

# Rapport de Synthèse du Projet : Outil de Credit Scoring Optimisé par le Coût Métier

Aubin Hérault, Gael Le Reun, Thomas Bertho

## Introduction et Objectif Stratégique

Ce projet visait à établir un outil de credit scoring transparent et robuste. L'objectif principal n'était pas lié aux métriques habituelles, mais à donner une décision strict d'accord de crédit ou non. Pour cela, la métrique clé choisie pour l'optimisation et l'évaluation n'était pas l'AUC (Area Under the Curve), mais le Coût Métier. Cette fonction objective a été paramétrée pour refléter l'asymétrie des risques en pénalisant dix fois plus la faute la plus coûteuse, à savoir l'octroi d'un crédit à un client qui fera défaut (Faux Négatif), que le rejet d'un client solvable (Faux Positif).

## Préparation des Données et Feature Engineering

Nous avons commencé par une exploration approfondie et la création de deux jeux de données distincts pour l'entraînement. Les 2 versions ont le même feature engineering, enrichi par des agrégations provenant de multiples sources (les différentes tables, les kernels Kaggle, etc). Pour la v1 nous avons gardé toutes les colonnes y compris des colonnes avec un taux de valeurs manquantes supérieur à 50%. À l'inverse, la version V2 a subi une suppression drastique de nombreuses colonnes jugées inutiles ou vides, comme celles relatives à l'habitat ou la soumission de documents.

Dans la v2 le traitement des variables catégorielles a été effectué de manière différenciée. L'encodage par One-Hot Encoding a été appliqué aux variables nominales, tandis que les variables ordinaires ont été transformées par Label Encoding, pour optimiser le modèle. Pour la v1 par contre les données ont été entièrement encodées avec du One-Hot Encoding.

Les données ont ensuite été séparées en ensembles Train, Validation et Test via une stratification, garantissant que la proportion de la classe cible (le défaut) soit identique dans chaque ensemble.

## Entraînement et Sélection du Modèle

Nous avons commencé par un entraînement initial sur les deux datasets. Ce test a rapidement identifié le modèle LightGBM comme l'algorithme le plus prometteur, surpassant largement les modèles de référence tels que le Dummy Classifier, Random Forest, et XGBoost, grâce à sa rapidité d'exécution et ses meilleurs résultats en prédiction.

Le LightGBM s'est distingué avec un score AUC de 0.788, le plaçant dans le haut des performances pour ce type de problématique. Plus important encore, sa métrique de Coût Métier s'est établie à environ 21 200, ce qui représente une amélioration significative par rapport aux autres modèles (environ 23 000) et surtout par rapport au modèle naïf (Dummy) qui atteignait 35 000. L'entraînement a révélé que les performances étaient similaires entre la V1 et la V2 ; nous avons donc conservé le modèle LGBM pour la suite de l'optimisation.

## Optimisation et Construction du Modèle Final

Pour affiner le modèle, nous avons concentré nos efforts sur l'optimisation des hyperparamètres du LightGBM en utilisant l'outil Optuna. L'objectif unique de cette recherche avancée était de minimiser le Coût Métier sur l'ensemble de Validation, s'assurant que le modèle converge vers la solution la plus rentable.

Nous avions initialement prévu d'utiliser la Validation Croisée Stratifiée avec 5 folds pour construire un modèle d'Ensemble plus robuste et généralisable. Cependant, les résultats obtenus avec l'Ensemble de CV se sont avérés moins performants que le modèle LightGBM unique et optimisé. Cela est probablement dû à un manque de données pour une être efficace, nous sommes donc rester sur notre meilleur modèle LightGBM unique avec ses paramètres optimisés. Ce modèle constitue la base de la solution final.

## Interprétation et Explicabilité (SHAP)

L'interprétation de l'influence des variables a été menée grâce à l'analyse SHAP. Cette démarche assure la transparence nécessaire à l'outil de credit scoring. L'analyse révèle que les facteurs déterminants pour la prédiction sur l'ensemble des clients, sont orientés vers les sources d'argent et le risque.

Trois variables se distinguent : d'abord, les Scores Externes (EXT\_SOURCE\_X), qui proviennent de sources d'information externes et sont de loin les variables les plus importantes. Un score faible augmente significativement la probabilité de défaut. Ensuite, l'Âge du Client (DAYS\_BIRTH) montre que les clients plus jeunes sont généralement associés à un risque accru. Enfin, le Montant de l'Annuité (AMT\_ANNUITY) qui est un indicateur direct de la charge de remboursement mensuelle et de la capacité du client à rembourser ses crédits.

