

NOTE METHODOLOGIQUE : **IMPLEMENTEZ UN MODELE DE SCORING**

**ESSIS KAROU JOSE**

**OPENCLASSROOMS : PROJET 7**

Sommaire

Introduction

**Prêt à dépenser** est une entreprise qui qui propose des crédits à la consommation pour des personnes. Elle souhaite mettre en œuvre un outil de scoring pour calculer la probabilité qu’un client rembourse son crédit ou pas en classifiant la demande en crédit accordé ou refusé. Ainsi pour réaliser ce projet, elle souhaite mettre en place un algorithme de classification à partir de différents types de données. Ces données sont d’origine variées (données comportementales, données provenant d'autres institutions financières).

Faisant partir de la team Data Scientist au sein de **Prêt à dépenser**, nous avons construit plusieurs modèles de classification et leur métrique afin d’en sélectionner le meilleur.

Par la suite nous ferons une interprétation globale et locale de notre modèle.

Enfin nous montrerons les limites de notre modèles et les pistes d’améliorations pour rendre ce modèle meilleur.

1. Features engineering

Notre jeu de données contient ces différentes tables :

application\_train

application\_test

bureau

bureau\_balance

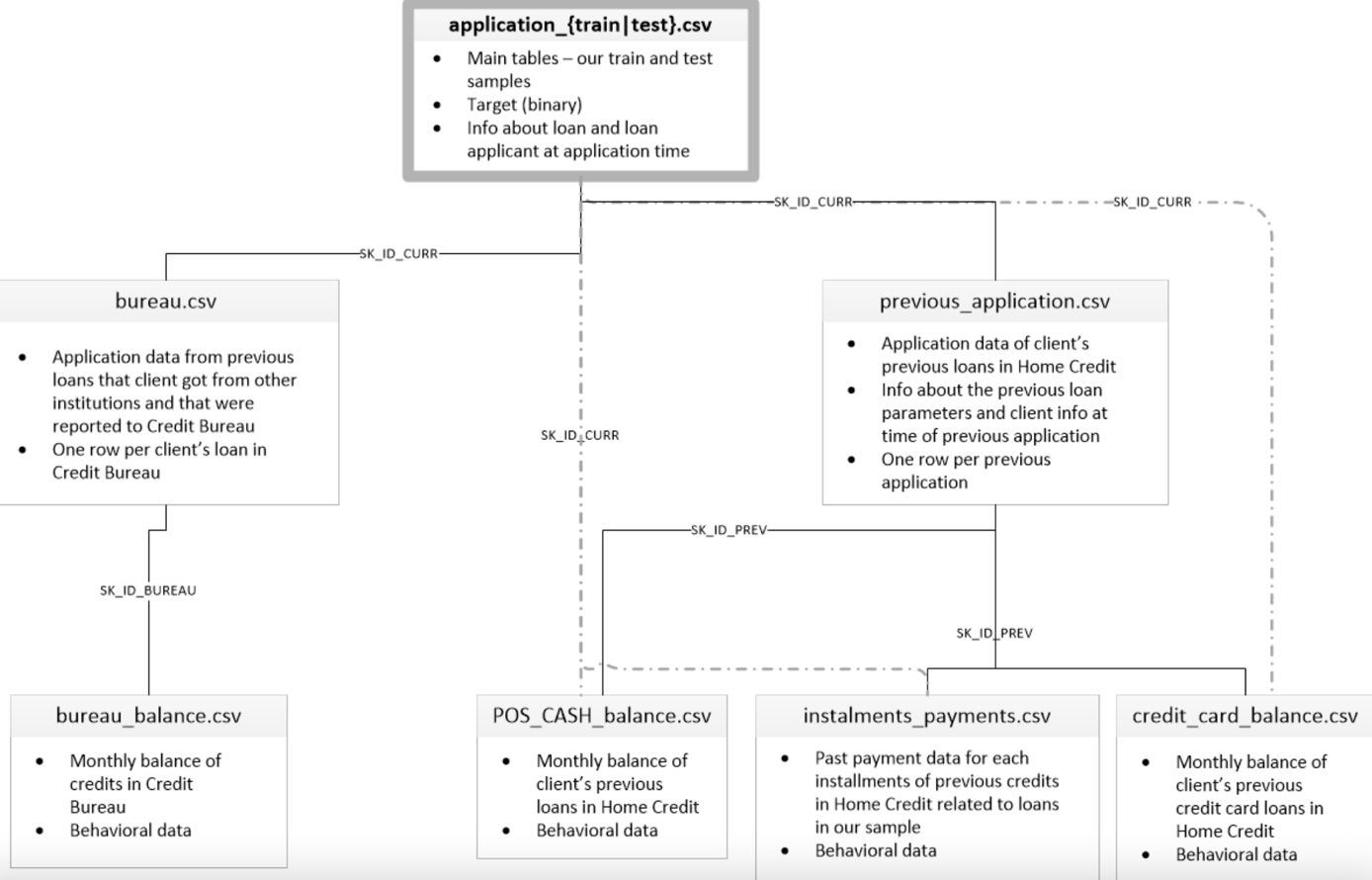
POS\_CASH\_balance

credit\_card\_balance

previous\_application

installments\_payments

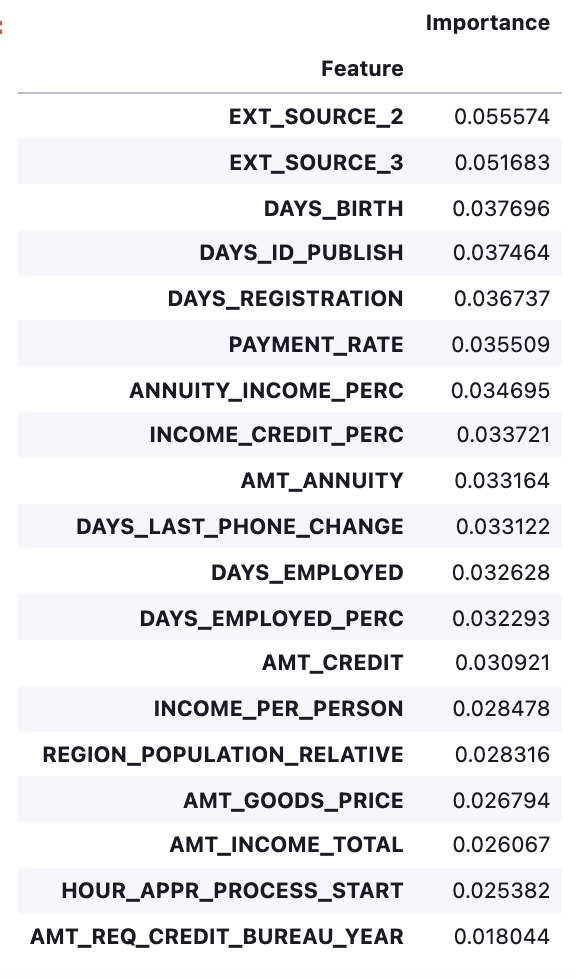
HomeCredit\_columns\_description



Ces tables contiennent des données de types catégorielles et des données de types numériques entiers et décimale.

Notre feature engineering à consister :

* Création de nouvelles variables à partir des données de types catégorielle en données types numériques entières par la méthode One hot encoding.
* Création de nouvelles (moyenne, sommes, maximum, minimum et size) à partir de l’agrégation de données types numériques entières et décimale.
