

# Formation Data Scientist

Projet 7

« Implémentez un modèle de scoring »

Soutenance de Projet  
Samir HINOJOSA

10 février 2021

Prêt à dépenser

Traitements initiaux



OPENCLASSROOMS

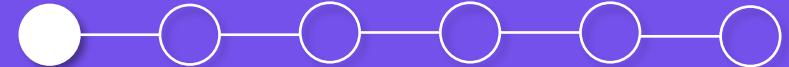
# Plan de la soutenance

1. Mission
2. Présentation du jeu de données
3. Kernel Kaggle
4. Optimisation du modèle
5. Tableau de bord
6. Conclusion



1.

# Mission



# Mission

**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



Développer un algorithme qui classifie la demande en crédit accepté ou rejeté.

Prendre en compte :

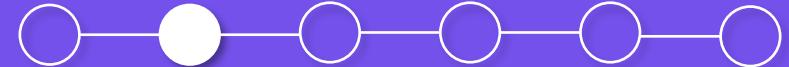
- Sélectionner un kernel Kaggle pour faciliter la préparation des données nécessaires à l'élaboration du modèle de scoring.
- Déployer le modèle de scoring comme une API.
- Construire un tableau de bord interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions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



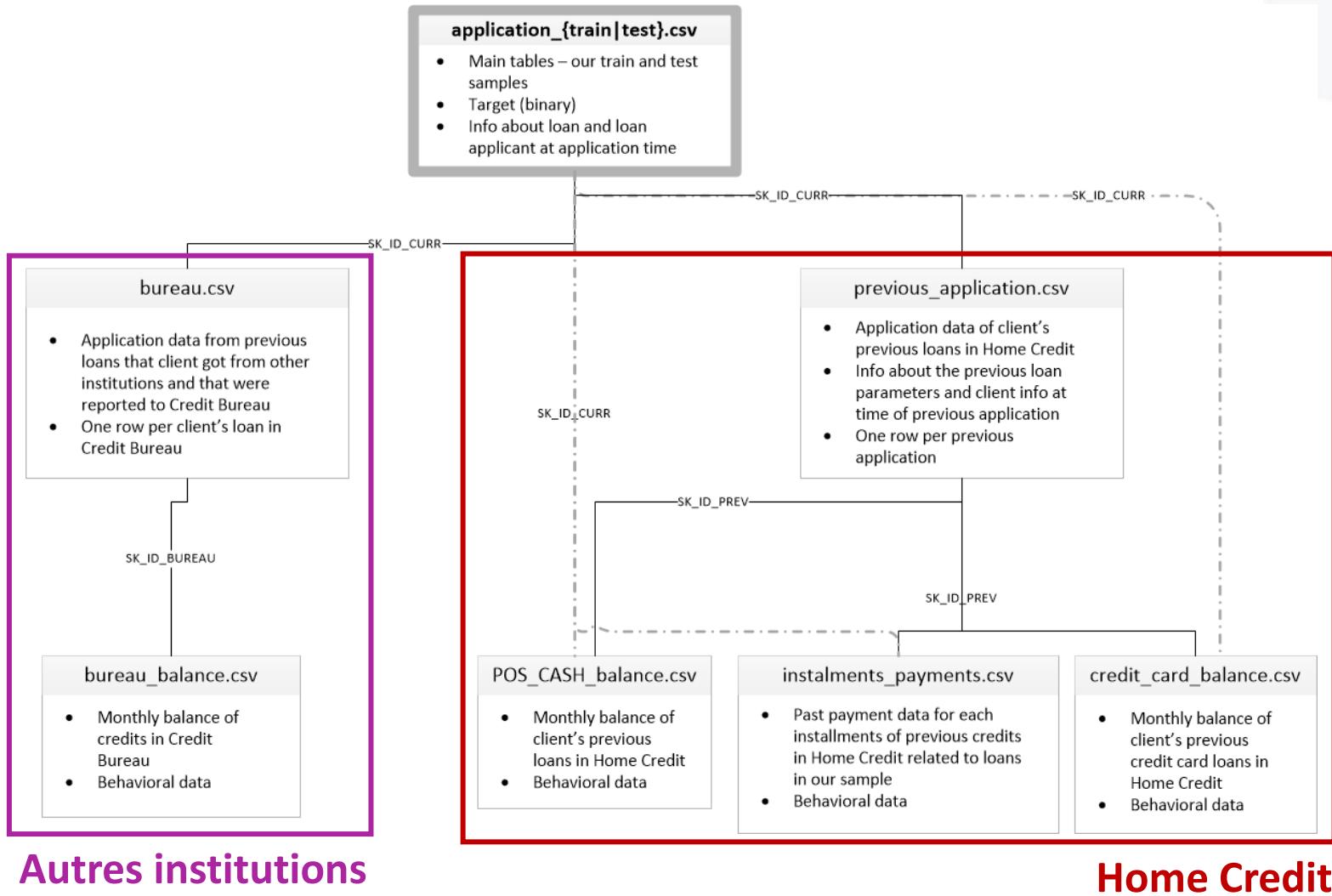
**HOME  
CREDIT**

2.

## Présentation du jeu de données



# Présentation du jeu de données<sup>1</sup>



## Généralités

- 7 sources de données
- 307511 clients
- 121 features  
âge, sexe, emplois, revenus, crédits précédents, etc.
- 24% valeurs manquantes

1 - <https://www.kaggle.com/c/home-credit-default-risk/data>

3.