Ces variables confirment que le modèle prend ses décisions en fonction d'indicateurs de risque classiques, tout en utilisant le Coût Métier pour optimiser le rendement financier, sans prendre en compte l'AUC comme les autres modèles sur Kaggle.

## Conclusion et Déploiement

La traçabilité de tout le processus est assurée par MLflow. Chaque run, chaque jeu de paramètres et chaque métrique sont consignés. Pour le déploiement nous avons réussi à créer un Docker et les tests API du notebook 04 ont montré que tout fonctionnait. Le modèle est bien capable de recevoir des données et de retourner des prédictions (Status Code 200) en temps réel.

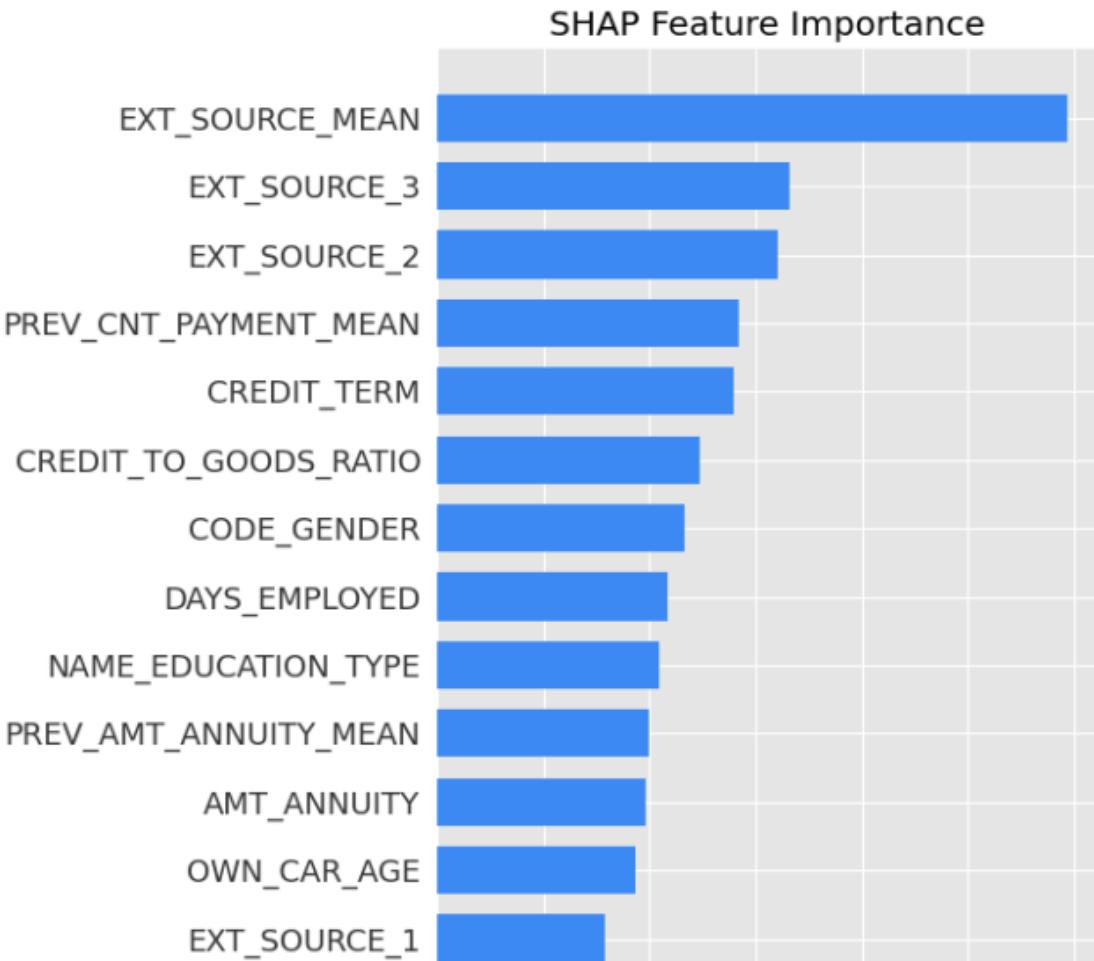


Figure pour représenter l'importance global des meilleures variables dans notre modèle final.