* Remplacement des outliers par des valeurs na (valeurs manquantes)
* Suppression des variables ayant des taux de données valeurs manquantes supérieurs à 80 pourcents
* Imputation par la médiane des variables ayant des taux de données valeurs manquantes inférieurs à 20 pourcents.
* Sélection des variables les plus importantes avec Random Forest Feature Importance



1. Méthodologie d'entraînement des modèles de classification

* Custom metrics

La fonction coût métier a pour objectif de minimiser les pertes. Nous estimons notre coefficient alpha à 5. Cela veut dire que nous estimons qu’il est 5 fois plus coûteux à la banque de d’octroyer un prêt à un client que ne peut pas rembourser que de ne pas le donner à un client qui peut rembourser un prêt. En résumé un faut négatif coute 5 fois plus cher que une faux positifs.

* Forêt aléatoire

Les algorithmes de Forêt aléatoire sont basés sur les arbres de décision. Il effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. En effet, au lieu d’essayer d’obtenir une méthode optimisée en une fois, on génère plusieurs prédicteurs avant de mettre en commun leurs différentes prédictions.

* Régression logistique

La régression logistique est un modèle statistique permettant d’étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s’agit d’un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien. Un modèle de régression logistique permet aussi de prédire la probabilité qu’un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l’optimisation des coefficients de régression

* Light GBM

[LightGBM](https://lightgbm.readthedocs.io/en/latest/) est un algorithme de machine learning basé sur le [Gradient Boosting Machine](https://lovelyanalytics.com/2016/09/12/gradient-boosting-comment-ca-marche/).