## Kernel Kaggle



# Kernel Kaggle

Le kernel utilisé pour le projet est  
LightGBM with Simple Features<sup>1</sup>



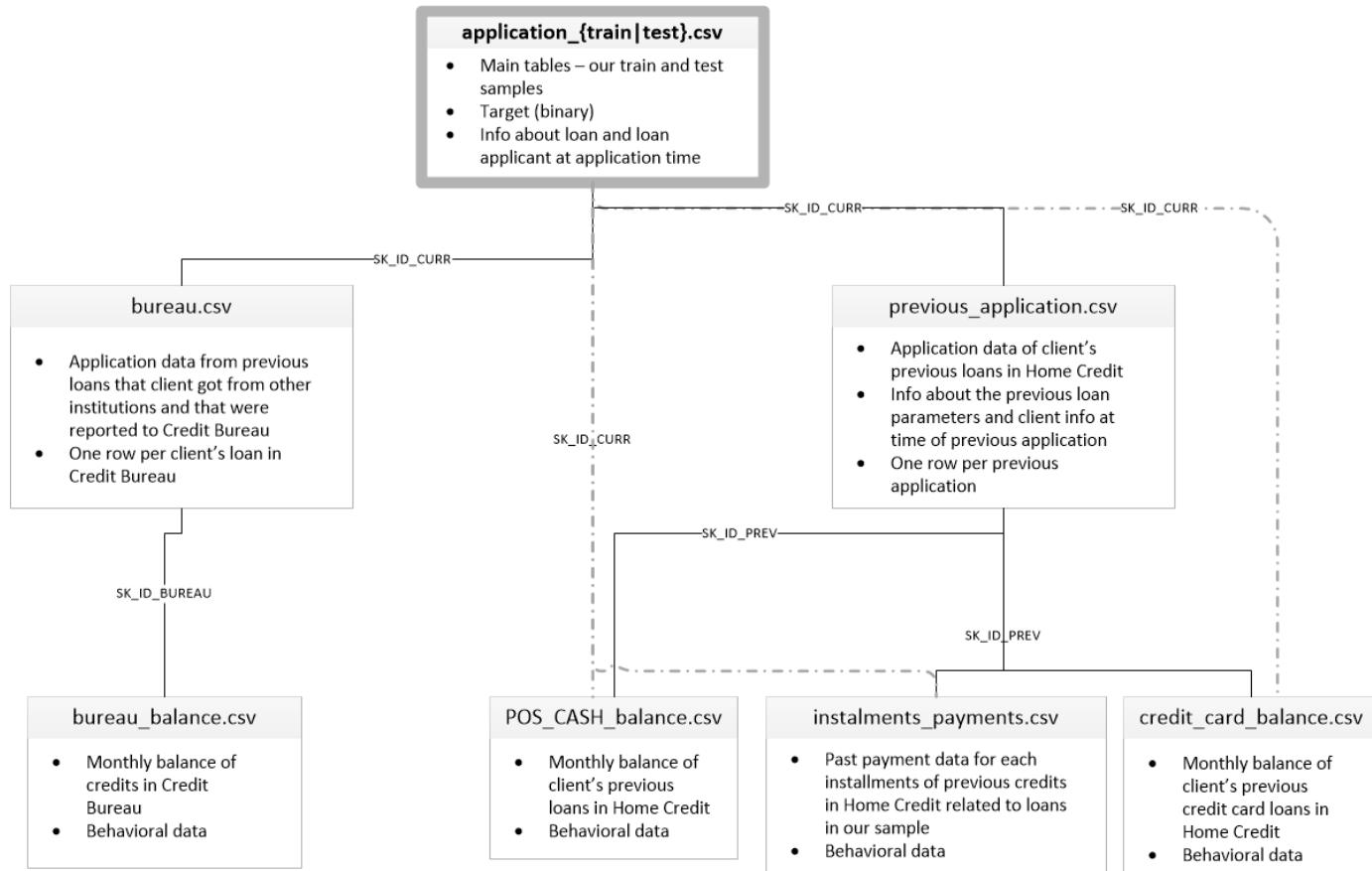
## Basé sur

- Un bon score dans la compétition
- Un Feature Engineering performant
- L'opportunité d'utiliser un Framework basé sur le Gradient Boosting - LightGBM
- Le kernel recommandé dans le projet

Rank	User	Notebook Name	Score	Comments	Prize
1	jsaguiar	LightGBM with Simple Features	0.79254	113	Gold
2	Thimoty	Home Credit_NICORUB	0.69684	1	Silver
3	AutoSimple	Auto Simple Ensemble Model	0.73308	35	Silver
4	CSE499_Tahmid	CSE499_Tahmid_NN_LGBM	0.77358	0	
5	StartHere	Start Here: A Gentle Introduction	0.754	534	Gold
6	ManuelFeatureEngineering	Introduction to Manual Feature Engineering	0.758	79	Gold
7	SimpleBlend	HomeCreditDefaultRisk_SimpleBlend_0.798	0.798	24	Silver

1 - <https://www.kaggle.com/jsaguiar/lightgbm-with-simple-features>

# LightGBM with Simple Features



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



Qu'est-ce qu'il fait ?

- Identification et traitement des variables catégorielles
- Création de nouvelles variables
  - $\text{DAYS_EMP\_ \%} = \text{DAYS\_EMP} / \text{DAYS\_BIRTH}$
  - Min, Max, Mean, Sum, Var
- Unification de jeux des données
  - +700 variables en total
- Modélisation avec LGBMClassifier

