NOTE MÉTHODOLOGIQUE

Implémenter un modèle de scoring

Table des matières

[I. Le contexte 2](#_Toc87100542)

[II. Les données 2](#_Toc87100543)

[III. Méthodologie 3](#_Toc87100544)

[a) Choix de l’algorithme 3](#_Toc87100545)

[b) Entraînement du modèle 3](#_Toc87100546)

[c) Fonction coût métier 4](#_Toc87100547)

[IV. Interprétabilité 6](#_Toc87100548)

[a) Interprétabilité globale 6](#_Toc87100549)

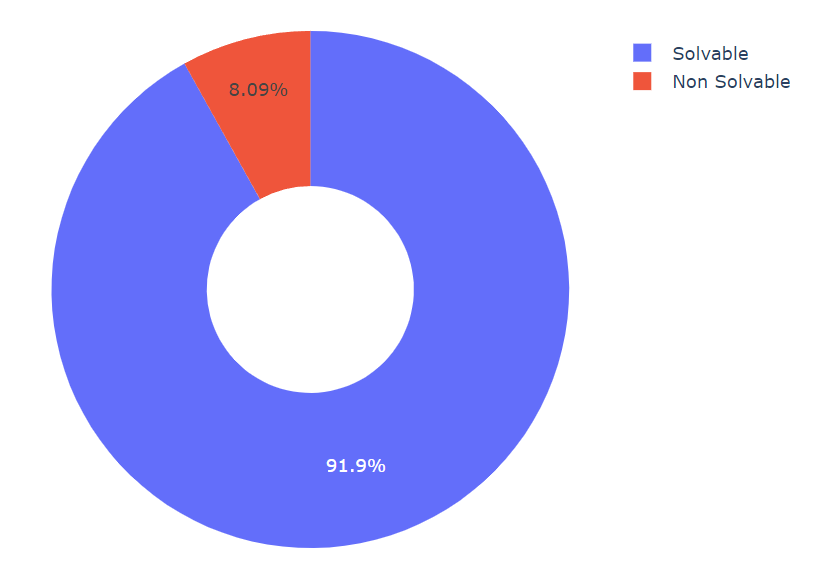
[b) Interprétabilité locale 6](#_Toc87100550)

# Le contexte

La société financière nommée "Prêt à dépenser" propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt. Cette entreprise souhaite développer un outil utilisant un modèle de scoring permettant d'obtenir la probabilité de défaut de paiement pour appuyer la décision d'accorder ou non un prêt à un client potentiel en s’appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.). Cet outil sera accompagné d'un Dashboard interactif permettant d'interpréter les prédictions faites par l'outil. Prédire si un client remboursera ou non un prêt ou s'il rencontrera des difficultés à le faire est un besoin commercial essentiel.

# Les données

Pour réaliser cet outil, c'est le service dédié à la fourniture de crédits à la population non bancarisée, Home Credit, qui fournit une base de données. Cette base de données regroupe des informations générales pour chaque client ainsi que des informations supplémentaires sur chaque prêt contracté par les clients. N'étant pas du métier de la banque, je me concentre essentiellement sur les données générales de chaque client comme l’âge, le sexe, les revenus, l’emploi, le logement, les informations de crédit en cours, des notations externes, etc. En observant de plus près la base de données, il apparaît un fort déséquilibre entre les personnes solvables et les personnes non solvables. Cette observation est représentée sur la figure ci-dessous.



# Méthodologie

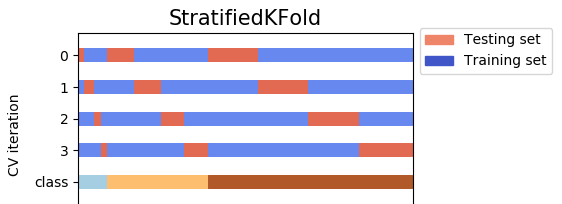
## Choix de l’algorithme

Le modèle doit prendre en entrée des données et donne en sortie une classe parmi deux (solvable ou non solvable). C’est donc un problème de classification binaire. L’algorithme Light GBM a été choisi pour sa capacité à gérer des datasets volumineux et pour ses performances.

