EXPOSÉ PROJET 4

CONSTRUIRE UN MODÈLE DE SCORING

Le 29 Octobre 2020

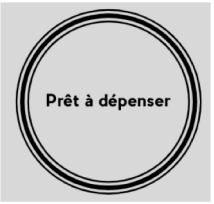
Zeineb Guizani

Plan de la présentation

1. Présentation de l'appel à projet
2. Description du jeu de données
3. Transformation du jeu de données
4. Comparaison et synthèse des résultats pour les modèles utilisés
5. Interprétabilité du modèle (Feature Importance / Lime)
6. Conclusion & perspectives

1- Présentation de l'appel à projet

- La société financière, « Prêt à dépenser » souhaite prédire la capacité de ses clients à rembourser un prêt.



- Une classification binaire entre 0 et 1.

- Développer un algorithme de scoring pour aider à décider si :

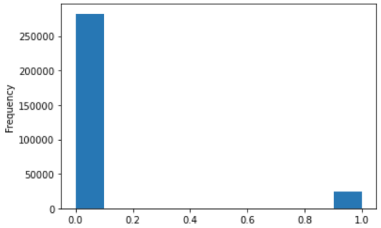
  le client peut rembourser (0) ou non (1) son prêt?

- Jeu de données "Home Credit" disponible sur Kaggle.
- Un kernel a été adopté pour la préparation de données.

II- Description du jeu de données

- Le jeu de données "Home Credit" décrit les détails financiers et bancaires des clients.

- Le dataframe a 307 511 exemples et 122 colonnes.
- Il est constitué de 106 entités numériques à savoir 65 de type float64 et 41 de type int64. Les entités de type objet sont au nombre de 16.
- Un dataframe déséquilibré:

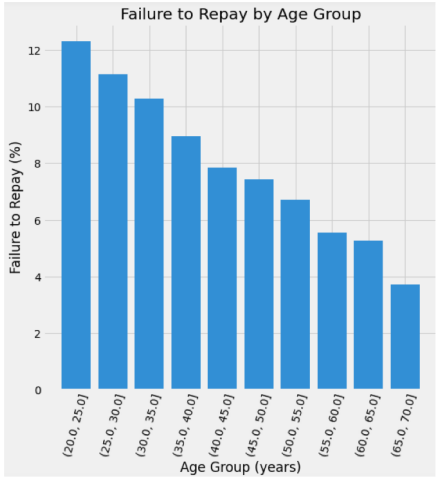


- Il y a 67 colonnes qui ont des valeurs manquantes (60% à 70%).
- Valeurs aberrantes: DAYS_EMPLOYED => (np.nan).

- Imputation simple Imputer avec la stratégie moyenne.

III- Transformation du jeu de données

- 1) Analyse exploratoire :
  - exemple : Effet de l'âge sur le remboursement



- 2) Transformation des variables catégorielles:
  - Label encoding pour toutes les variables catégorielles avec seulement 2 catégories
  - One hot encoding pour toutes les variables catégorielles avec plus de 2 catégories.

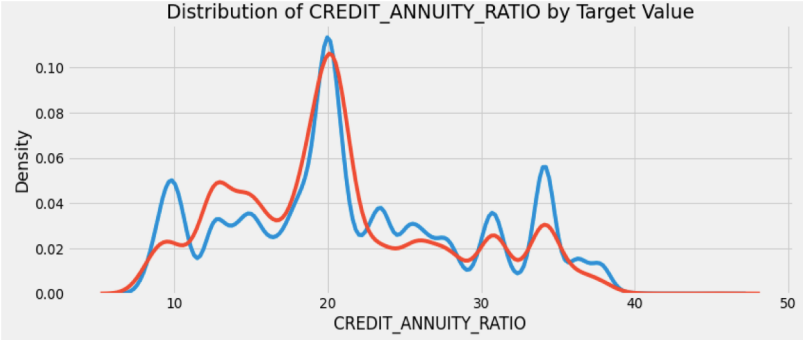
- 3) Création de nouvelles variables (Domain knowledge features):
  - CREDIT_ANNUITY_RATIO: le pourcentage du montant du crédit par rapport à la rente de prêt d'un client
  - CREDIT_GOODS_PRICE_RATIO: le pourcentage du montant du crédit par rapport au revenu d'un client
  - CREDIT_DOWNPAYMENT: la durée du paiement en mois
  - AGE_INT: le rapport entre l'âge du client et le nombre de jours par an