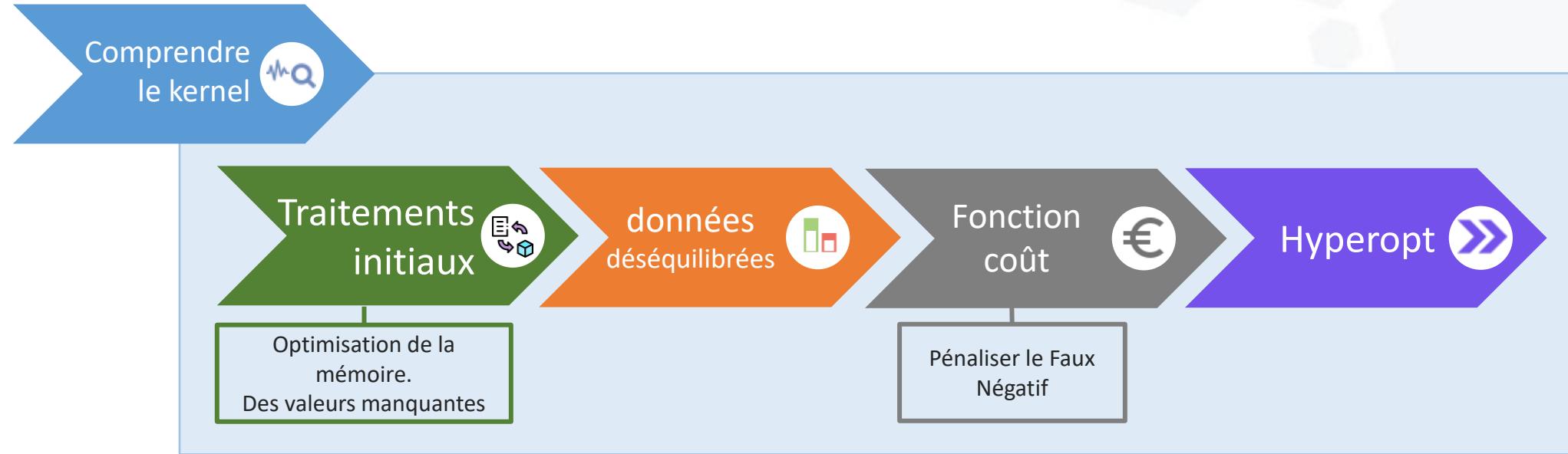


4.

## Optimisation du modèle



# Optimisation du modèle



Sera pris en compte le Feature Engineering déjà fait dans le kernel choisi



# Traitements initiaux



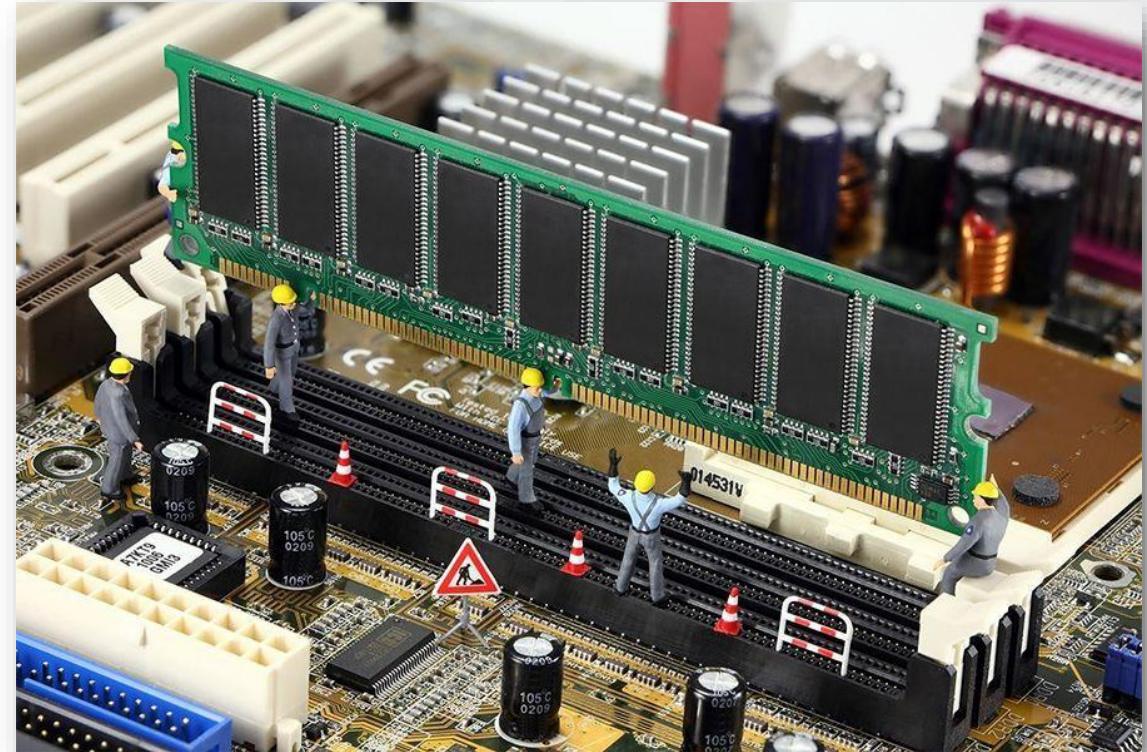
## Optimisation de la mémoire

- Reduction de types de colonnes (Float64 à float32)
  - Utilisation de la mémoire :  
2,1 GB à 941 MB



