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

Aubin Hérault, Gael Le Reun, Thomas Bertho

Introduction et Objectif Stratégique

Ce projet vise à établir un outil de crédit scoring robuste et transparent pour une institution financière. L'objectif principal était d'aligner la décision d'octroi de crédit sur la réalité économique de l'entreprise. Pour cela, la métrique clé choisie pour l'optimisation et l'évaluation n'est pas l'AUC (Area Under the Curve), mais le Coût Métier. Cette métrique a été paramétrée pour refléter l'asymétrie des risques, 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).

Preprocessing et traitement des données

La démarche a commencé par l'exploration et la préparation des données, en créant deux jeux de données distincts. La version V1 et v2 ont eu un feature engineering intensif à partir de multiples sources (historique de prêts externes, paiements par versements, etc.), souvent depuis les kernels kaggles.

La v1 a gardé toute les colonnes même celles avec plus de 50% de missing values et les variables catégorielles ont toutes été encodé par One-Hot Encoding.

La v2 quant à elle a subi une suppression de nombreuses colonnes comme celles en rapport avec l'habitat et celles qui disaient si un des 20 documents avait été envoyé qui nous paraissaient inutile car vide a 90% du temps. L'encodage des variables catégorielles a été fait en One-Hot Encoding ou en Label Encoding en fonction du type de variable : ordinal ou nominal.

Entrainement et choix des modèles

Nous avons commencé par un benchmark initial qui a rapidement identifié le modèle LightGBM comme l'algorithme le plus prometteur, surpassant les modèles de référence (Dummy, Random Forest, XGBoost) grâce à sa rapidité et à ces meilleurs résultat de prédiction.

Au niveau des résultats les modèles avaient presque les mêmes sur les 2 versions des dataset. Le meilleurs modèle a été LightGBM avec un score AUC de 0,788 ce qui nous place proche des meilleurs code kaggle qui ont 0,81 au maximum. La métrique métier quant à elle est d'environ 21 200 au mieux en comparaison un modèle dummy est à 35 000 et les autre modeles a 23 000.

Optimisation et Construction du Modèle de Production

Pour améliorer notre modèle nous avons utilisé différents moyen. Pour commencer nous voulions améliorer les hyperparamètres du LightGBM en utilisant Optuna, un outil de recherche avancée. L'objectif unique de cette optimisation était de minimiser le Coût Métier sur l'ensemble de Validation, garantissant ainsi que le modèle apprenne à réduire le risque financier.

Ensuite nous voulions utiliser la validation croisé avec 5 fold pour avoir un ensemble de modeles qui serait plus robuste mais les résultat étaient moins bon surement a cause du manque de données.

Nous sommes donc restés sur notre meilleur modèle avec ses paramètre optimisé.

Interprétation et explicabilité

L'interprétation de l'influence des variables a été réalisée grâce à l'analyse SHAP. Cette analyse révèle que les facteurs déterminants pour la prédiction de défaut, en moyenne sur l'ensemble des clients, sont :

1. Le Score Externe (EXT_SOURCE_X) : Ces scores de risque provenant de sources d'information externes sont, de loin, les variables les plus importantes. Un score faible augmente significativement la probabilité de défaut.
2. L'Âge du Client (DAYS_BIRTH) : Les clients plus jeunes sont, de manière générale, perçus comme présentant un risque accru.
3. Le Montant de l'Annuité (AMT_ANNUITY) : La charge de remboursement mensuelle est un indicateur direct de la capacité du client à honorer ses engagements.

Ces variables montrent que le modèle se base sur des indicateurs de risque classiques, mais les pondère de manière à optimiser le rendement financier, s'éloignant des jugements purement basés sur l'AUC.

