



**FORMATION DATA SCIENTIST**

|  |  |
| --- | --- |
|  | **NOTE METHODOLOGOGIQUE** |

**Implémentez un modèle de scoring**

**Projet 7**

**Samir HINOJOSA DIAZGRANADOS**

Février 2022

Table des matières

[1. Introduction 3](#_Toc96510239)

[2. Présentation des données 3](#_Toc96510240)

[3. Modélisation 4](#_Toc96510241)

[3.1. Sélection du Kernel pour la Feature Engineering 4](#_Toc96510242)

[3.2. Optimisation du modèle 4](#_Toc96510243)

[3.2.1. Traitements initiaux 4](#_Toc96510249)

[3.2.2. Données déséquilibrées 5](#_Toc96510250)

[3.2.3. Modèles par défaut 5](#_Toc96510251)

[3.2.4. Modélisation finale 6](#_Toc96510252)

[3.2.4.1. Fonction coût 6](#_Toc96510253)

[3.2.4.2. Résultats de modélisation 6](#_Toc96510254)

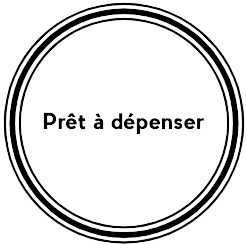
[4. Interprétation du modèle 8](#_Toc96510255)

[4.1. Interprétation globale 8](#_Toc96510256)

[4.2. Interprétation locale 8](#_Toc96510257)

[5. Limites et améliorations possibles 9](#_Toc96510258)

# **Introduction**

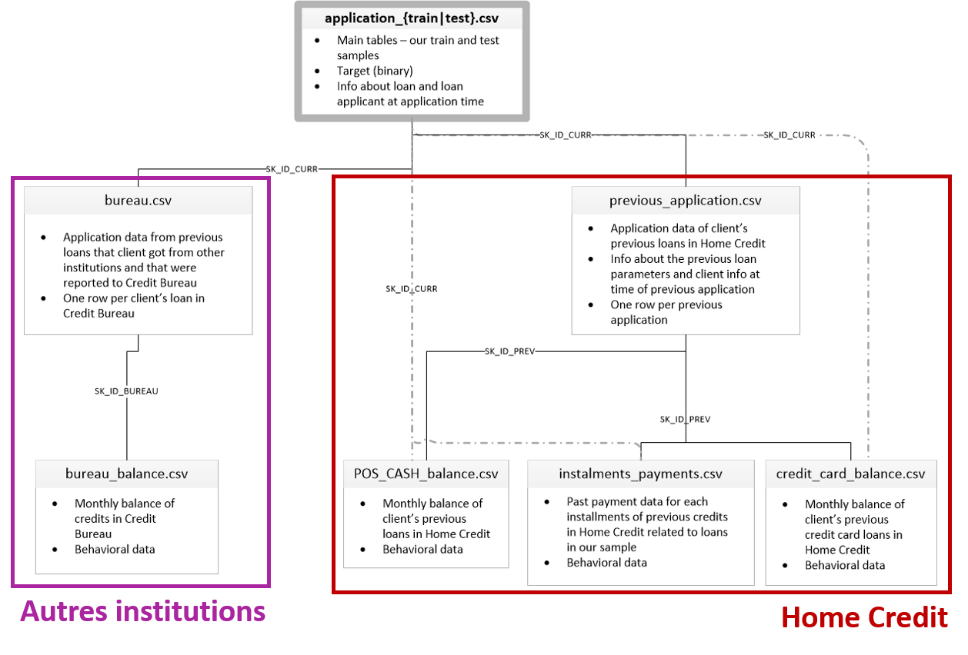
La société « **Prêt à dépenser** » souhaite mettre en œuvre un outil de « **scoring crédit** » pour calculer la probabilité qu’un client rembourse son crédit d’après diverses données.

La mission est de développer un algorithme qui classifie la demande en crédit accepté ou rejeté.

Prendre en compte :

* Sélectionner un kernel Kaggle pour faciliter la préparation des données nécessaires à l’élaboration du modèle de scoring.
* Déployer le modèle de scoring comme une API.
* Construire un tableau de bord interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle de scoring

# **Présentation des données**

****

Sept (7) tables de données sont mises à disposition pour la démarche.