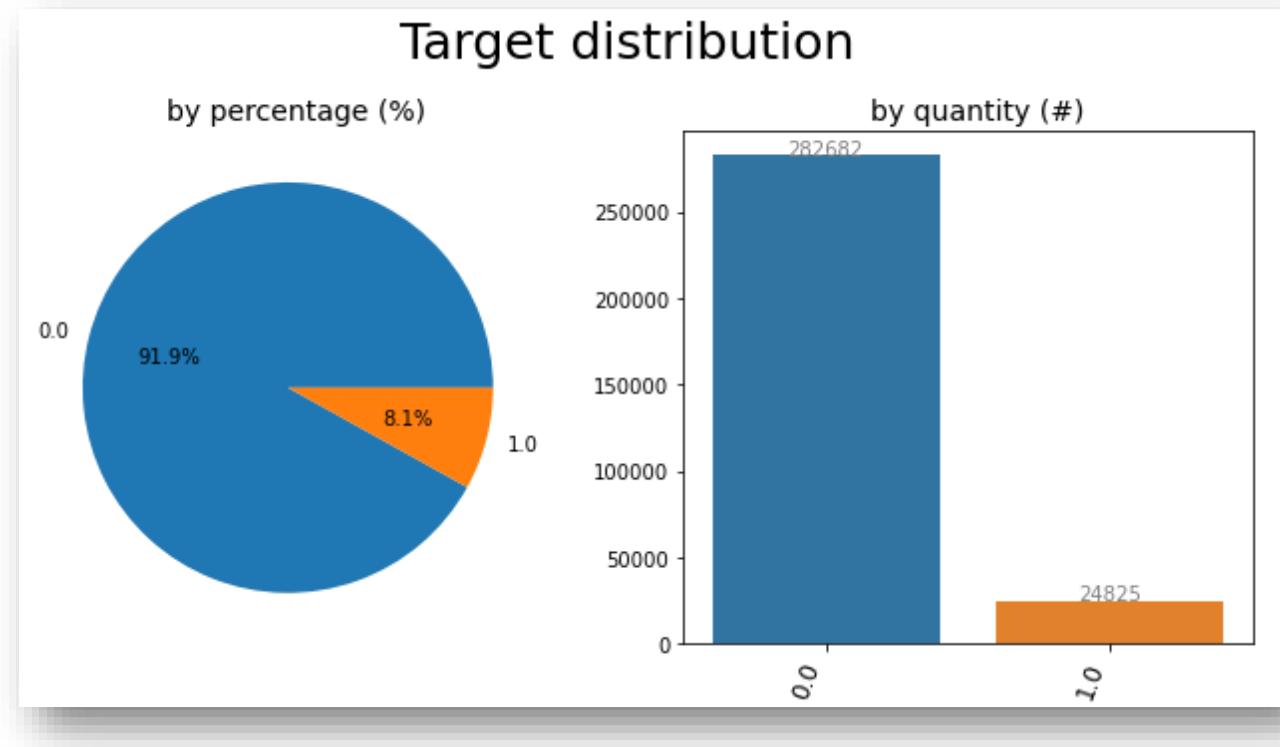
## Traitements de valeurs manquantes

- Traitements des valeurs infinies
  - Imputation de valeur manquante
- Traitements des valeurs manquantes
  - Imputation de la moyenne basée sur la colonne



# Le jeu de données est déséquilibré

données  
déséquilibrées

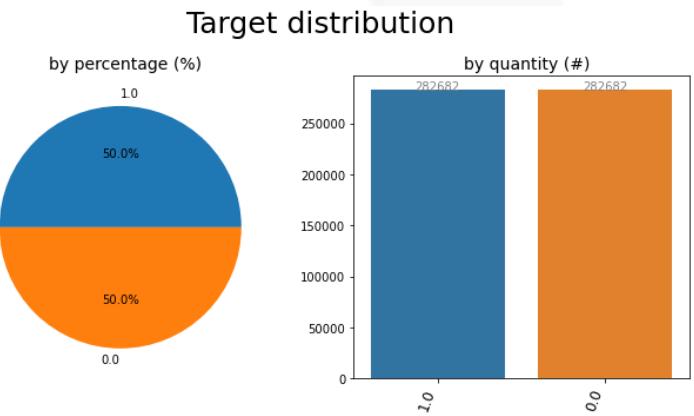


- (0) Ce sont des prêts qui ont été remboursés.
- (1) Ce sont des prêts qui n'ont pas été remboursés



## Utilisation du Oversampling - SMOTE

- pour augmenter les données dans la classe minoritaire



# Pénaliser des Faux Négatifs

Fonction  
coût



		Classe réelle
		0
Classe prédictive	0	TN
1	FP	
		1
FN		TP



## Fonction coût

Taux	Valeur
TN (vrais négatifs)	1
TP (vrais positifs)	1
FP (faux positifs)	0
FN (faux négatifs)	-10

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.

```
# Total of default and not default cases
total_not_default = TN + FP      # Not default cases
total_default = TP + FN          # Default cases

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)
```

# Optimisation des hyperparamètres

Hyperopt 

Paramètres basés sur les paramètres déjà existants  
dans le kernel choisi

```
N_ESTIMATORS = [8000, 10000, 12000]
NUM_LEAVES = [32, 34, 36]
MAX_DEPTH = [7, 8, 9]
```

```
space_params = {
    "n_estimators" : hp.choice("n_estimators", N_ESTIMATORS),
    "learning_rate" : hp.uniform("learning_rate", 0.002, 0.003),
    "num_leaves" : hp.choice("num_leaves", NUM_LEAVES),
    "max_depth" : hp.choice("max_depth", MAX_DEPTH),
    "reg_alpha" : hp.uniform("reg_alpha", 0.041545473, 0.051),
    "reg_lambda" : hp.uniform("reg_lambda", 0.0735294, 0.0835294),
    "min_split_gain" : hp.uniform("min_split_gain", 0.0222415, 0.0322415),
    "min_child_weight" : hp.uniform("min_child_weight", 39.3259775, 49)
}
```



Pour optimiser la performance et le temps

- StandardScaler()
- LGBMClassifier
  - colsample\_bytree=0.8
  - subsample=0.8
  - Is\_unbalance=False

Le kernel a déjà un cross-validation mis en œuvre

4.a.