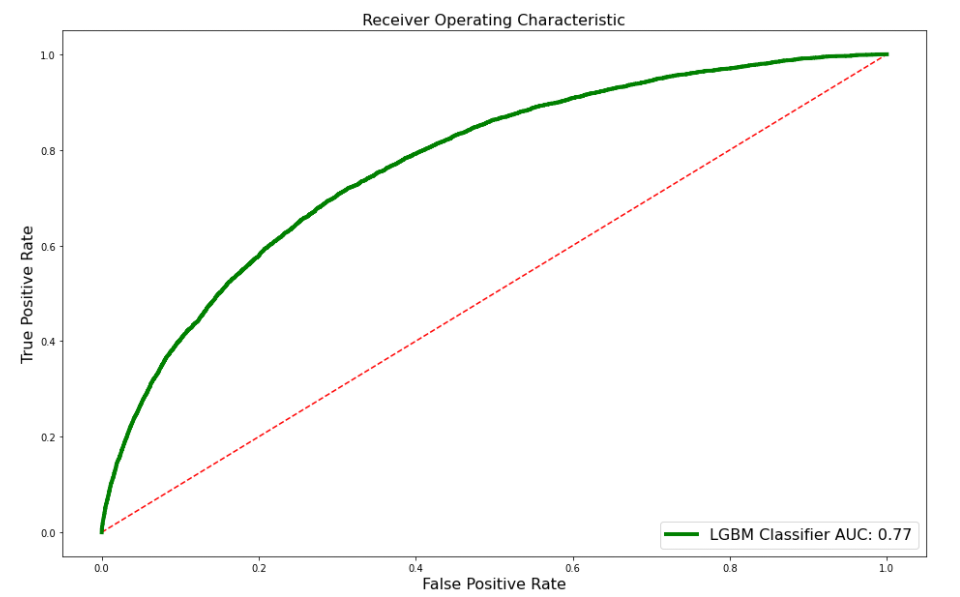
## Entraînement du modèle

Une validation croisée a été utiliser pour la recherche des hyperparamètres afin d’obtenir un modèle qui arrive à généraliser le problème. Cette validation croisée est de plus stratifiée pour avoir un nombre de classe égales entre les jeux de données séparé. Le modèle pourra donc ajuster ses hyperparamètres en s’entrainant à tour de Rôle sur les différentes parties du jeu de données sans être en surapprentissage.

*Exemple à trois classes :*



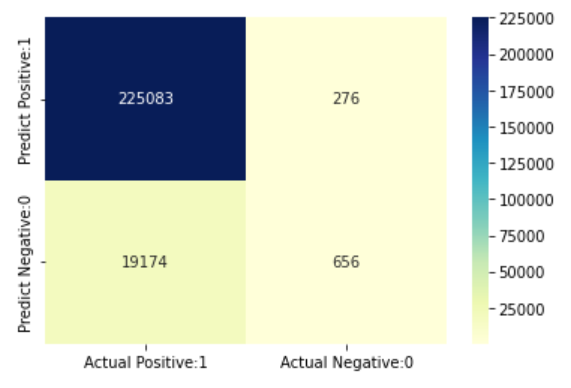
Cette première validation croisée cherche à optimiser l’AUC (aire sous la courbe).