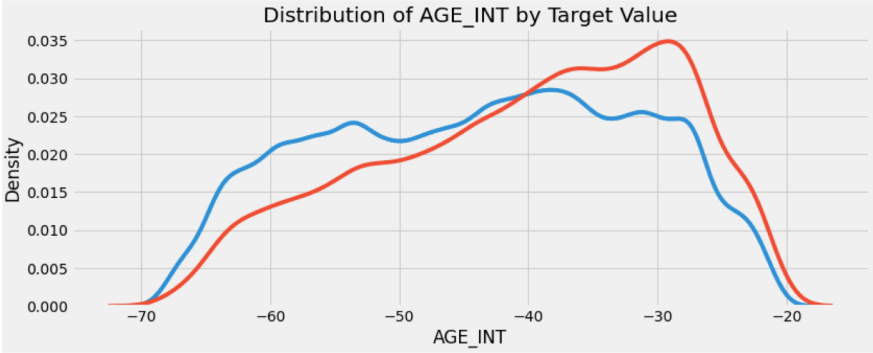
III- Transformation du jeu de données

- Corrélation entre variables et Target:

CREDIT_DOWNPAYMENT	-0.065407
CREDIT_ANNUITY_RATIO	-0.032102
CREDIT_GOODS_PRICE_RATIO	0.069427
AGE_INT	0.078234
TARGET	1.000000

- Les tracés KDE colorés par la valeur de TARGET:



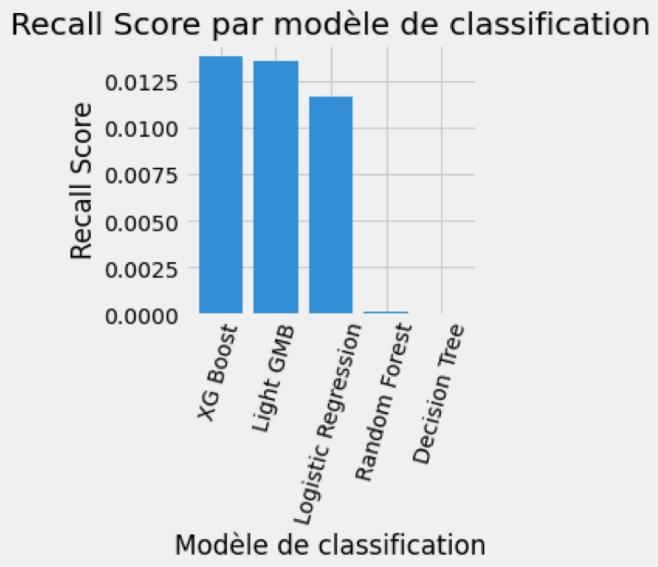


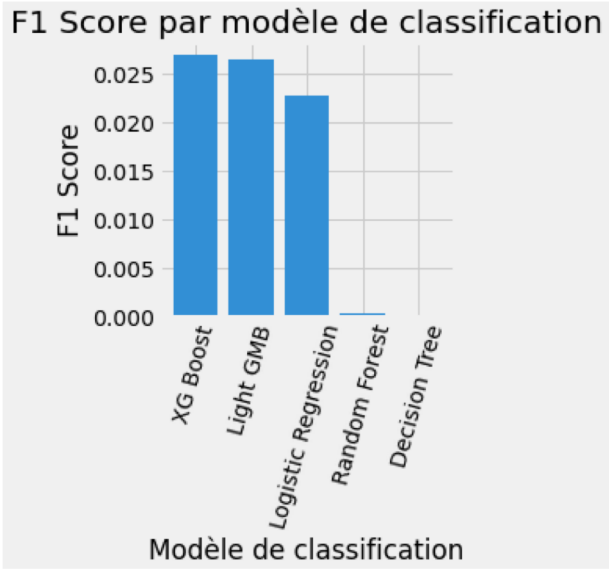
IV- Comparaison et synthèse des résultats