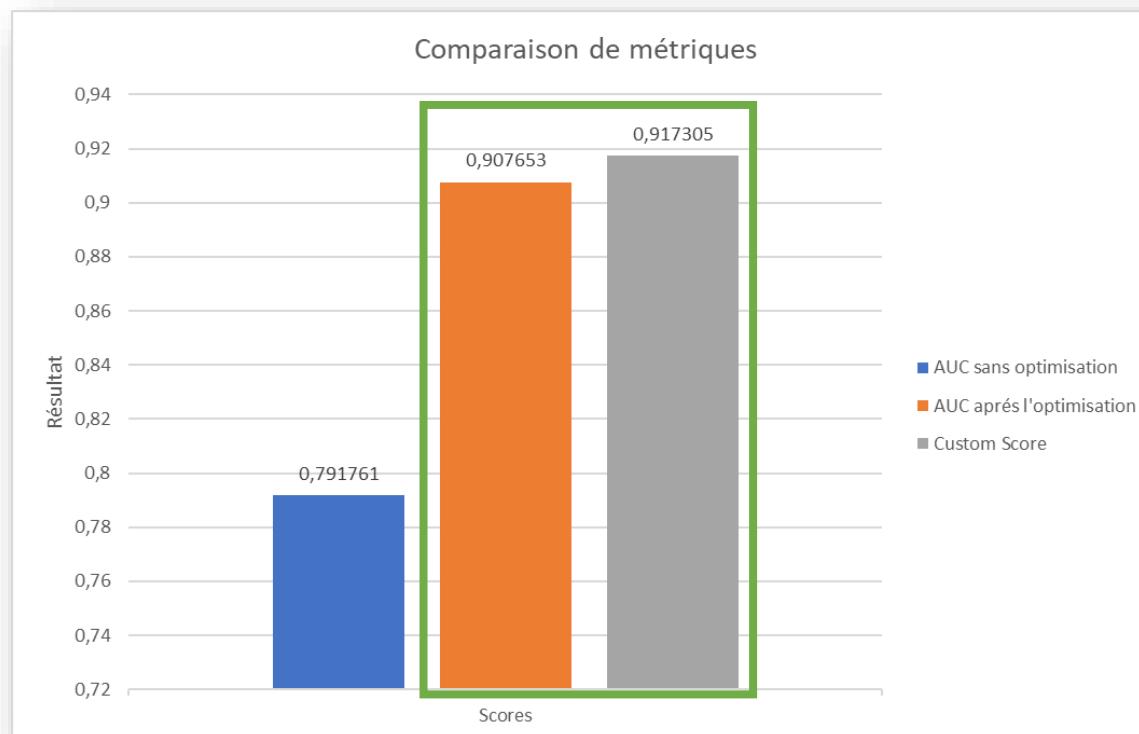
## Résultats d'optimisation



# Optimisation des hyperparamètres

Hyperopt ➤

Les résultats obtenus après avoir fait l'optimisation sont meilleurs que ceux du précédent



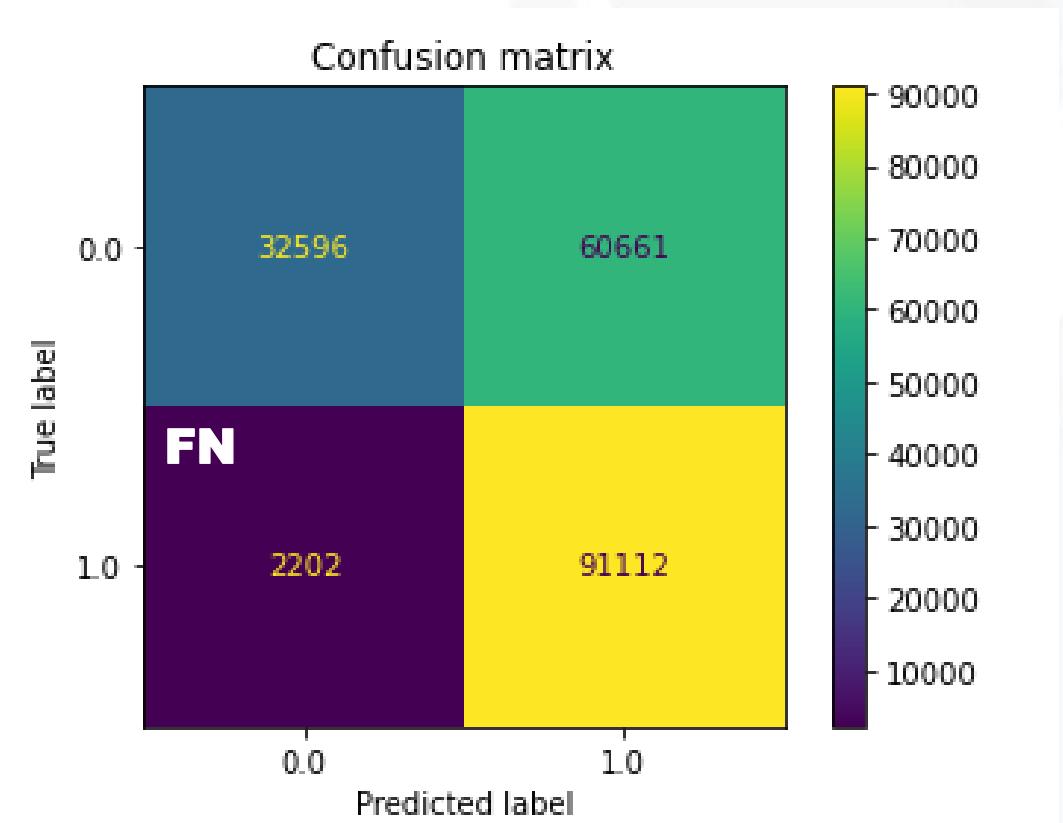
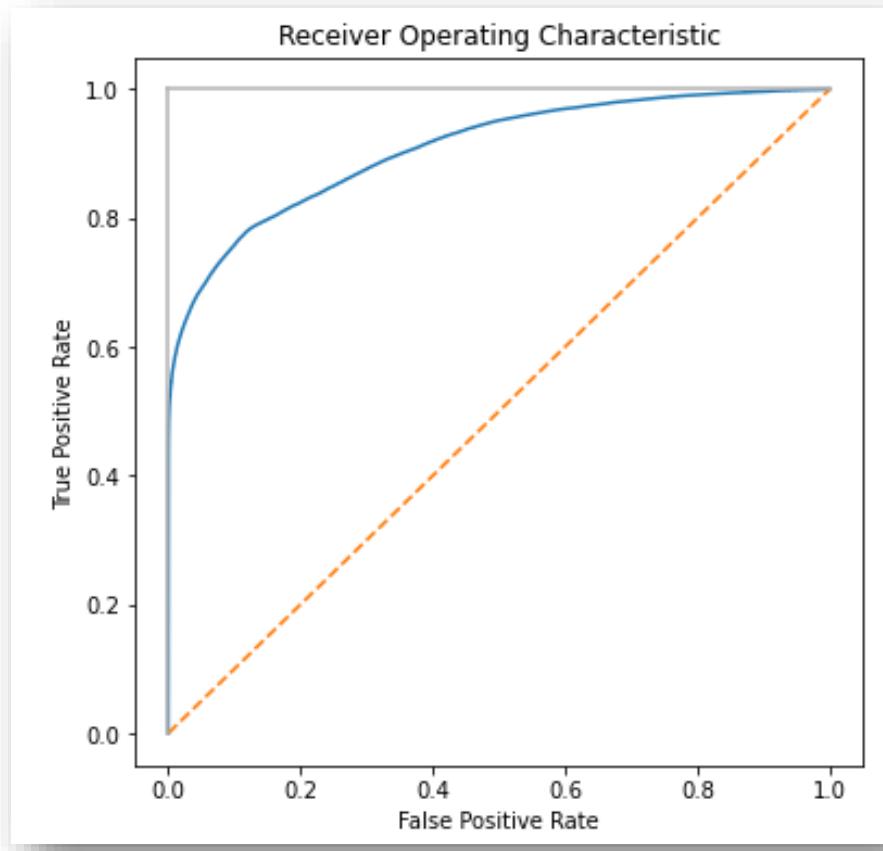
Meilleur résultat avec Hyperopt

