



**FORMATION DATA SCIENTIST**

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

**Implémentez un modèle de scoring**

**Projet 7**

**Samir HINOJOSA DIAZGRANADOS**

Février 2022

Table des matières

[**1.** **Introduction** 3](#_Toc95122559)

[**2.** **Présentation des donnes** 3](#_Toc95122560)

[**3.** **Modélisation** 4](#_Toc95122561)

[a. Sélection du Kernel 4](#_Toc95122562)

[b. Optimisation du modèle 4](#_Toc95122563)

[1. Préparation des données 4](#_Toc95122564)

[2. Données déséquilibrées 4](#_Toc95122565)

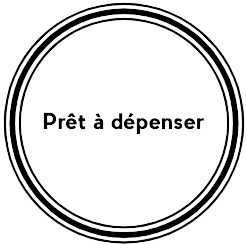
[3. Fonction coût 4](#_Toc95122566)

[4. Hyperopt 4](#_Toc95122567)

[**4.** **Interprétation du modèle** 4](#_Toc95122568)

[**5.** **Limites et les améliorations possibles** 4](#_Toc95122569)

# **Introduction**

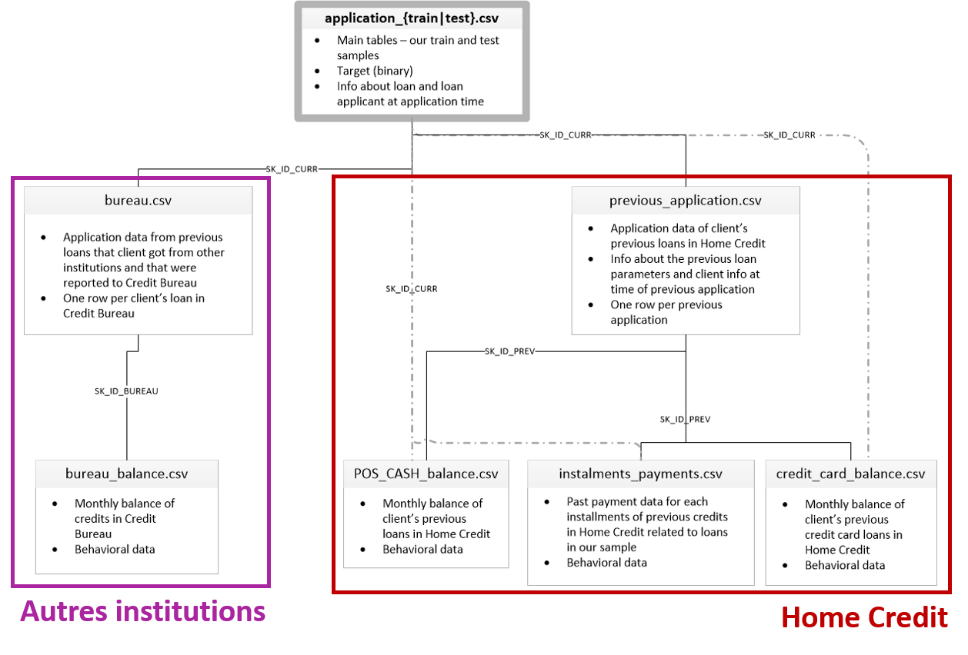
La société « **Prêt à dépenser** » souhaite mettre en œuvre un outil de « **scoring crédit** » pour calculer la probabilité qu’un client rembourse son crédit d’après diverses données.

La mission est développée un algorithme de classification 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 donnes**

****

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

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

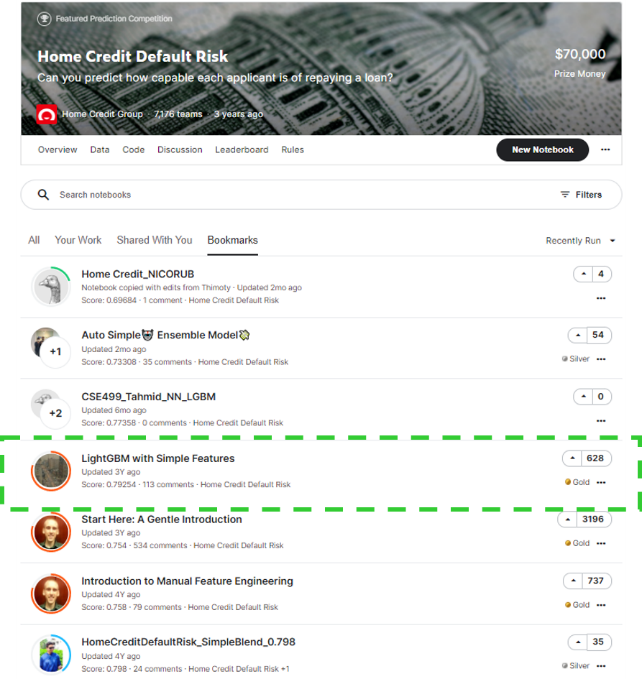
* Bureau : Antécénts de crédit du client (autres institutions financières)
* Bureau balance : Soldes mensuelles de crédits précédents

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 précédemment
* Credit\_card\_balance: Solde mensuel des cartes de crédit antérieures

# **Modélisation**

## Sélection du Kernel

Le kernel utilisé pour le projet 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
* L’opportunité d'utiliser un Framework basé sur le Gradient Boosting
* 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

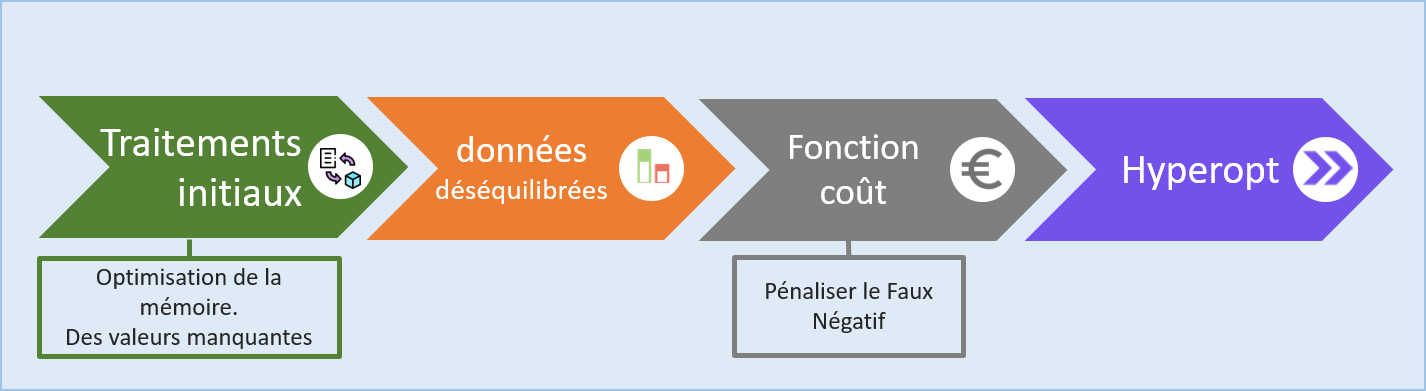
Min, Max, Mean, Sum, Var

* Unification de jeux des données

+700 variables en total

* Modélisation avec LGBMClassifier

## Optimisation du modèle



Pour faire l’optimisation du modèle, 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 le types de colonnes

### Données déséquilibrées

### Fonction coût

### Hyperopt

# **Interprétation du modèle**

# **Limites et les améliorations possibles**