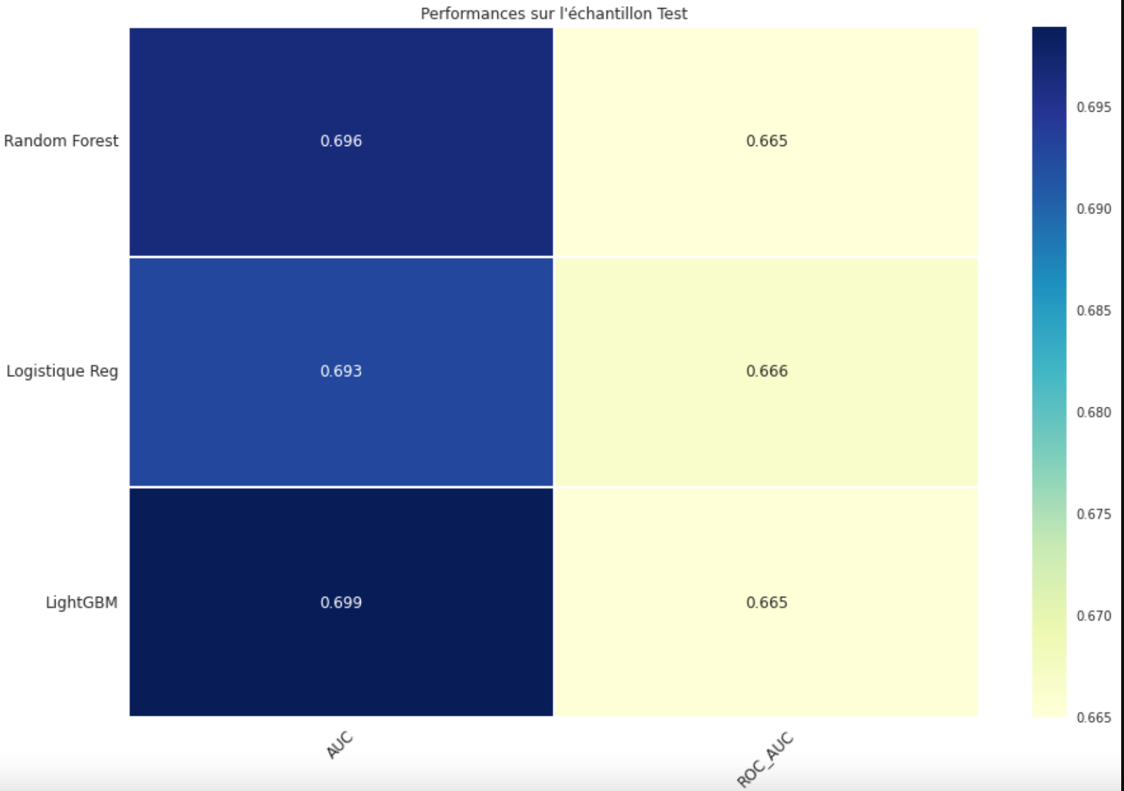
Dans le Gradient Boosting, les arbres sont construits 1 à 1, avec chaque nouvel arbre qui vient corriger les prédictions incorrectes de l’arbre précédent alors que avec le bagging de la forêt aléatoire combinent les résultats à la fin du processus en calculant la moyenne.