Paramètres	Meilleur résultat
learning_rate	0.002021947556803579
max_depth	9
min_child_weight	44.68618422455195
min_split_gain	0.030970825122649367
n_estimators	8000
num_leaves	36
reg_alpha	0.045341569610647205
reg_lambda	0.08049459639521307

Paramètres basés sur ceux déjà existants dans le kernel choisi

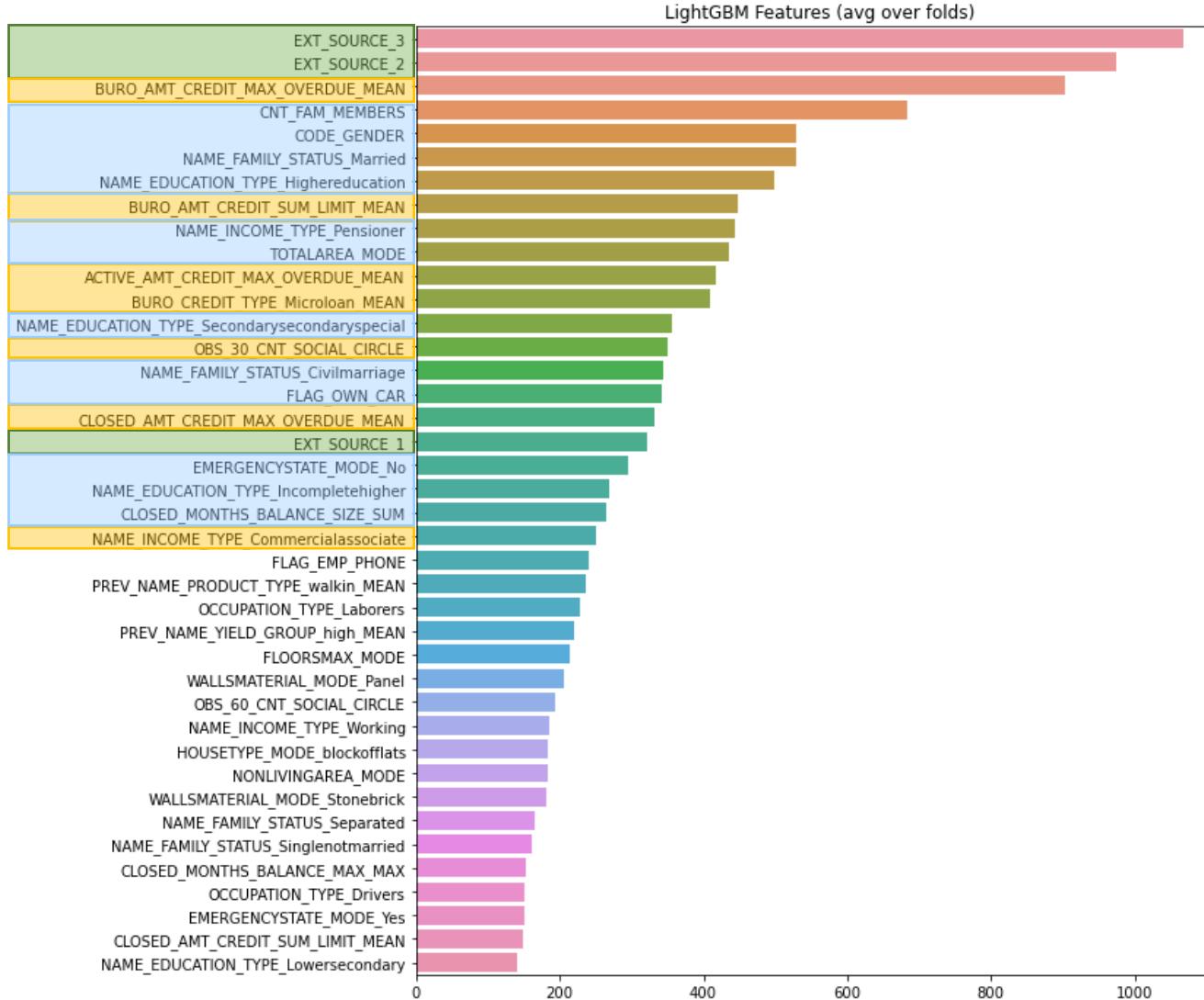
# Le modèle est performant

Parce qu'il s'agit d'un ensemble de données équilibré,  
ROC-AUC sera pris en compte. ROC-AUC est 0,907633



# Interprétation du modèle

Hyperopt ➤



Les variables les plus importantes proviennent d'une source externe



