - DAYS\_EMPLOYED\_PERCENT : pourcentage de jours travaillés.
3. **Encodage des variables catégorielles** : Les variables catégorielles ont ensuite été encodées (Label Encoding pour les binaires, One-Hot Encoding pour les autres).
4. **Stratégie d'agrégation** : Compte tenu des relations "un-à-plusieurs" (ex: un client possède plusieurs crédits passés), nous avons appliqué une agrégation en cascade. Les tables de détails (comme *installments\_payments*) ont d'abord été groupées par identifiant de prêt (*SK\_ID\_PREV*) pour extraire des statistiques (moyenne, somme, min, max). Ces résultats ont ensuite été regroupés par client (*SK\_ID\_CURR*) pour obtenir des indicateurs globaux.

### Vérification et Gestion des données

Après la fusion, l'intégrité des ensembles d'entraînement (Train) et de test (Test) a été validée pour assurer la cohérence des colonnes. Deux points critiques ont été adressés par la suite :

- **Déséquilibre des classes** : La cible est fortement déséquilibrée (92 % de clients sans défaut de paiement contre 8 % en défaut). Ce biais a été traité par une pondération des classes (`class_weight='balanced'`) lors de l'entraînement.
- **Valeurs manquantes** : Aucune suppression massive n'a été effectuée afin de préserver l'information. Les modèles basés sur des arbres de décision (LightGBM, XGBoost) gérant nativement ces absences, nous avons choisi de conserver ces données en l'état.

## Modélisation et métrique métier

**Définition de la métrique métier :** Afin de disposer d'une évaluation pertinente, nous avons élaboré une fonction de coût personnalisée alignée sur les enjeux business. Celle-ci pénalise dix fois plus un "Faux Négatif" (octroi d'un crédit à défaut, entraînant une perte de capital) qu'un "Faux Positif" (refus d'un client solvable, représentant un manque à gagner).

Cette stratégie vise à minimiser le risque financier global, acceptant une sélectivité accrue au détriment du volume de crédits accordés.

**Choix des modèles et Stratégie d'entraînement** Deux approches ont été retenues :

1. **Modèle de référence (Baseline)** : Une Régression Logistique a été établie pour fixer un seuil de performance minimal. Ce modèle ne gérant pas les valeurs manquantes nativement, il a été intégré dans un pipeline incluant une étape d'imputation par la médiane.
2. **Modèles avancés (Gradient Boosting)** : Nous avons sélectionné LightGBM et XGBoost. Si XGBoost est réputé pour sa robustesse et sa précision, LightGBM se distingue par sa rapidité d'exécution et son efficacité mémoire sur de grands volumes de données.

**Optimisation des Hyperparamètres** Pour chaque modèle avancé, nous avons comparé une version avec les paramètres par défaut et une version optimisée. L'optimisation de LightGBM a été réalisée via une recherche bayésienne avec le framework Optuna. Après 50 itérations (*trials*), la configuration optimale retenue est la suivante :

## Résultats obtenus

Performance du meilleure Modèle (LightGBM) Le modèle final sur le jeu de validation affiche les performances suivantes :

Un ROC AUC de 0.786, ce qui nous indique que le modèle a une bonne capacité à discriminer les instances positives et négatives, notamment en prenant en compte les faux positifs et les faux négatifs ainsi qu'un seuil de 0.51.

Nous avons obtenu ce seuil après avoir analysé le score métier obtenu pour tous les seuils entre 0 et 1 avec un pas de 0.01. On retrouve ~0.5 car nos classes sont balanced.

Pour la métrique f1\_score nous obtenons 0.3, ce résultat montre que notre modèle a tendance ne pas accorder de crédit même s'il le pouvait car il ne veut pas prendre de risque du au fort cout d'un non remboursement.

On a également un recall de 0.68 donc notre modèle reconnaît 68% des bons clients.

La dernière métrique importante est la précision pour laquelle on obtient 0.19 ce qui est faible. La signification est que sur tous les mauvais payeurs identifier seulement 19% sont réellement des mauvais payeurs.

## Conclusion

Ce projet a permis de développer un outil de Credit Scoring robuste, reposant sur l'algorithme LightGBM optimisé. Avec une ROC AUC de 0.786, le modèle démontre une capacité satisfaisante à discriminer les clients solvables des profils à risque.

L'alignement avec la stratégie financière de l'entreprise est respecté : en pénalisant fortement les défauts de paiement (coût x10), le modèle adopte un comportement prudent.

- Sécurité du capital : Le seuil de décision de 0.51 (cohérent avec l'équilibrage des classes) permet de sécuriser les actifs en évitant les profils douteux.
- Le compromis Risque/Volume : Cette prudence se traduit par un Recall de 0.68 (bonne détection des cibles) mais une Précision de 0.19. Cela indique que pour éviter les impayés, l'entreprise accepte de refuser un certain nombre de clients potentiellement solvables (manque à gagner) afin de garantir un taux de défaut minimal.

En somme, l'outil déployé répond efficacement à l'objectif prioritaire de minimisation du risque financier, fournissant une aide à la décision fiable pour l'octroi de crédits.

## Annexe

Metric	Value	Module
cr_mean_auc	0.786025542999705	model
cr_mean_business_cost	23544.4	model
cr_mean_f1	0.2999019629997127	model
cr_mean_precision	0.1933714624281403	model
cr_mean_recall	0.68335042740966	model
cr_std_auc	0.00259805486179011	model
cr_std_business_cost	175.8791946489817	model
cr_std_f1	0.002057516005273599	model
cr_std_precision	0.0016474995144214976	model
cr_std_recall	0.004152791564687023	model

Parameter	Value
booster_type	gbdt
class_weight	None
categorical_hybrid	0.01
importance_type	split
is_balanced	True
learning_rate	0.0146
max_depth	11
min_child_samples	20
min_child_weight	0.001
min_split_gain	0.0
num_leaves	51
n_estimators	536
n_jobs	3
objective	None