

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. Feature Engineering / Kernel Kaggle
4. Modélisations effectuées
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



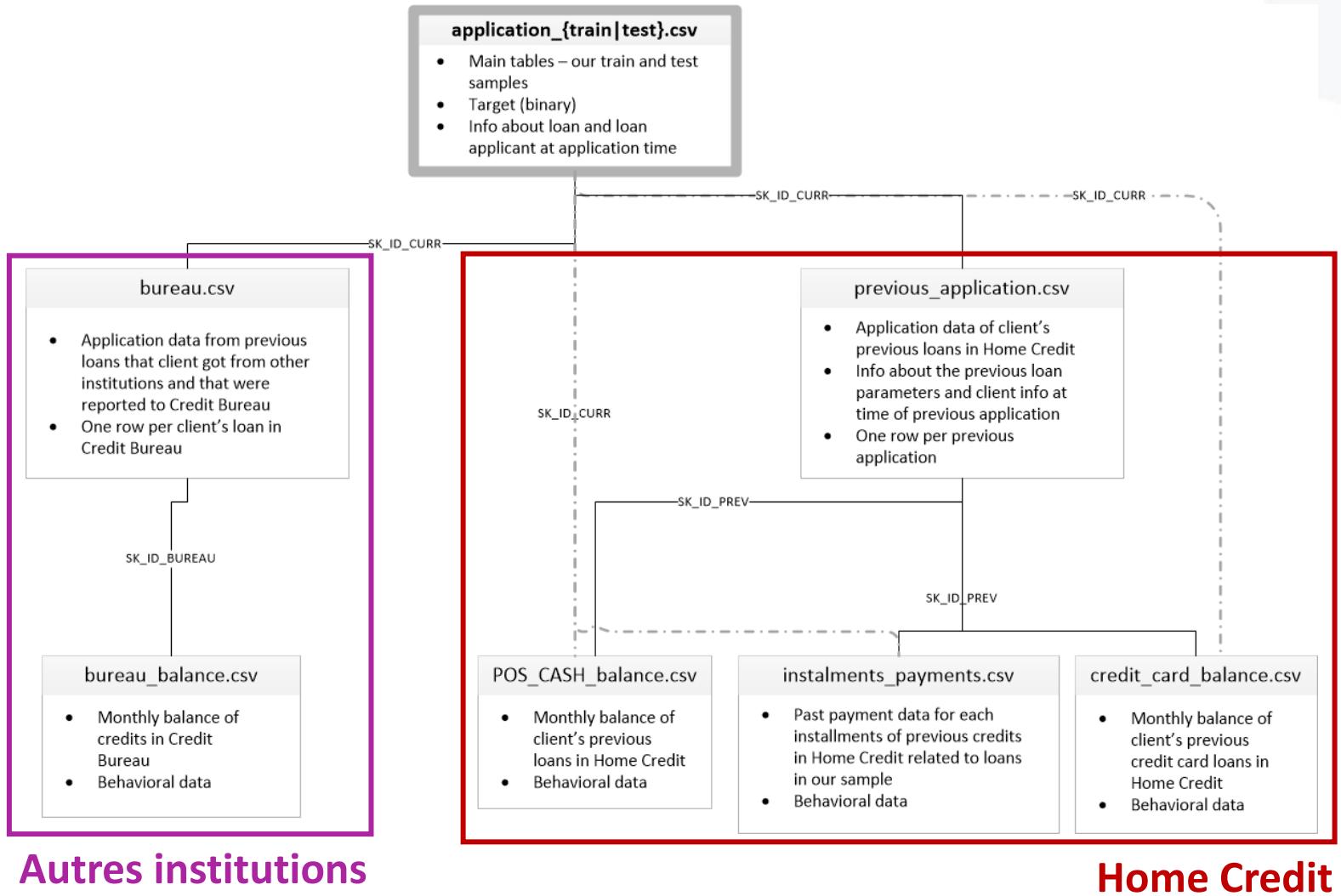
**HOME
CREDIT**

2.

Présentation du jeu de données



Présentation du jeu de données¹



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.