Il y a de tables de données qui viennent des autres institutions :

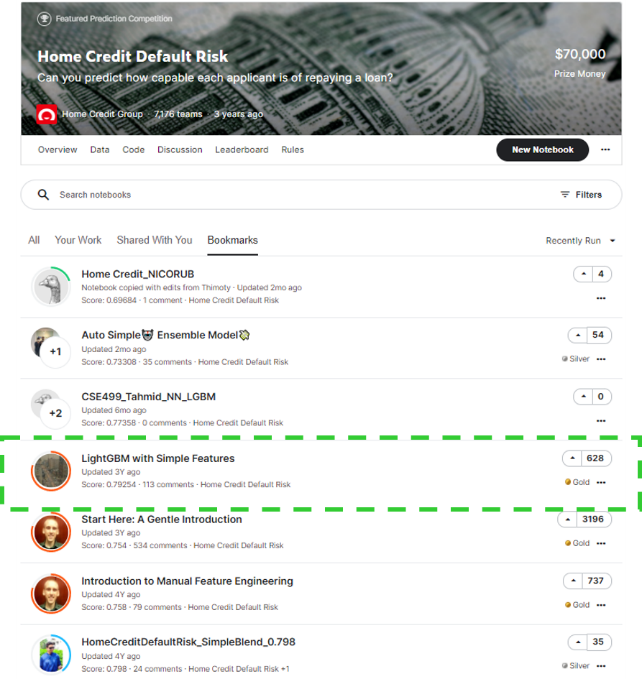
* Bureau : Antécénts de crédit du client (autres institutions financières)
* Bureau\_balance : Soldes mensuelles de crédits précédents

Aussi, on peut trouver de tables de données propre de « **Prêt à dépenser** »

* Previos\_application: Demandes de crédit antérieures
* Pos\_cash\_balance: Des bilans mensuels des anciens points de vente et des prêts cash
* Installements\_payments: Historique de remboursement des crédits précédents
* Credit\_card\_balance: Solde mensuel des cartes de crédit antérieures

# **Modélisation**

## **Sélection du Kernel pour la Feature Engineering**

Le kernel utilisé pour le preprocessing est

[LightGBM with Simple Features](https://www.kaggle.com/jsaguiar/lightgbm-with-simple-features).

La sélection est basée sur

* Un bon score dans la compétition
* Un Feature Engineering performant
* Le kernel recommandé dans le projet

Tout au long du kernel, des traitements sont faits pour obtenir des nouvelles données

* Identification et traitement des variables catégorielles   
    
  Création de nouvelles variables

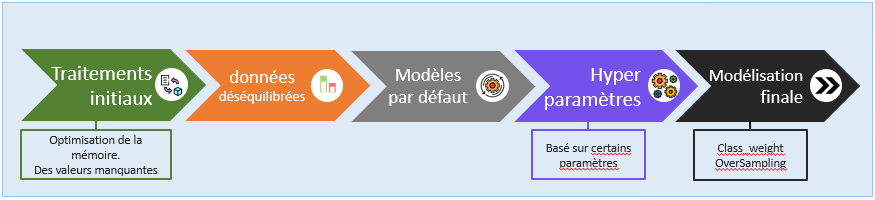
DAYS\_EMP\_% = DAYS\_EMP / DAYS\_BIRTH

Min, Max, Mean, Sum, Var

* Unification de jeux des données

+700 variables en total

## **Optimisation du modèle**



Pour faire la modélisation, on va prendre en compte le Feature Engineering déjà fait dans le kernel choisi.



### **Traitements initiaux**

Tout d’abord, il faut faire une optimisation d’usage de la mémoire parce que l’utilisation actuelle est 2,1 GB. Alors, on va transformer les types de colonnes

* **if** df[col]**.**dtype **==** "int64" **and** df[col]**.**nunique() **==** 2:

df[col] **=** df[col]**.**astype("int8")

* if df[col]**.**dtype **==** "float64" **and** df[col]**.**min() **>=** **-**2147483648 **and** df[col]**.**max() **<=** 2147483648:

df[col] **=** df[col]**.**astype("float32")

À travers cette transformation, l'usage de la mémoire est passé de 2.1Gb à 941 Mb.

D’ailleurs, on a fait aussi, ***le traitement des valeurs manquantes*** à travers :

* Suppression des colonnes avec 20 % ou plus de valeurs manquantes
* L’imputation de la moyenne basée sur la colonne.

### **Données déséquilibrées**

L’analyse exploratoire a montré que la mission a un problème de classification déséquilibrée.

Une modification de l’ensemble des données est possible avant d’entraîner le modèle prédictif afin d’équilibrer les données (égaliser les clases).