- Séparation des données en deux bases train/test (70% vs 30%).
- Standardiser des données.
- Modèles de classification à utiliser:
  - régression logistique, 'C':[1e-3,1e-2,0.1, 1, 10]
  - arbre de décision:
    - "n_estimators": [10, 30, 50, 100], "criterion": ["gini", "entropy"], "max_depth": [5, 10, 20],
  - random forests,
    - 'criterion':['gini','entropy'], 'max_depth': np.arange(3, 15)
  - Xgboost
    - 'max_depth':[3, 4, 5], 'min_child_weight':[1, 3, 5]
  - Lightgbm.
    - "n_estimators": [50, 75, 100], "boosting_type": ["gbdt", "dart", "goss"], "max_depth": [1, 2, 5]
- Application du Cross-validation avec GridSearchCV afin de calculer et d'optimiser les hypermaramètres.

IV- Comparaison et synthèse des résultats

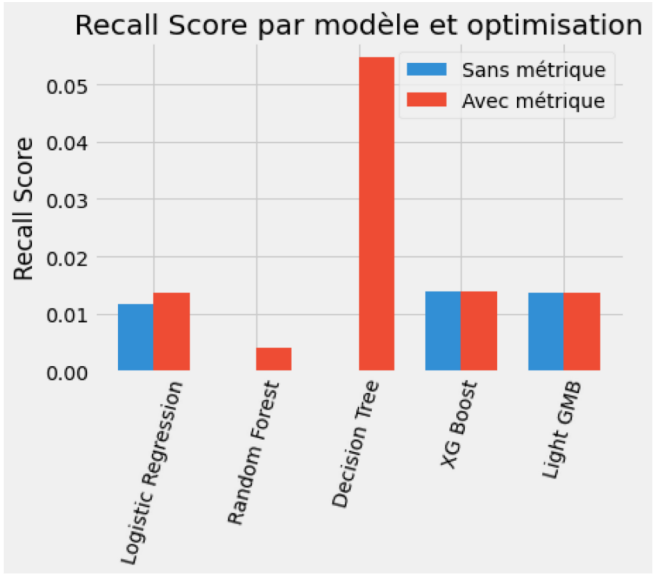
- **Sans améliorations:**
  - Utilisation de f1_score dans le scoring.