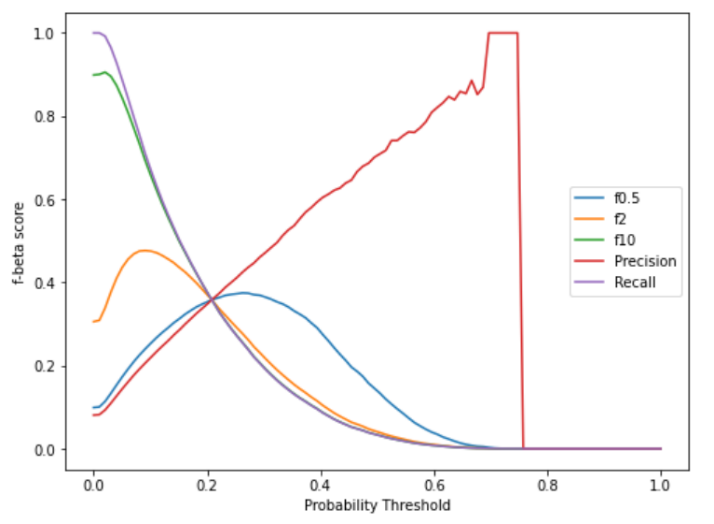


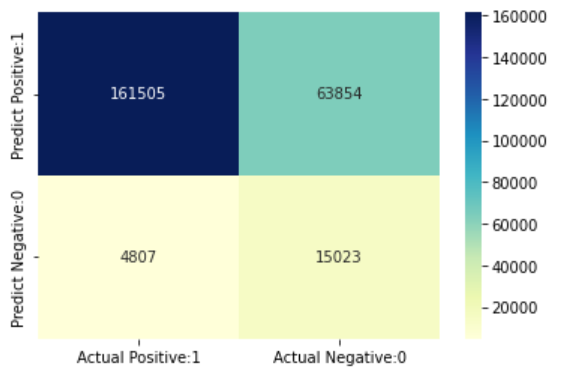
## Fonction coût métier

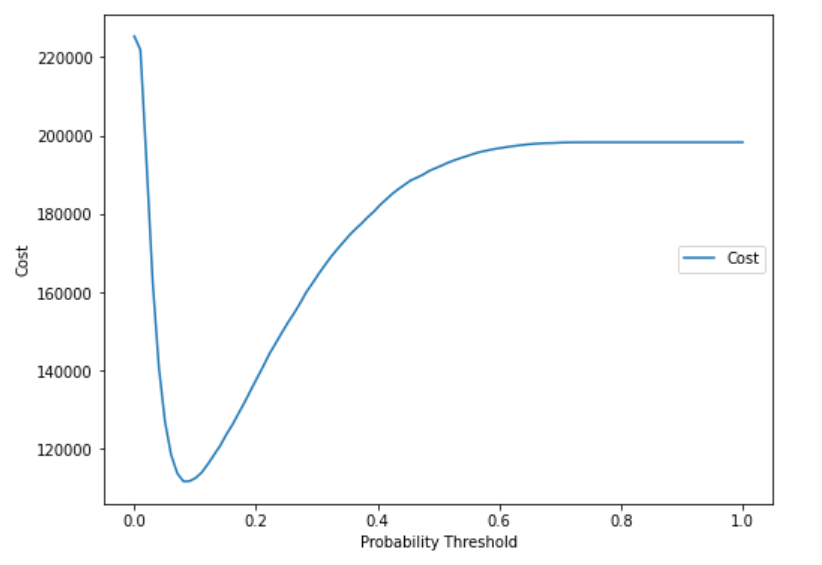
La problématique « métier » est de prendre en compte qu’un faux positifs (bon client considéré comme mauvais = crédit non accordé à tort, donc manque à gagner de la marge pour la banque) n’a pas le même coût qu’un faux négatif (mauvais client à qui on accorde un prêt, donc perte sur le capital non remboursé). Un faux négatif est environ 10 fois plus coûteux qu’un faux positif. La fonction à minimiser est donc:

*Coût métier = 10 \* FN + FP*

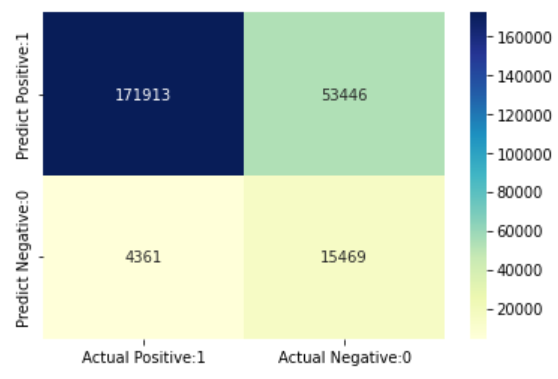
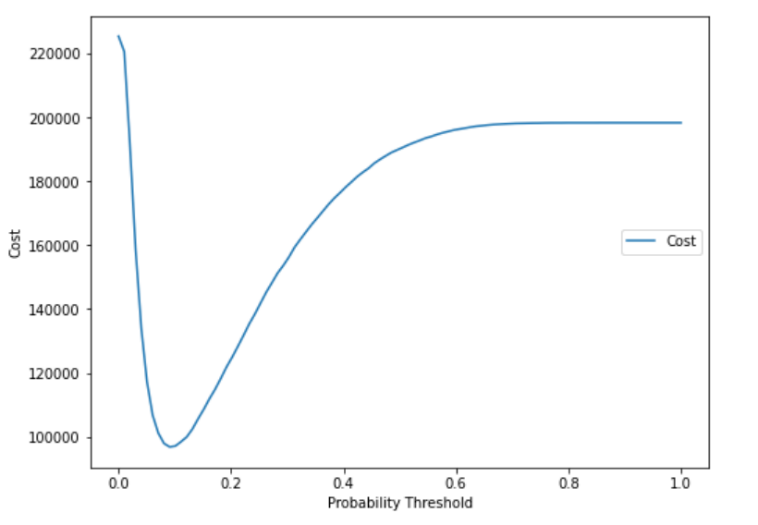


Pour ce premier exemple la fonction coût nous retourne 192 016. Après optimisation du seuil, cette fonction coût descend à 111 924 pour un seuil de 0,1.





En optimisant directement les hyperparamètres avec la validation croisée pour cette fonction coût il est encore possible de minimiser cette fonction coût :



En ajustant le seuil à 0.09, la fonction coût descend à 97 056.

Pour créer un score facilement interprétable, nous allons faire l’opération suivante :

On aura ainsi un score allant de 0 à 100. Plus il sera petit, plus le client sera à risque. Le nouveau seuil sera donc = 91.

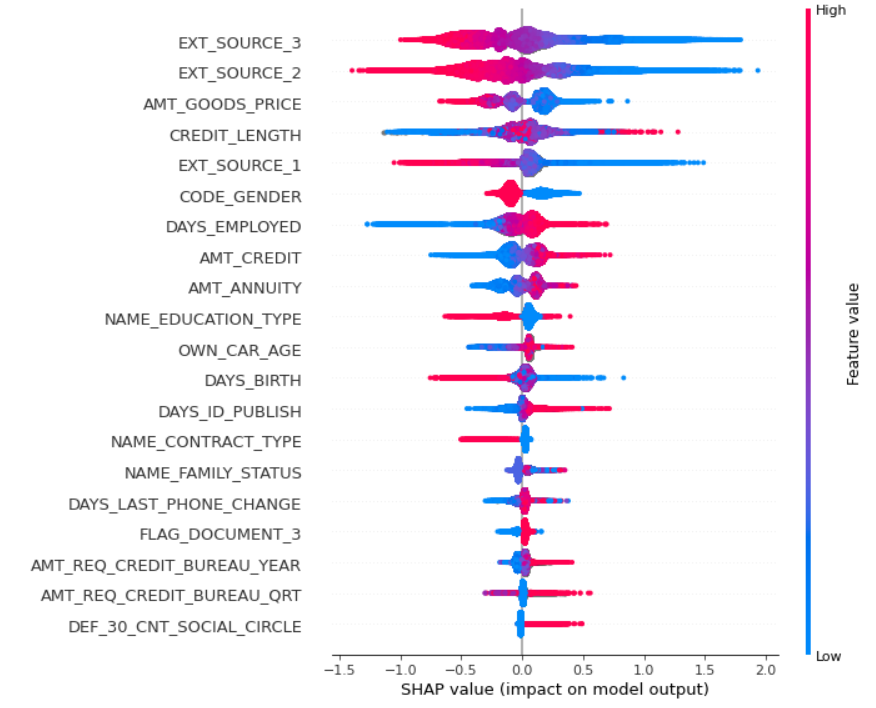
Il faudra ainsi un score supérieur ou égal à 91 pour pouvoir accorder un prêt.