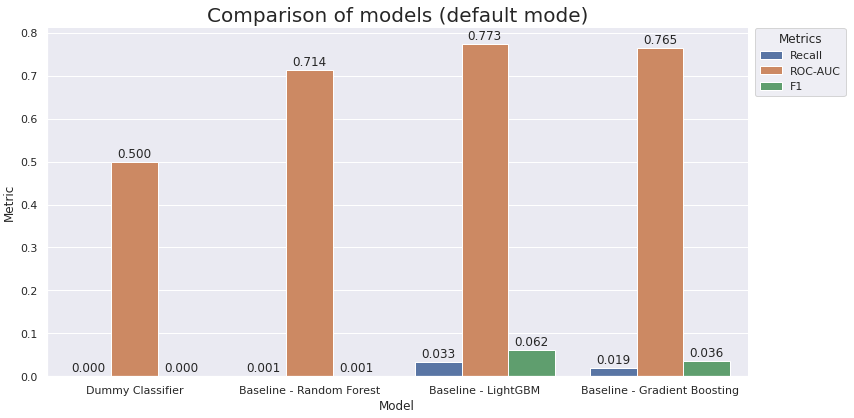
(0) Ce sont des prêts qui ont été remboursés et (1) Ce sont des prêts qui n'ont pas été remboursés

Dans ce cas, on va utiliser deux stratégiques :

* Une stratégie appelée rééchantillonnage (re-sampling), en se concentrant sur la méthode du *sur-échantillonnage (Oversampling) avec SMOTE[[1]](#footnote-1) (Synthetic Minority Oversampling Technique).*
* Utilisation de class\_weight comme paramètres pour faire une pondération inversement proportionnellement à la fréquence des classes

### **Modèles par défaut**

Au moment de réaliser la modélisation en prenant en compte les valeurs par défaut, le LightGMB a obtenu le meilleur résultat.

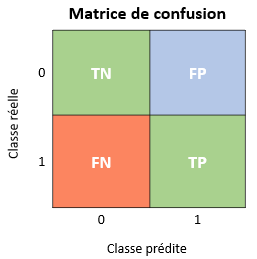


### **Modélisation finale**

On va utiliser trois algorithmes (Random Forest, LightGBM et Gradient Boosting).

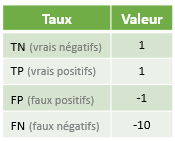
### **Fonction coût**

Pour ce projet, une fonction coût a été développée dans l’objectif de pénaliser des Faux Négatifs.



* **TN (Vrais Négatifs)** : des prêts qui ne sont pas en défaut et ont été prédits correctement
* **TP (Vrais Positifs) :** sont en défaut et ont été prédits correctement
* **FP (Faux Positif)** : des prêts qui ne sont pas en défaut et ont été prédits de manière incorrecte
* **FN (Faux Négatif)** : des prêts qui sont en défaut et ont été prédits de manière incorrecte

Un Faux Positif (FP) constitue une perte d'opportunité pour la banque, à la différence d'un Faux Négatif (FN) qui constitue une perte pour créance irrécouvrable.

  
Ces valeurs de coefficients signifient que les Faux Négatif (FN) engendrent des pertes 10 fois supérieures aux autres.

Ces poids ont été fixés arbitrairement et il est tout à fait envisageable de les modifier en fonction de l’optique métier.

*# Total of default and not default cases*

total\_not\_default = TN + FP *# Not default cases*

total\_default = TP + FN *# Default cases*

*# Calculating the gains based on the rate/values*

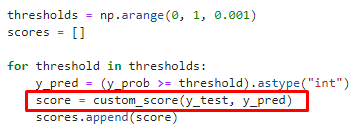
gain\_total = TN\*TN\_rate + TP\*TP\_rate + FP\*FP\_rate + FN\*FN\_rate