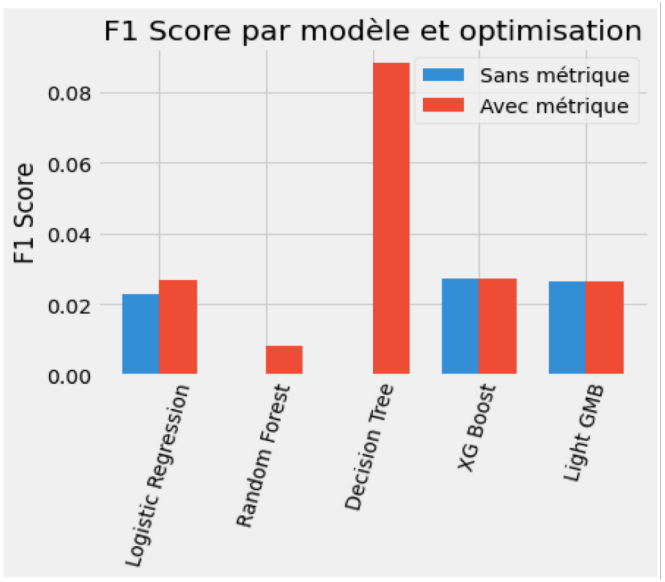


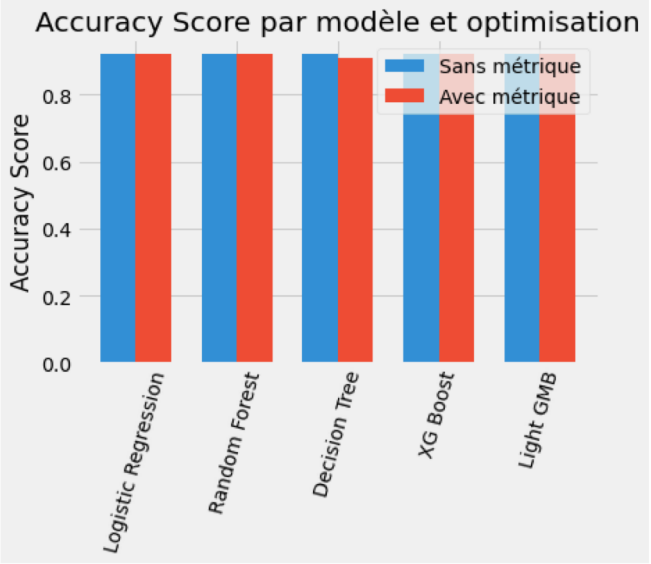


IV- Comparaison et synthèse des résultats

- **Métrique métier personnalisée :**
  - FP = perte des intérêts non acquis.
  - FN = perte des intérêts et perte de crédits.
    - min {metric_metier} = max { -(FP + 10 * FN) }