Feature Engineering / Kernel Kaggle



Feature Engineering / Kernel Kaggle

Le kernel utilisé pour le preprocessing est
LightGBM with Simple Features¹



Basé sur

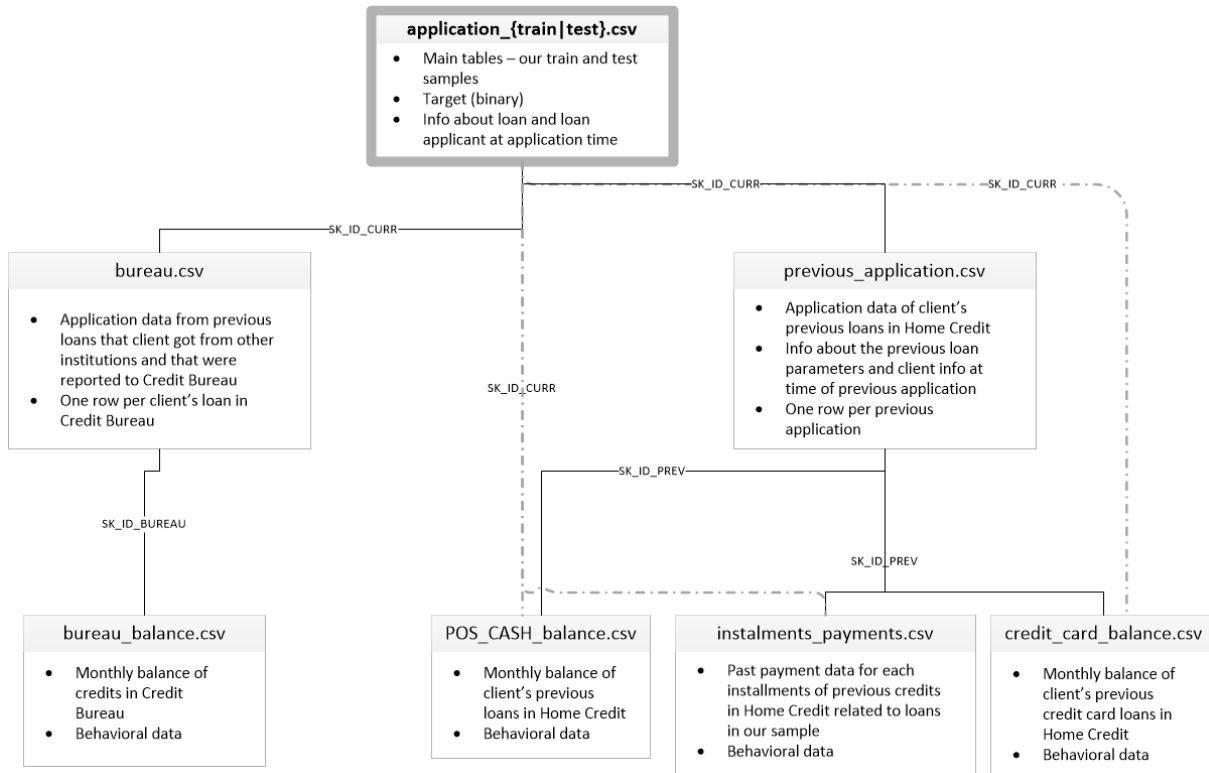
- Un bon score dans la compétition
- Un Feature Engineering performant
- Le kernel recommandé dans le projet

The screenshot shows the Kaggle competition interface for 'Home Credit Default Risk'. At the top, it displays the competition title, a featured prediction competition, and a \$70,000 prize money. Below the title, it shows the Home Credit Group, 7,176 teams, and 3 years ago. A navigation bar includes Overview, Data, Code, Discussion, Leaderboard, Rules, New Notebook, and more. A search bar allows searching for notebooks, and a filters button is available. The main area lists notebooks, with the 'Bookmarks' tab selected. The list includes:

- Home Credit_NICORUB (Notebook copied with edits from Thimoty · Updated 2mo ago · Score: 0.69684)
- Auto Simple Ensemble Model (Updated 2mo ago · Score: 0.73308)
- CSE499_Tahmid_NN_LGBM (Updated 6mo ago · Score: 0.77358)
- LightGBM with Simple Features** (Selected item, updated 3Y ago · Score: 0.79254 · 113 comments)
- Start Here: A Gentle Introduction (Updated 3Y ago · Score: 0.754 · 534 comments)
- Introduction to Manual Feature Engineering (Updated 4Y ago · Score: 0.758 · 79 comments)
- HomeCreditDefaultRisk_SimpleBlend_0.798 (Updated 4Y ago · Score: 0.798 · 24 comments)

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 ?

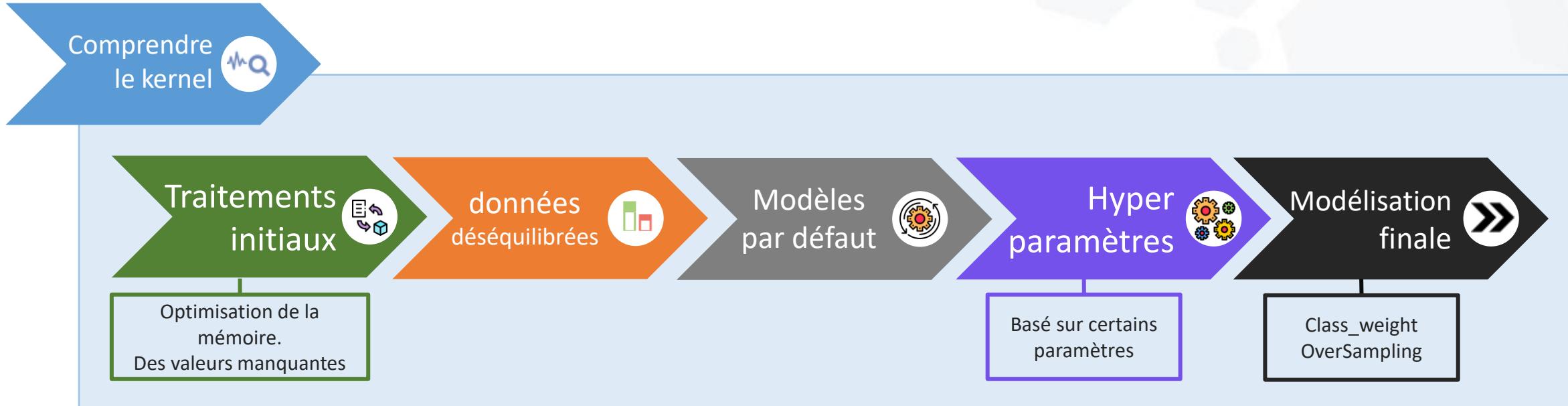
- Identification et traitement des variables catégorielles
- Crédit de nouvelles variables
 - $\text{DAYS_EMP_ \%} = \text{DAYS_EMP} / \text{DAYS_BIRTH}$
 - Min, Max, Mean, Sum, Var
- Unification de jeux des données
 - +700 variables en total