gain\_maximun = total\_not\_default\*TN\_rate + total\_default\*TP\_rate

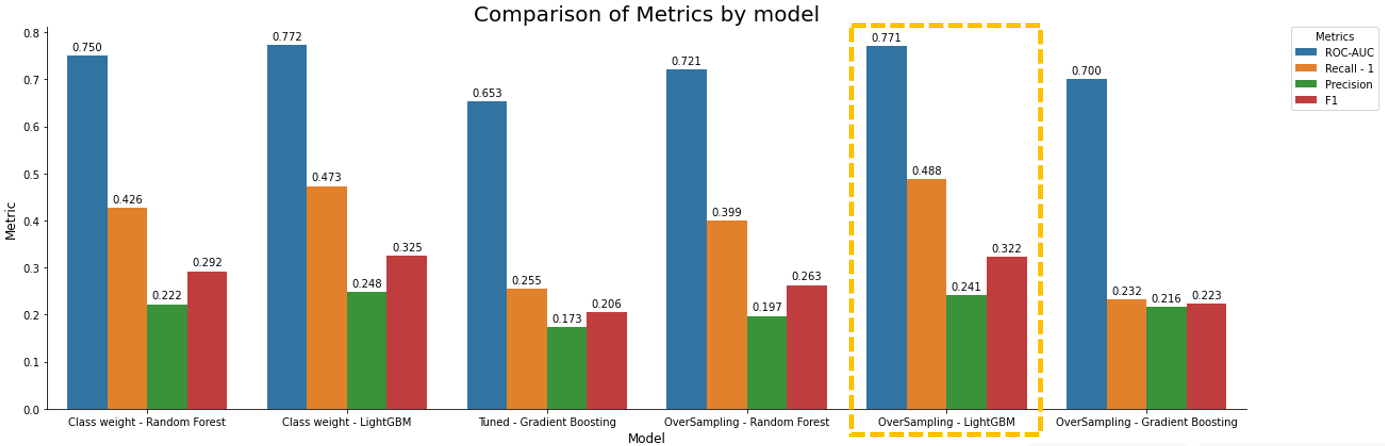
gain\_minumun = total\_not\_default\*TN\_rate + total\_default\*FN\_rate

*# Normalize to get score between 0 (baseline) and 1*

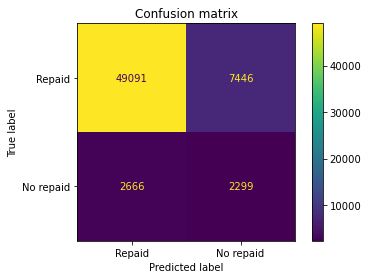
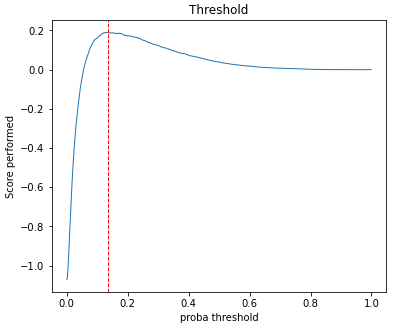
score = (gain\_total - gain\_minumun) / (gain\_maximun - gain\_minumun)

La fonction coût sera utilisé au moment de calculer le seuil

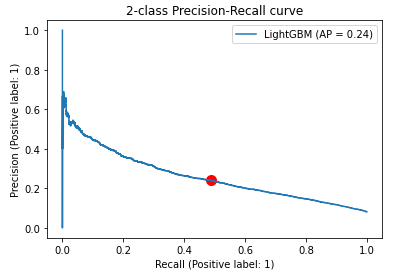
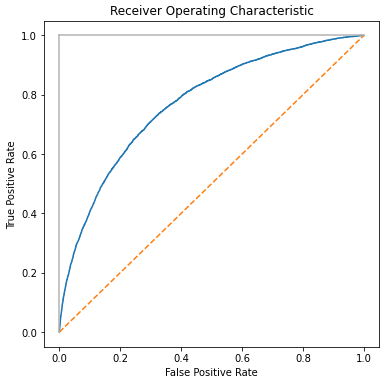
### **Résultats de modélisation**



Voici, on peut noter que les résultats sont très similaires. À gauche ceux sont de modèles avec Class weight. À droite ce sont de modèles avec OverSampling. Il faut dire que le Gradient Boosting n’a pas le paramètres Class weight. Basé sur la métrique recall - 1, on a choisi le modèle OverSampling – LightGBM pour continuer.