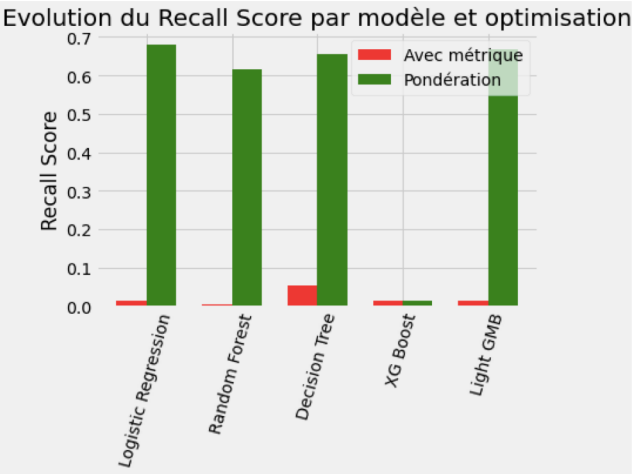


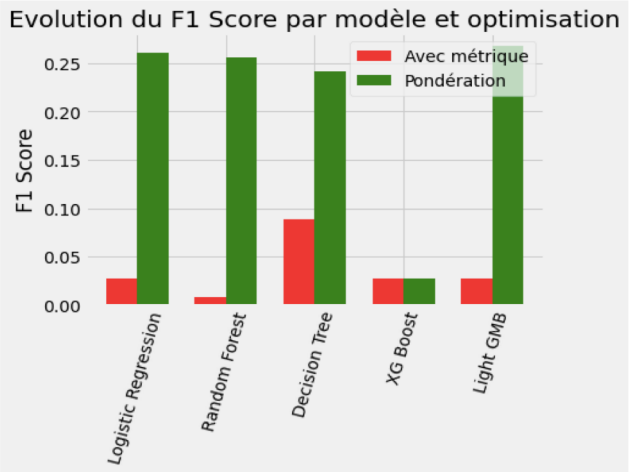


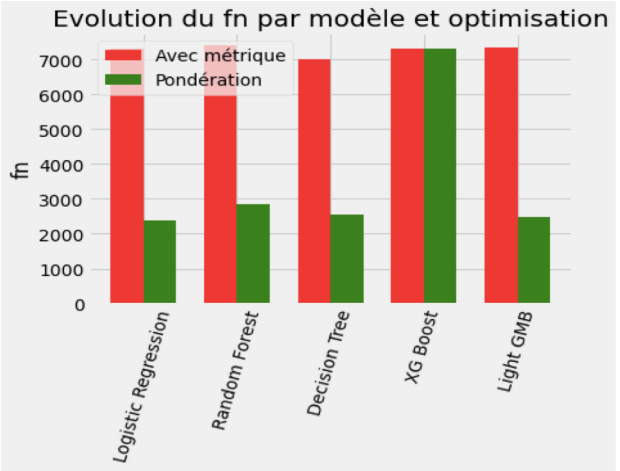


IV- Comparaison et synthèse des résultats

- **Gestion du déséquilibre des classes avec la pondération (classweight)**
  - Le scoring est fait à la base de la métrique personnalisée.



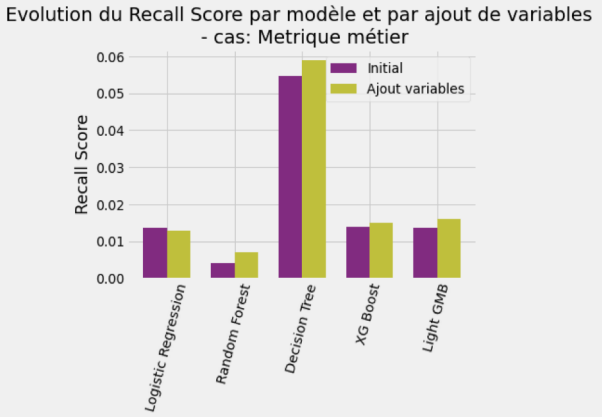


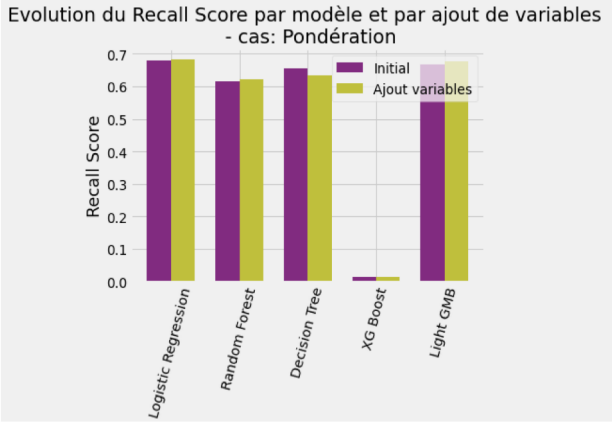


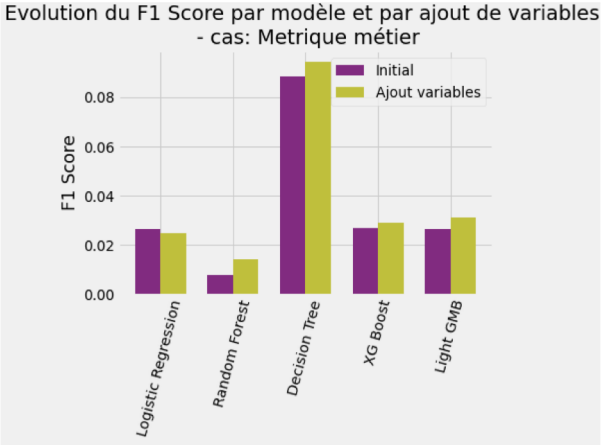
IV- Comparaison et synthèse des résultats