# Interprétabilité

## Interprétabilité globale

Nous allons utiliser la librairie SHAP pour interpréter notre modèle final. SHAP (SHapley Additive exPlanations) est une approche théorique des jeux pour expliquer la sortie de tout modèle d'apprentissage automatique. Il relie l'allocation de crédit optimale aux explications locales en utilisant les valeurs de Shapley classiques de la théorie des jeux et leurs extensions associées

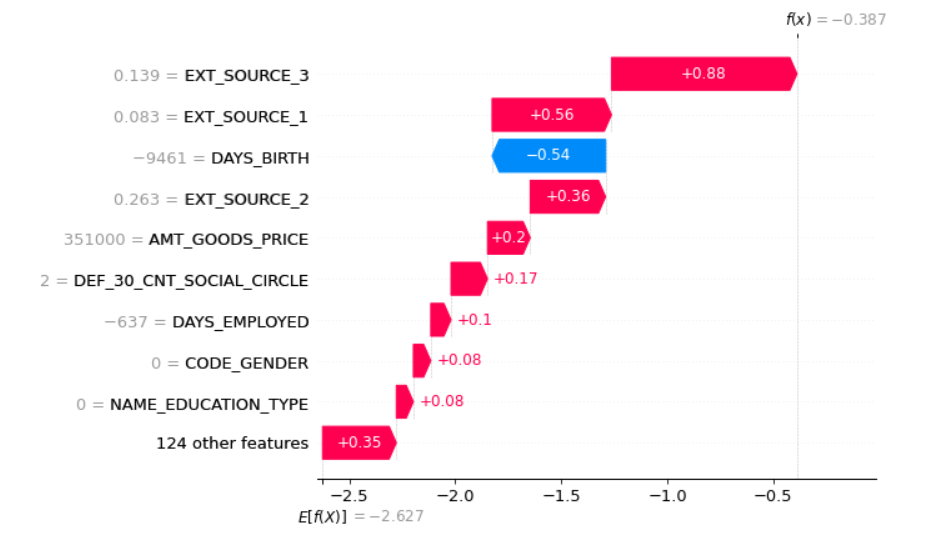
Importance des features et leur impact sur le modèle :



On constate que les note extérieurs (EXT\_SOURCE 1, 2, 3) ont beaucoup d’importance dans le résultat du modèle.