4.

Modélisations effectuées



Modélisations effectuées



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

Traitements initiaux



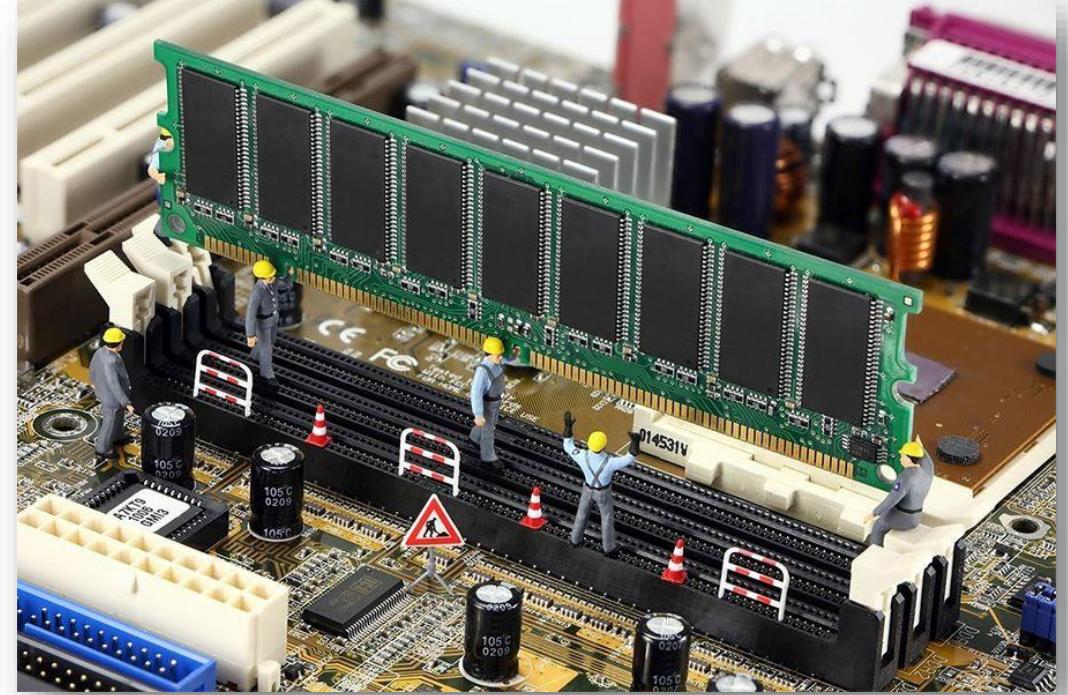
Optimisation de la mémoire

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



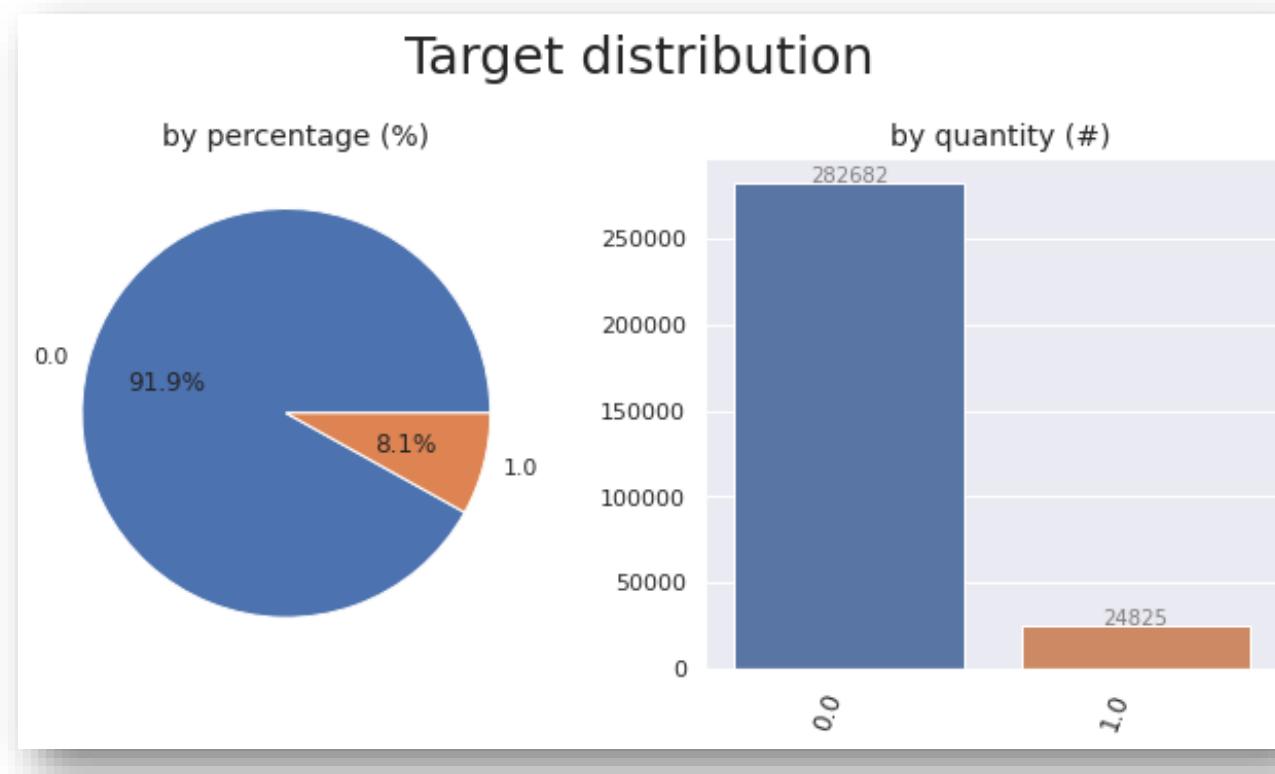
Traitements de valeurs manquantes