Le Feature Engineering a ajouté la valeur au moment de la modélisation



Pris en compte des différentes variables.

*Variables personnelles, Variables bancaires, Variables externes*

5.

## Tableau de bord



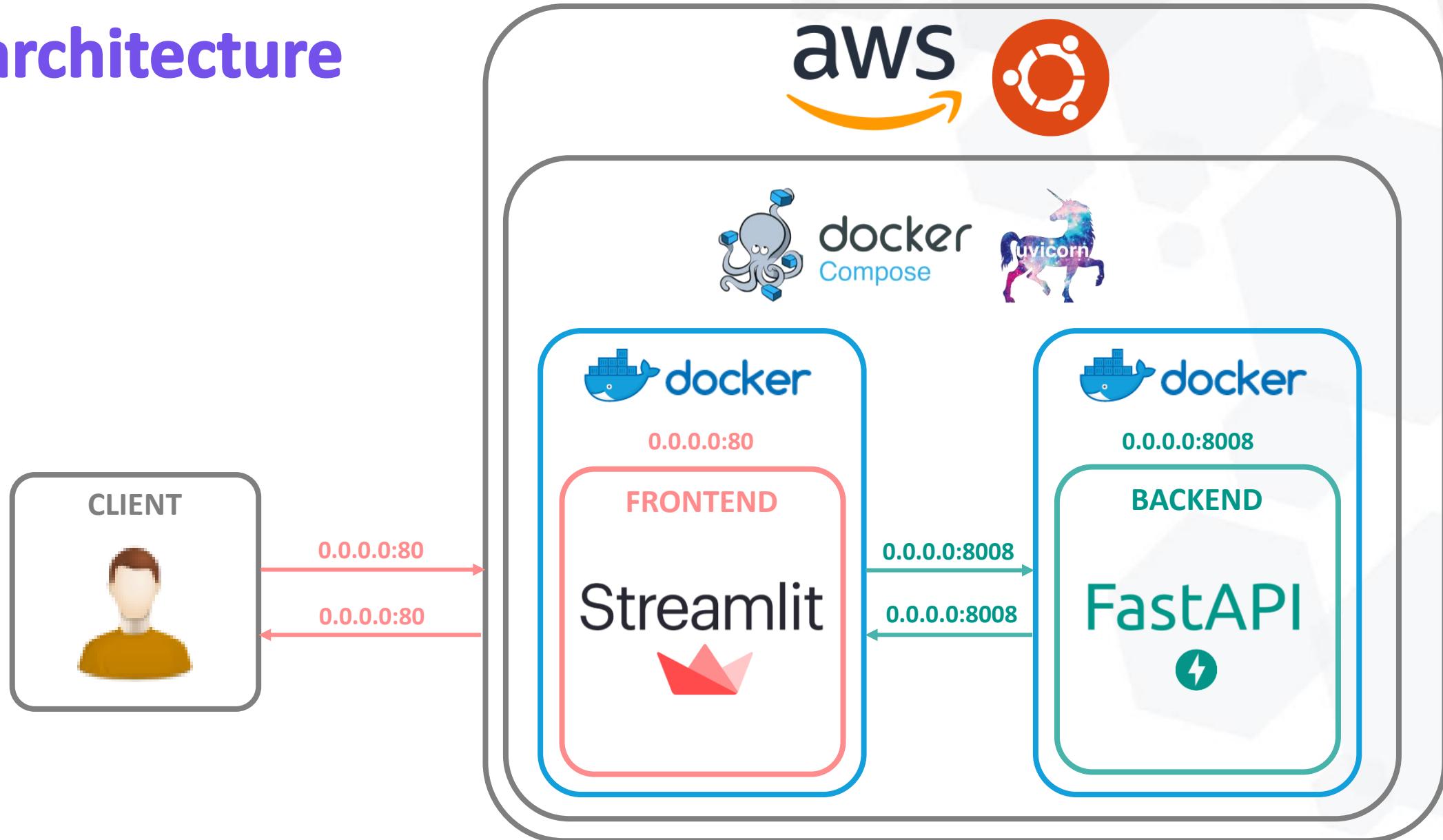
# Outils utilisés



Dashboard: <http://home-credit.samirhinojosa.com>

Repository: <https://github.com/samirhinojosa/OC-P7-implement-a-scoring-model>

# L'architecture



# EndPoints - API

## Home Credit Default Risk 0.1.0 OAS3

[/openapi.json](#)

Obtain information related to probability of a client defaulting on loan.