## Interprétabilité locale

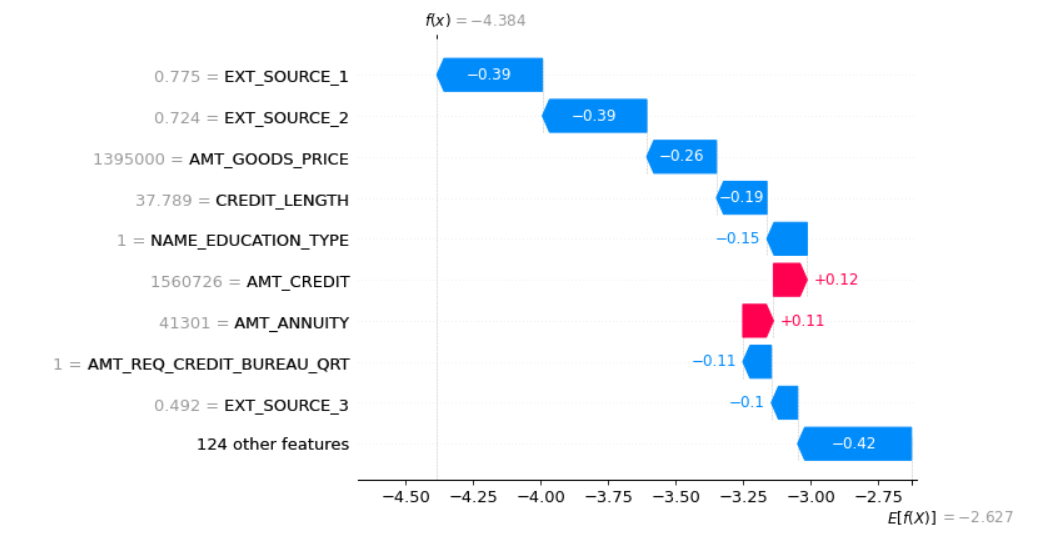
Comparons maintenant un client à risque et un client que l’on prédit non risqué :

Le client 100002 obtient un score de 52 donc que l’on prédit non solvable :

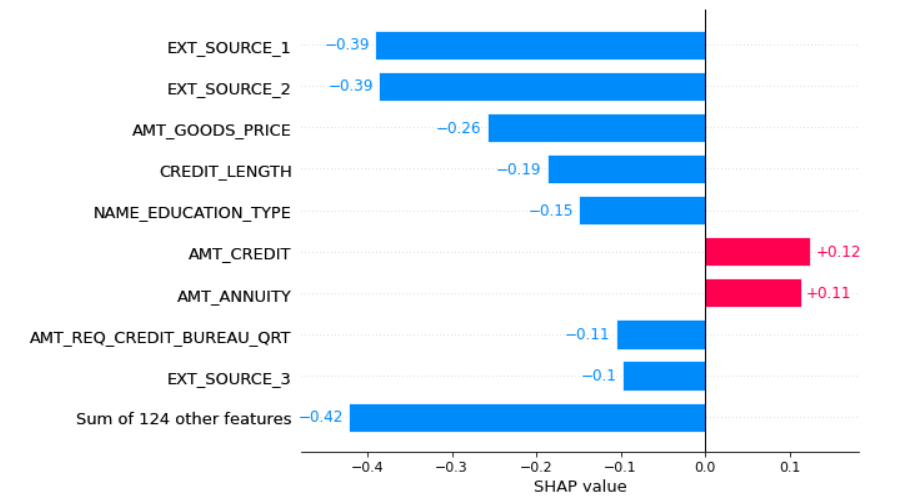
Grace à SHAP on peut facilement visualiser quelle variable lui fait défaut et à quel point.

Le client 100009 obtient lui un score de 99 et est donc prédit comme solvable :

*Waterfall plot*



*Bar plot*



# Limites et amélioration

L’optimisation des modèles a été effectué sur la base d’une métrique personnelle (cost = 10 x FN + FP) où l’on considère que prêter à un client non solvable est 10 fois plus couteux que de ne pas prêter à un client solvable. Ce coefficient a été choisi arbitrairement et pourrais être améliorer en calculant la perte en argent pour chacune des classes ce qui améliorerait significativement le gain de la banque.

Ce modèle prend en compte toutes les features comme le sexe et est donc pas construit “éthiquement”. Pour construire un modèle plus éthique il serait intéressant d’enlever certaines features discriminantes et de comparer ses prédictions avec notre modèle “non éthique” pour voir s’il est vraiment nécessaire de garder ses features discriminantes et faire un modèle plus éthique.

Le Dashboard interactif pourrait être améliorer pour être plus fluide, rapide et répondre au mieux aux attentes des besoins du banquiers et de ses clients.