- Enlever les colonnes avec 20 % ou plus 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 de class weight

- pour faire une pondération inversement proportionnellement à la fréquence des classes



Utilisation du Oversampling - SMOTE

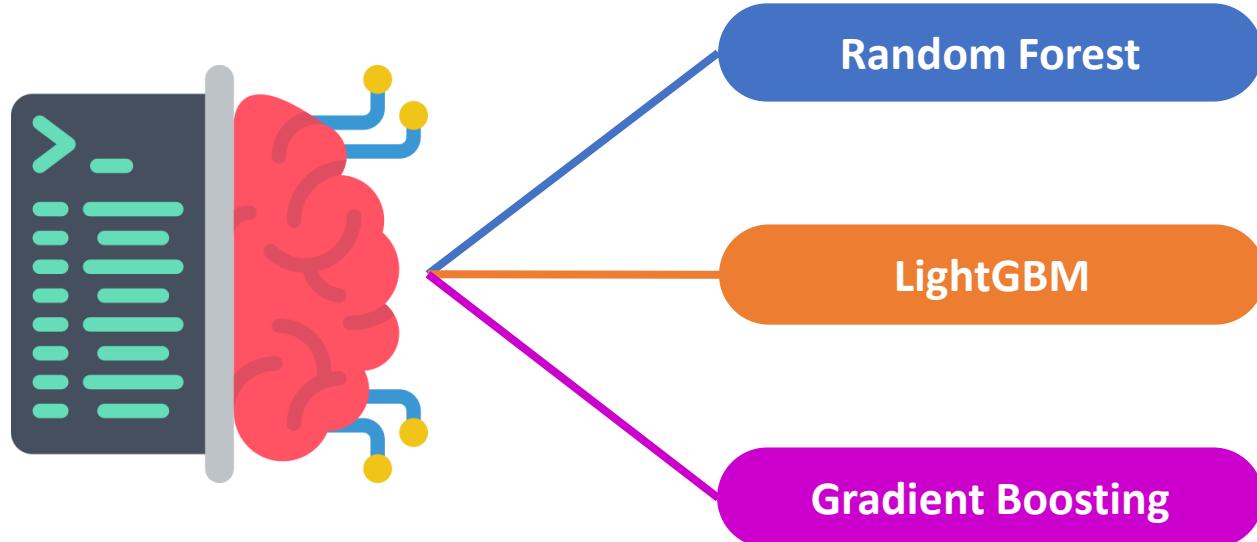
- pour augmenter les données dans la classe minoritaire