### default

`GET /api/clients` Clients Id

`GET /api/clients/{id}` Clients

`GET /api/predictions/clients/{id}` Predict

`GET /api/statistics/amtIncomes` Statistical Amt Income

`GET /api/statistics/amtCredits` Statistical Amt Credit

`GET /api/statistics/ages` Statistical Age

`GET /api/statistics/yearsEmployed` Statistical Years Employed

### Schemas

`HTTPValidationError` >

`ValidationError` >

Swagger : <http://home-credit.samirhinojosa.com:8008/docs> 

# Tableau de bord

## Prêt à dépenser - Default Risk

### Client selection

Select a client to obtain information related to **probability** of a client **paying the loan**. In addition, you can analyze some stats.

Client id list

139321

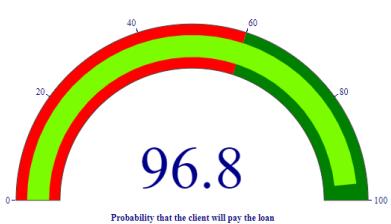
See stats

See stats will take more time. Are you sure ?

Predict

### Client information

Based on the client's information, the credit application is **accepted!**



Client id:

139321

Gender:

Woman

Age:

64

Children:

0

Own realty:

No

Own car:

No

Years employed:

6

Anual income:

\$ 121,500.00

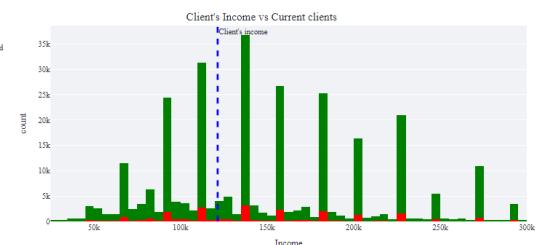
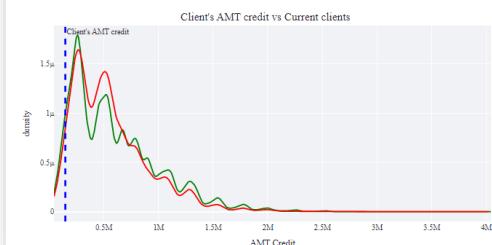
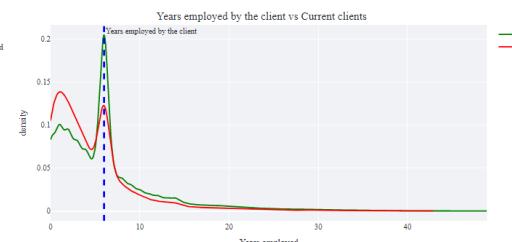
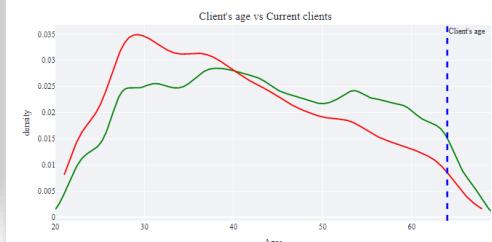
AMT credit:

\$ 148,365.00

## Client avec crédit accepté

### General statistics

Below, you can see some general statistics about clients who repaid and do not repay the loan



Dashboard: <http://home-credit.samirhinojosa.com>

# Tableau de bord

## Prêt à dépenser - Default Risk

### Client selection

Select a client to obtain information related to probability of a client paying the loan. In addition, you can analyze some stats.

Client Id list

455326

See stats

See stats will take more time. Are you sure ?

Predict

### Client information

Based on the client's information, the credit application is not accepted!



Probability that the client will pay the loan

Client id:

455326

Gender:

Man

Age:

42

Children:

1

Own realty:

Yes

Own car:

Yes

Years employed:

11

Anual income:

\$ 112,500.00

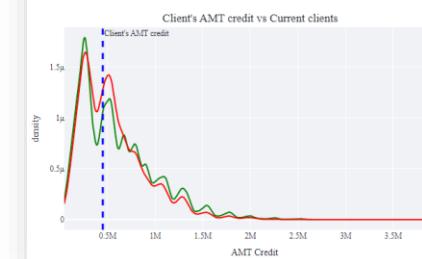
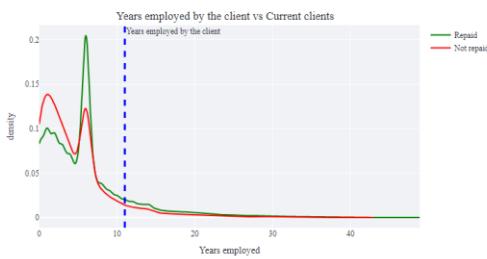
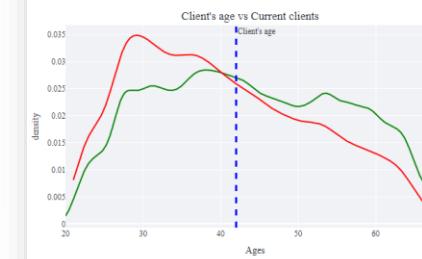
AMT credit:

\$ 450,000.00

## Client avec crédit rejeté

### General statistics

Below, you can see some general statistics about clients who repaid and do not repay the loan



Dashboard: <http://home-credit.samirhinojosa.com>

6.

## Conclusion



# Les conclusions de la mission



## Le modèle

- Il faut faire une analyse exploratoire au début pour bien comprendre les données.
- Il est nécessaire de faire un réduction de données pour éviter le « *Curse of Dimensionality* »  
*Il faut prendre en compte une méthode comme Feature Selection par Scikit-Learn*
- Une analyse du composant principal « PCA » peut apporter des avantages
- Aller plus loin dans l'hyperparamétrisation du modèle



## Tableau du bord

- Permettre la sélection de diverses variables au moment de la réalisation du graphique
- Ajouter des informations sur les variables sélectionnées

# Avez-vous des questions ?



**MERCI**

Soutenance de Projet  
Samir HINOJOSA

10 février 2022



**OPENCLASSROOMS**

# Annexes