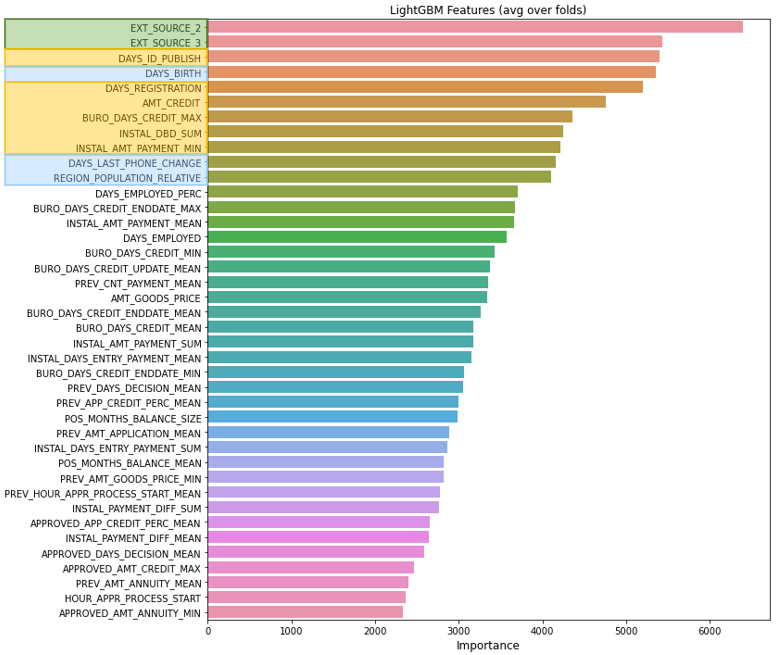
Le seuil est bas, ça veut juste dire que le modèle a tendance à prédire trop de faux négatifs et positifs



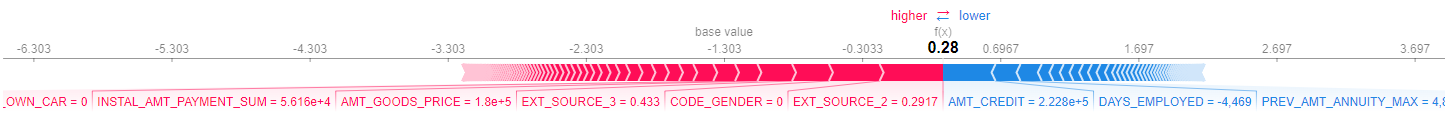
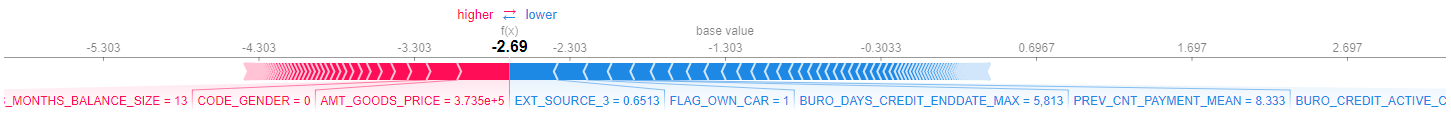
Le modèle est acceptable avec un ROC-AUC de 0,771 et un Recall pour 1 de 0,488.

# **Interprétation du modèle**

## **Interprétation global**

*  Les variables les plus importantes proviennent d'une source **externe.** Ce sont des scores normalisés provenant d'autres institutions.
* Le Feature Engineering a ajouté la valeur au moment de la modélisation. On peut le remarquer au travers des terminaisons de noms de Features (mean, max, etc.).
* Le modèle a pris en compte les différentes variables   
  ***Variables personnelles****,* ***Variables bancaires****,* ***Variables externes.***

## **Interprétation locale**

* Une observation qui est en défaut  
    
  *Nous pouvons remarquer que ext\_source\_2, ext\_source\_3 et amt\_goods\_price ont une forte influence dans le résultat pour cette observation poussant la prédiction à droite.*
* Une observation qui est en défaut

*Dans ce cas, ext\_source\_3 appuie la prédiction à gauche*

# **Limites et améliorations possibles**

* Il faut faire une analyse exploratoire au début pour bien comprendre les données.
* Il est nécessaire de faire une réduction de données pour éviter le « Curse of Dimensionality ».   
  *Il faut prendre en compte plusieurs méthodes de Feature Selection par Scikit-Learn, PCA, T-SNE, etc.*
* Aller plus loin dans l'hyperparametrisation du modèle
* Essayer d’autres algorithmes comme OneClassSVM, etc.
* Essayer plusieurs coefficients dans le custom score pour avoir un modèle plus ou moins strict

1. SMOTE (Synthetic Minority Oversampling Technique) : Consiste à synthétiser des éléments pour la classe minoritaire, à partir de ceux qui existent déjà en choisissant aléatoirement un point de la classe minoritaire et à calculer les k plus proches voisins de ce point. [↑](#footnote-ref-1)