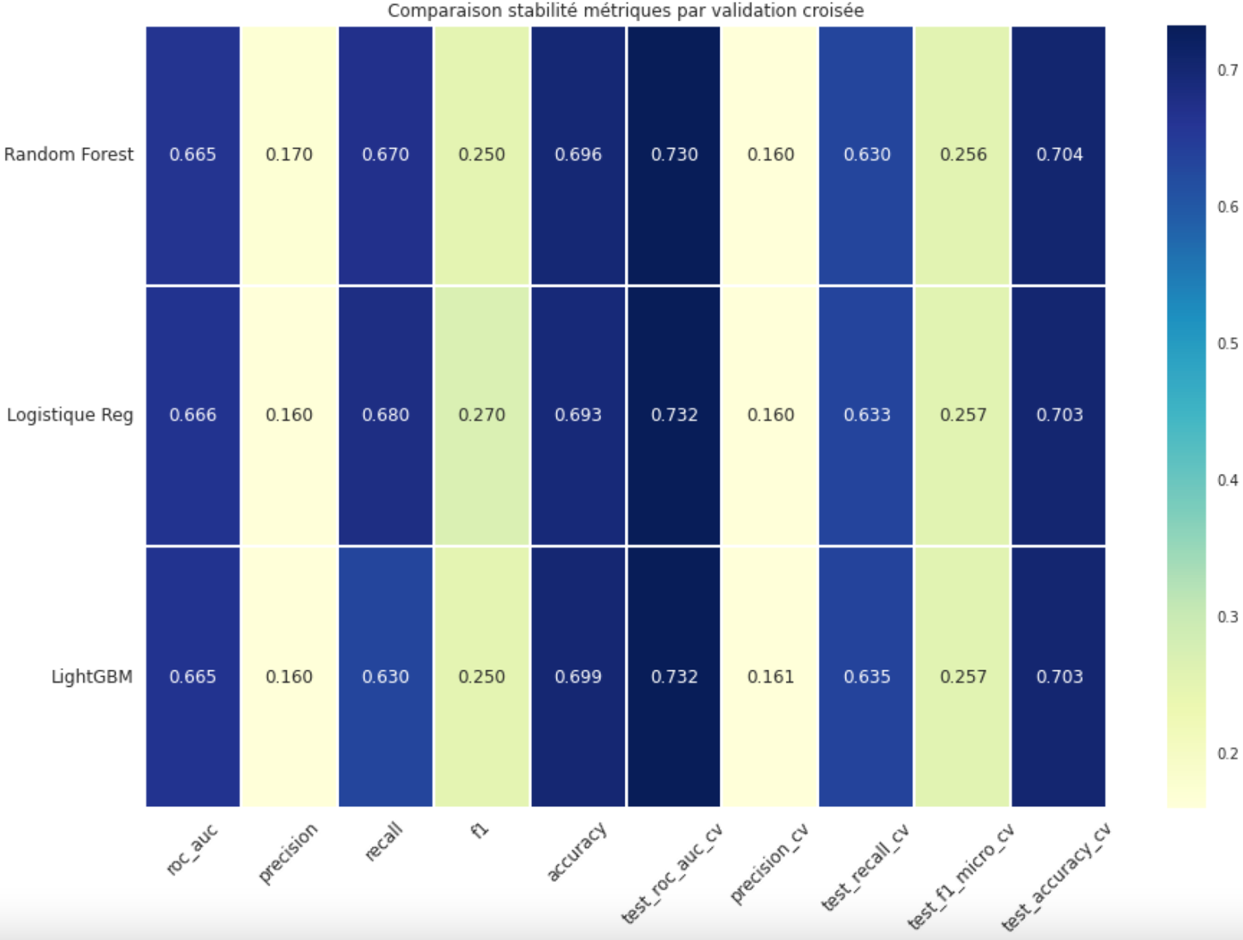
* L’optimisation des hyperparamètres

Le choix des meilleurs hyperparamètres de notre modèle a été possible grâce à la méthode d’optimisation numérique.

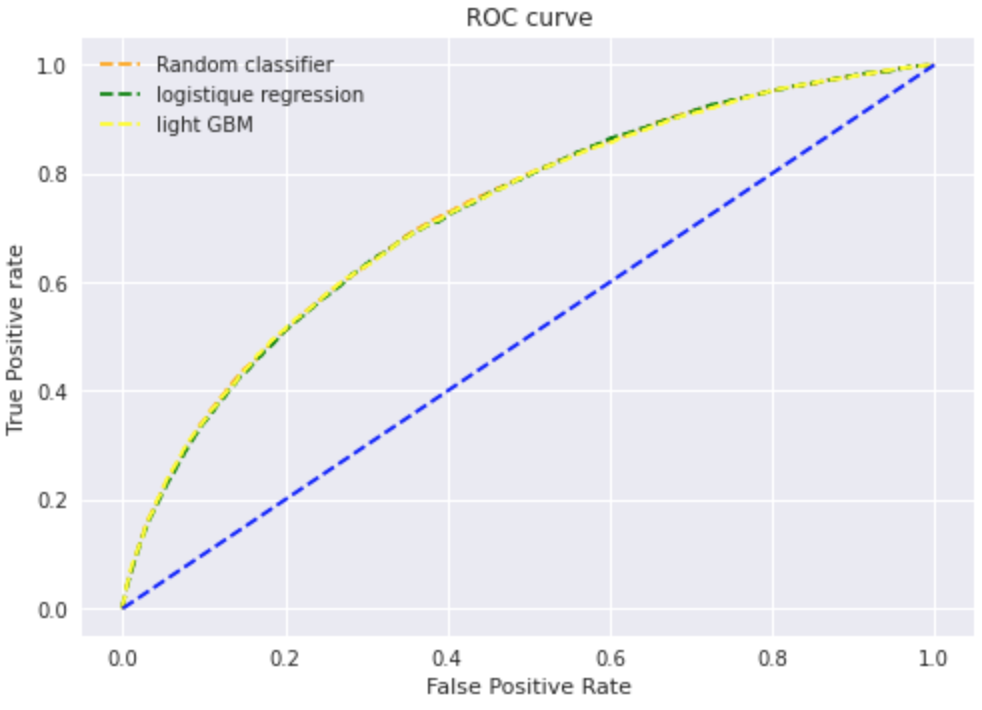
1. Evaluation du modèle et choix du modèle