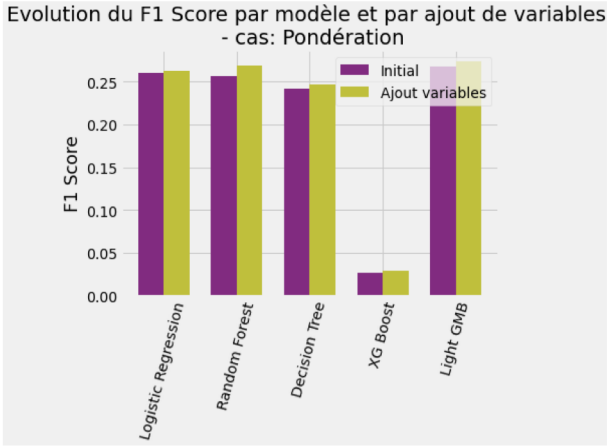
❖ Ajout de nouvelles variables

- Le scoring est fait à la base de la métrique personnalisée.









Synthèse de l'analyse de données

Interprétation globale :

- Le "recall" et le "F1_score" augmentent
- Le "Fn" et la "métrique métier" diminuent

- Light GMB est le meilleur puisqu'il fait un compromis entre le nombre de variables et la classification binaire.

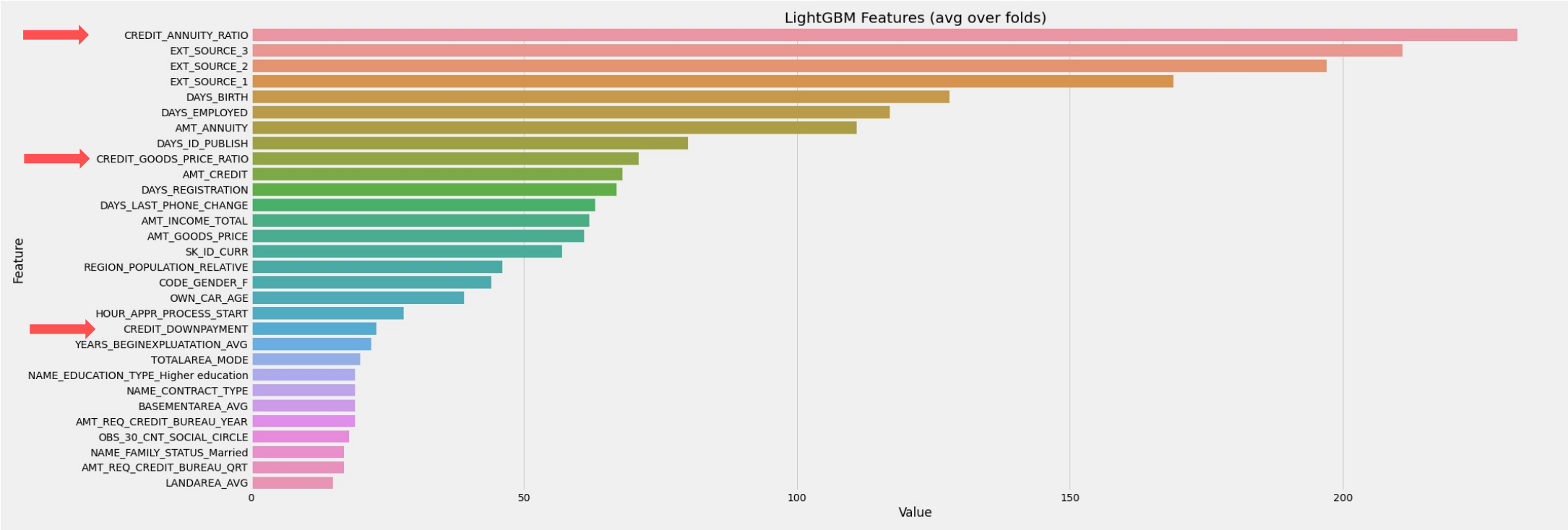
- Les résultats de XGBoost ne sont plus intéressants dès qu'il y a ajout des améliorations.

- Random forest et de l'Arbre de décision donnent de bon résultats avec les optimisations (244 variables).