Conclusion et Déploiement

La robustesse et la traçabilité du projet sont assurées par l'utilisation de MLflow. Chaque étape d'entraînement, chaque paramètre et chaque score sont consignés dans un Run, permettant une reproductibilité totale. Le modèle final est enregistré dans le Model Registry de MLflow. L'étape ultime du projet consiste à le rendre opérationnel via MLflow Serving, permettant au service de crédit d'interroger directement le modèle via une API REST rapide et fiable pour l'approbation instantanée des dossiers mais non n’avons pas réussi à le mettre en place.

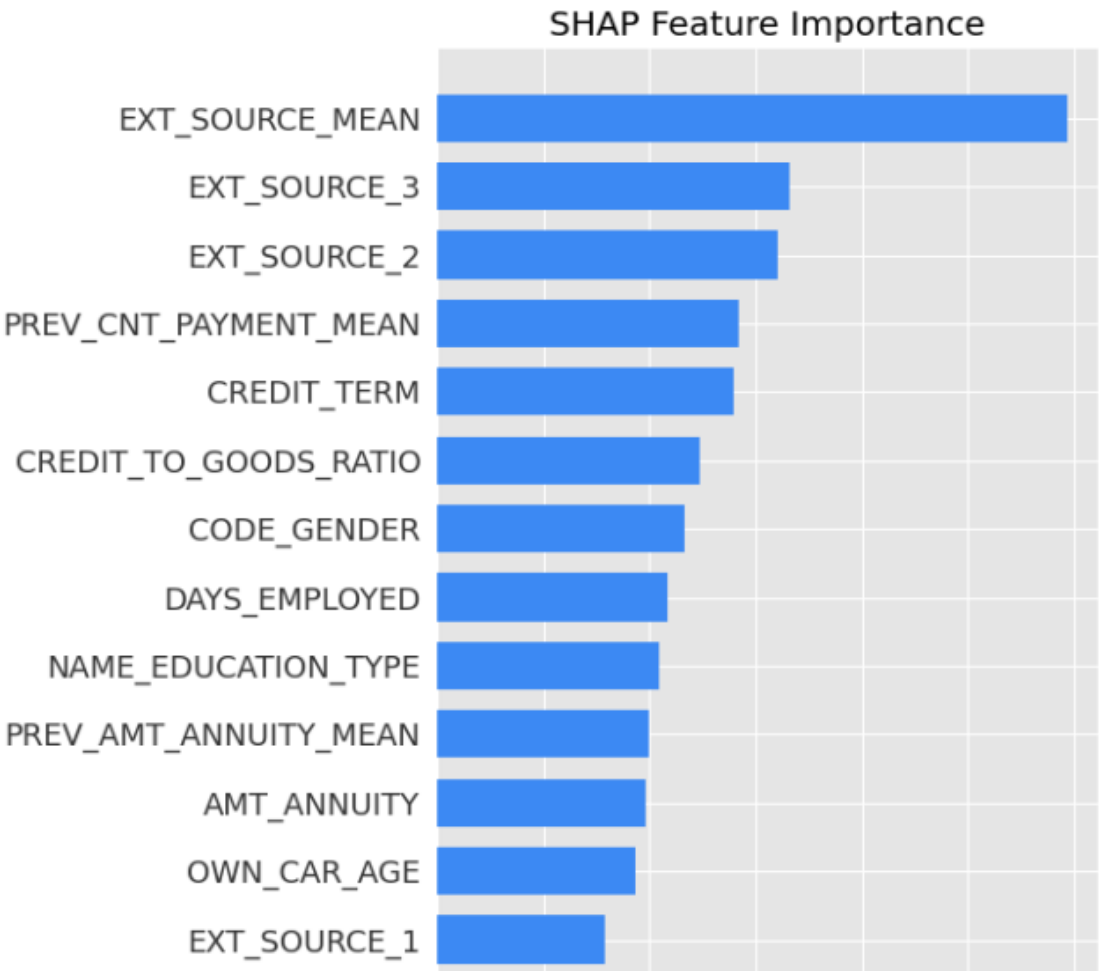

Figure pour représenter l'importance global des meilleurs variable dans notre modèle final.