* Comparaison des métriques d’évaluation de chaque modèle



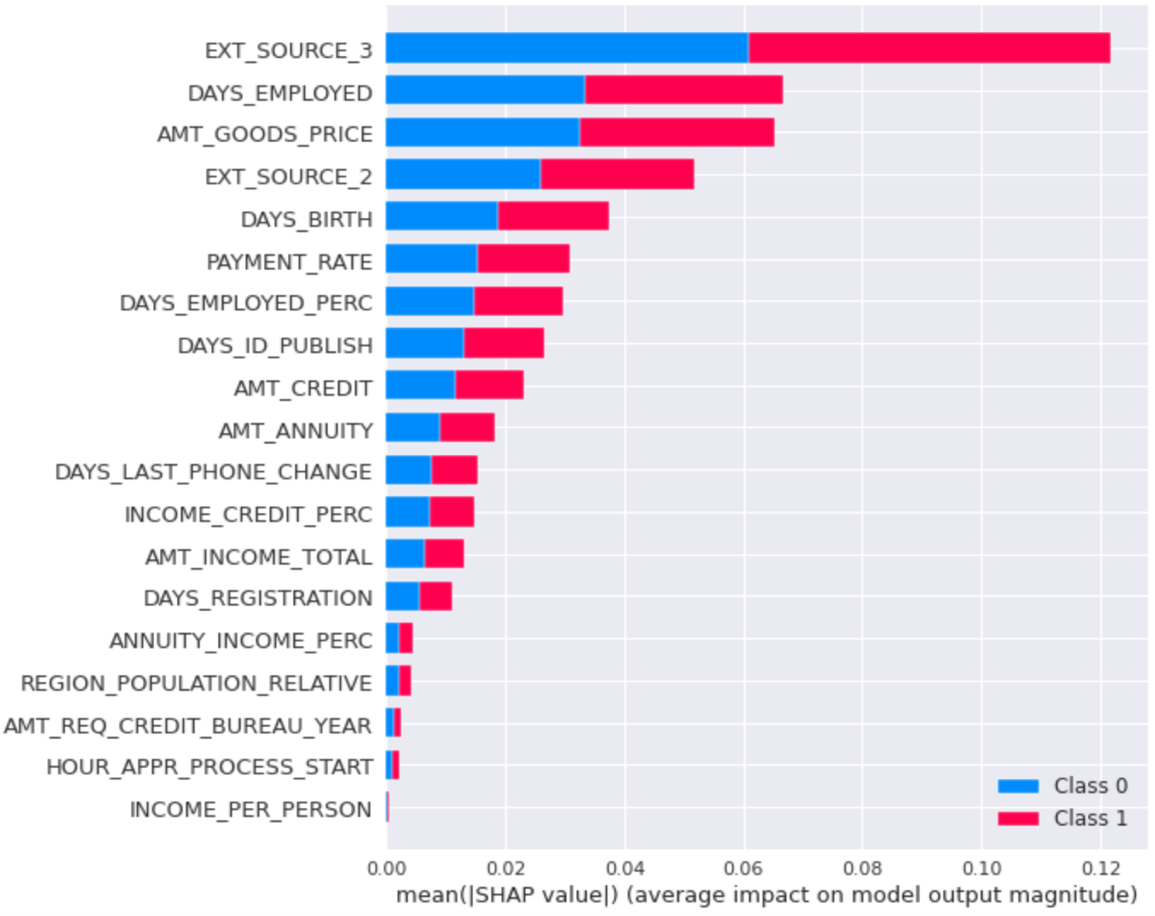


Le modèle lightGBM à les meilleures métriques mais il faut noter que la différence avec les autres modèles est faible. La cross validation de nos modèles montre également que nos modèles sont plutôt stables.