Les modèles par défaut

Modèles
par défaut



Un problème de classification déséquilibré

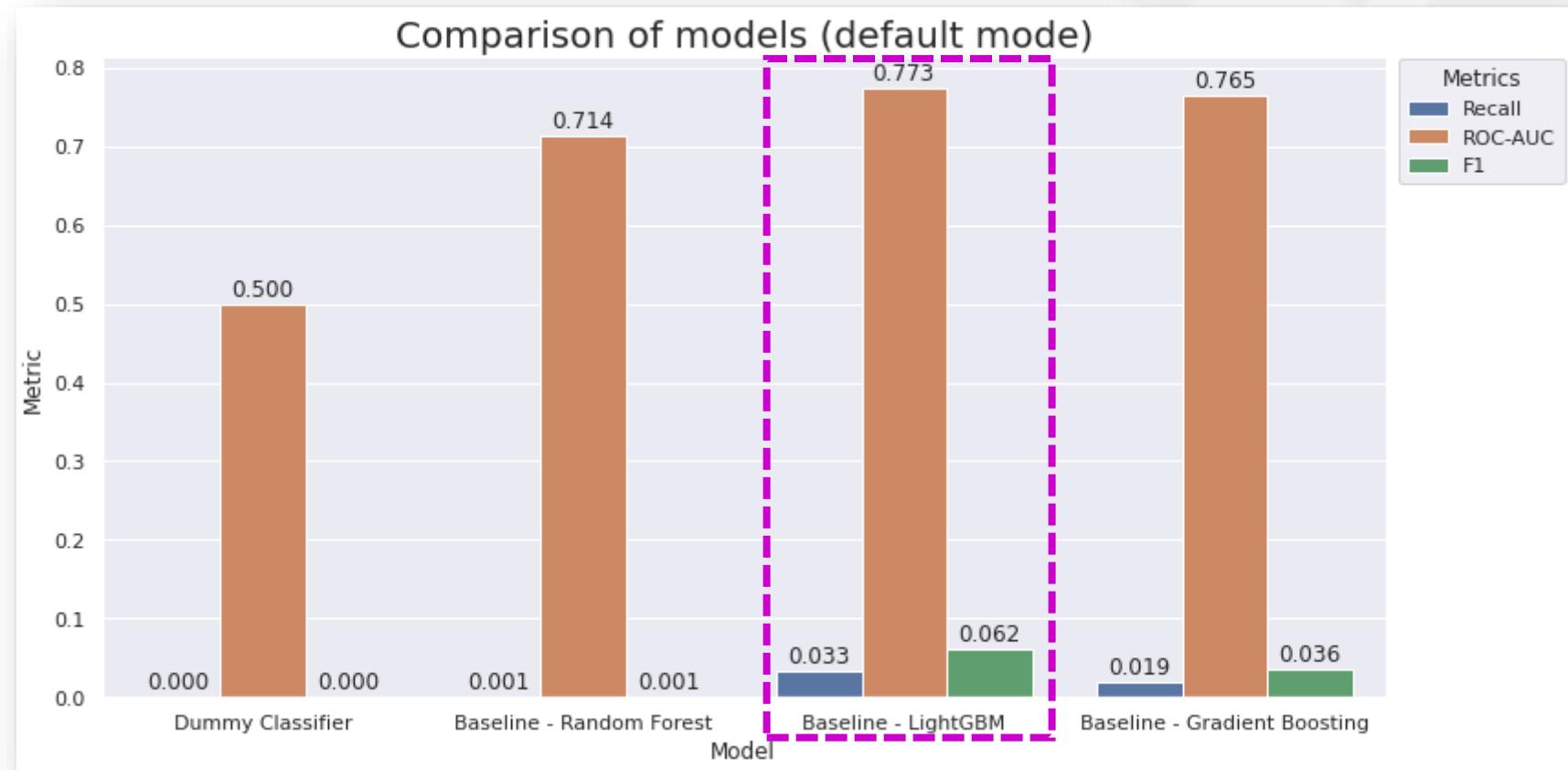


Métriques sélectionnées

- ROC-AUC
- Recall
- F1
- Precision
- Custom Score (fonction coût)
- Duration



En mode par défaut, le LightGMB a le meilleur résultat



Pénaliser des Faux Négatifs

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.

Matrice de confusion

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



Fonction coût

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

```
# 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

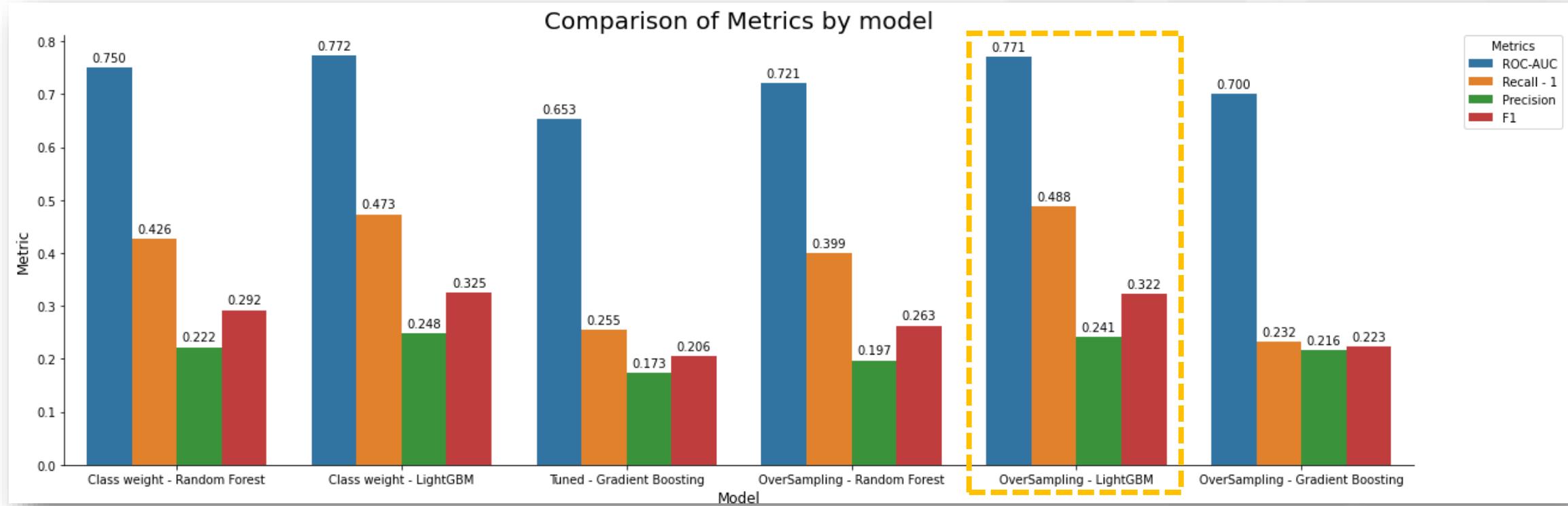
# normalize to get score between 0 (baseline) and 1
score = (gain_total - gain_minumun) / (gain_maximun - gain_minumun)

thresholds = np.arange(0, 1, 0.001)
scores = []

for threshold in thresholds:
    y_pred = (y_prob >= threshold).astype("int")
    score = custom_score(y_test, y_pred)
    scores.append(score)
```