- La régression logistique se porte à la perfection à ce type de base de données (classification binaire).

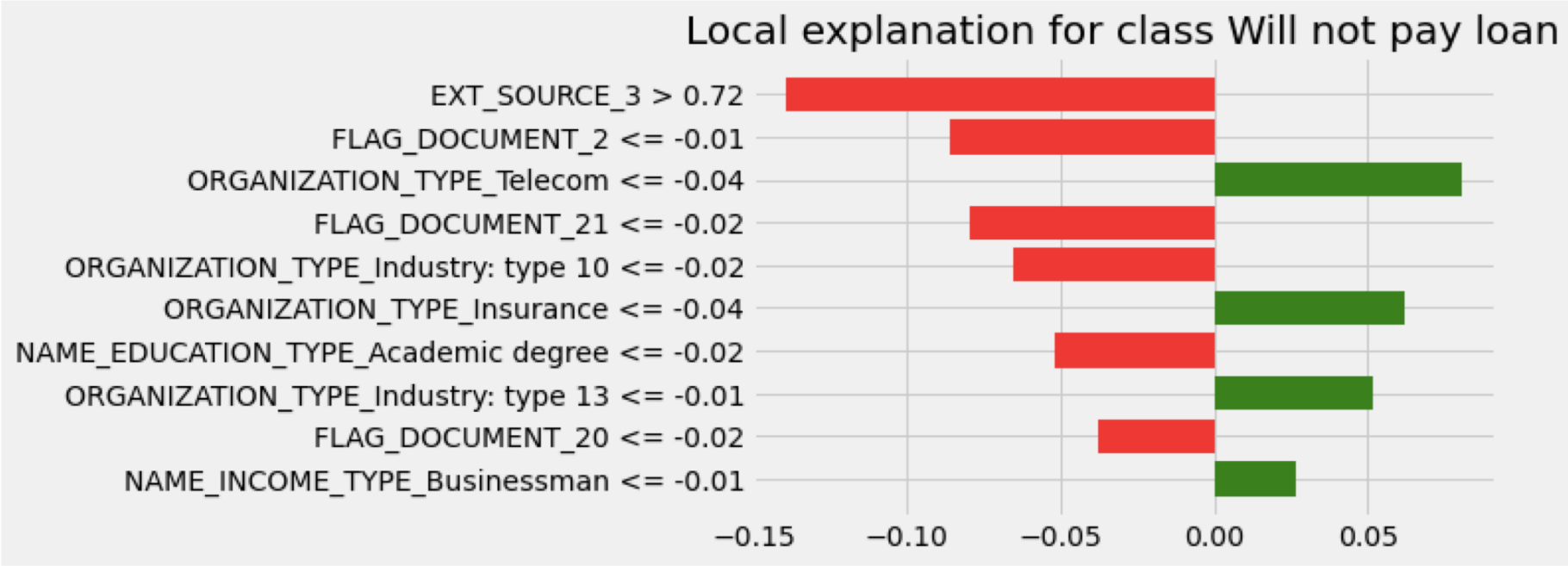
IV- Interprétabilité du modèle

- **Fonction « Feature Importance »**
  - Meilleur modèle. « Light GBM » avec les meilleurs hyper-paramètres en utilisant la pondération.



IV- Interprétabilité du modèle

- **Package « LIME»**
  - Meilleur modèle. « Light GBM » avec les meilleurs hyper-paramètres en utilisant la pondération.



Conclusion & perspectives

Conclusion :

- Stratégies importantes pour l'amélioration des prédictions:
  - Choix de variables
  - Création d'une métrique
  - Gestion du problème du déséquilibre des classes
  - Choix du bon algorithme
  - Validation croisée & choix des meilleurs hyper-paramètres

Perspectives :

- Combiner d'autres fichiers csv fournis par HomeCredit
- Une meilleure préparation des données
- Appliquer la technique d'augmentation des données

MERCI DE VOTRE ATTENTION