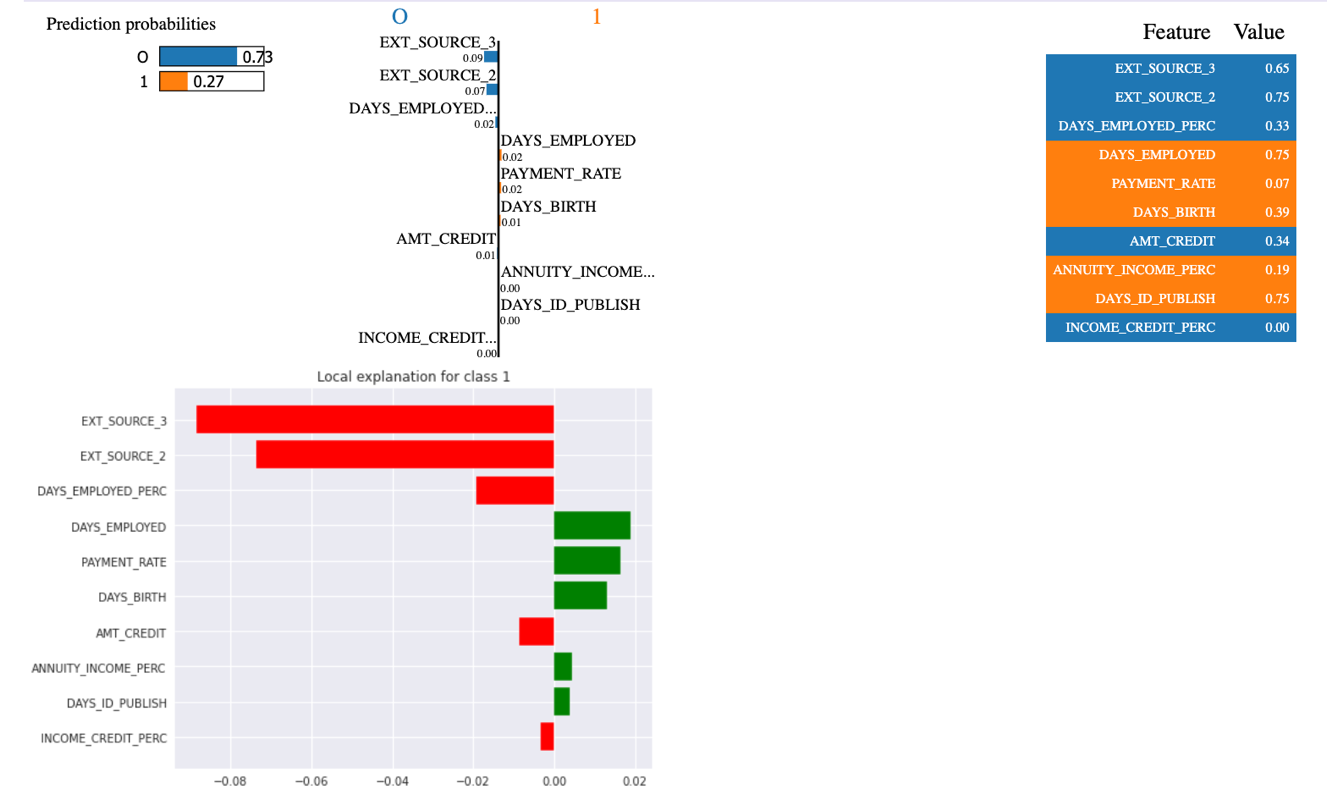


1. Interprétation du modèle

* Interprétation globale du modèle



* Interprétation locale du modèle



1. Limites et améliorations possibles

Sur la base test, notre modèle affirme que 60000 clients ont la capacité de rembourser leur crédit et que 4700 clients ne sont pas en mésure de rembourser leur crédit.

Par contre son estimation sur les faux négatifs est de 2800 clients et de 28000 sur les faux positifs. Notre octroie à 2800 clients des prêts qu’ils ne pourront pas rembourser et cette perte peut être grandes selon le montant des prêts. C’est également une perte pour l’institution bancaire de ne pas octroyer un prêt à des clients qui peuvent rembourser.

Une des pistes d’améliorations serait de redéfinir la fonction métier afin de pouvoir bien identifier et cerner les faux positifs et les faux négatifs.

L’idée serait de baisser considérablement les faux négatifs pour bien afin de les avoir dans les vrais négatifs et de baisser les faux négatifs pour les avoir dans les vrais positifs.