La fonction coût sera utilisé au moment de calculer le seuil

Les résultats sont très similaires



Le Gradient Boosting n'a pas le paramètres Class weight

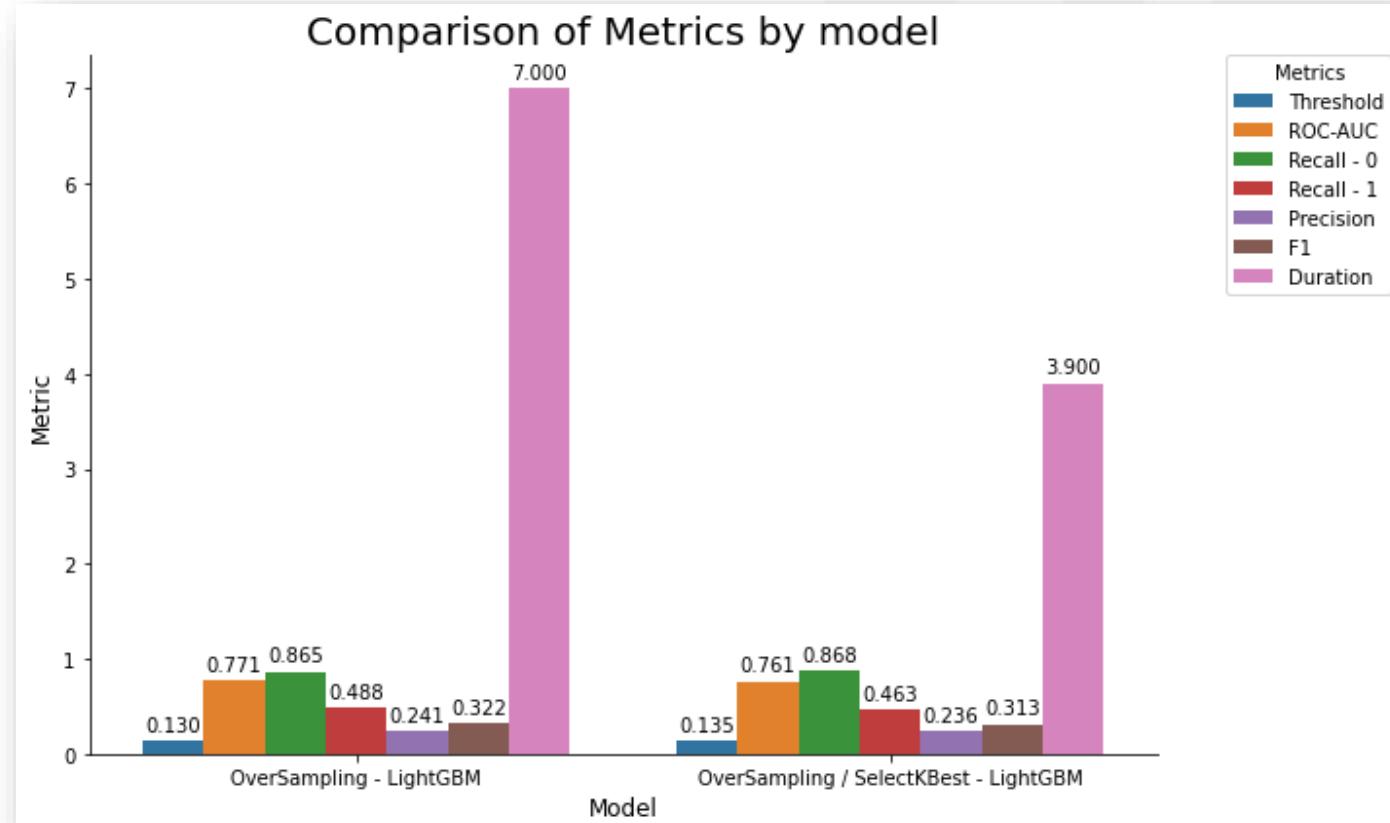
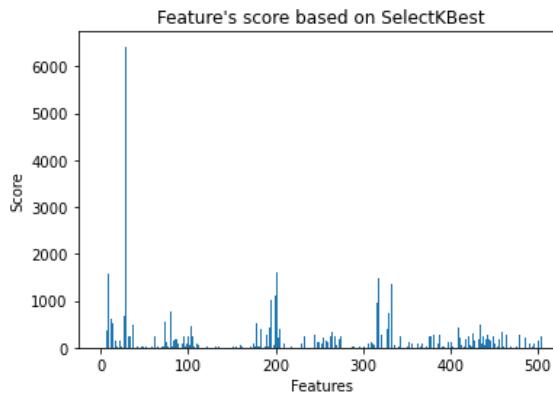
4.a.

Feature selection



Il y a un réduction de temps

SelectKBest k=150

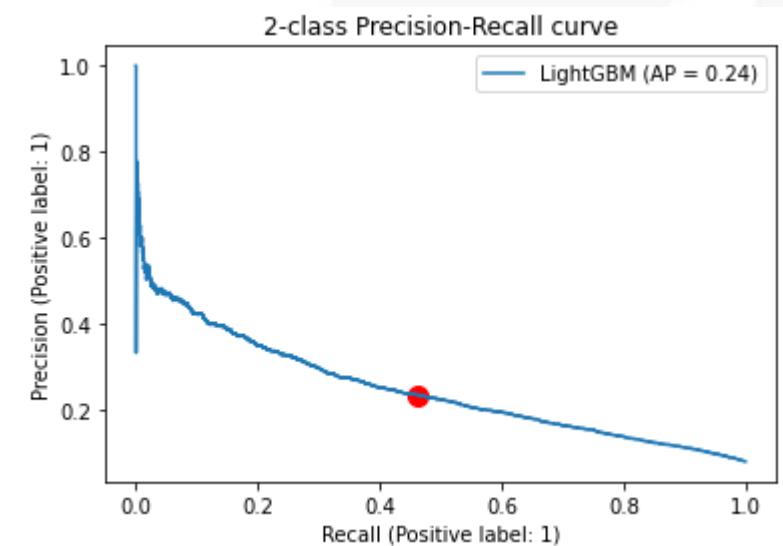
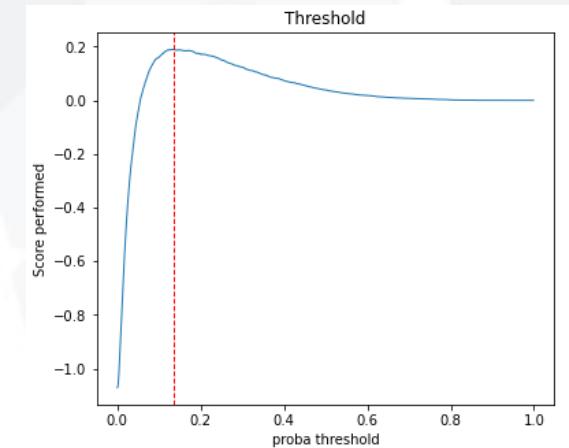
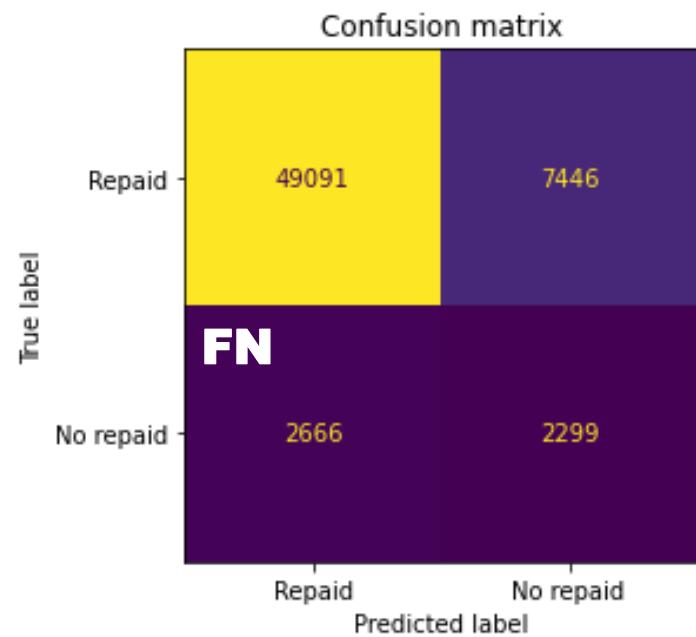
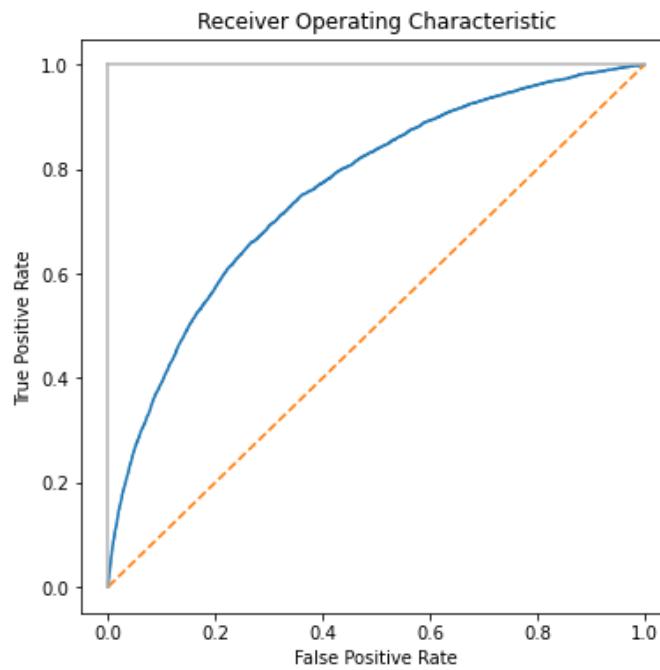


Le modèle est acceptable

Parce qu'il s'agit d'un ensemble de données équilibré,

ROC-AUC est 0,771

Recall-1 est 0,488

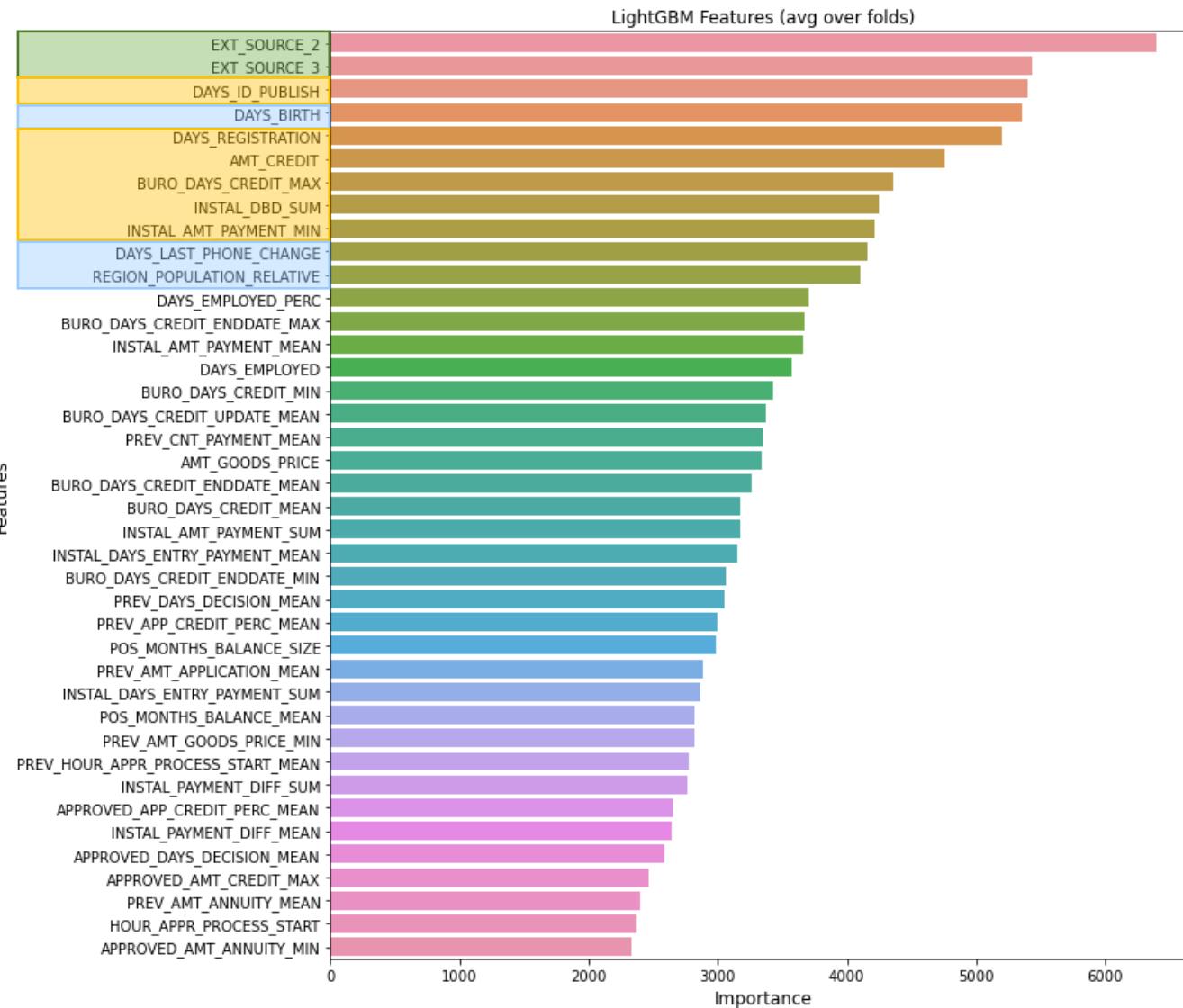


4.b.

Interprétation du modèle



Interprétation globale



Les variables les plus importantes proviennent d'une source externe



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



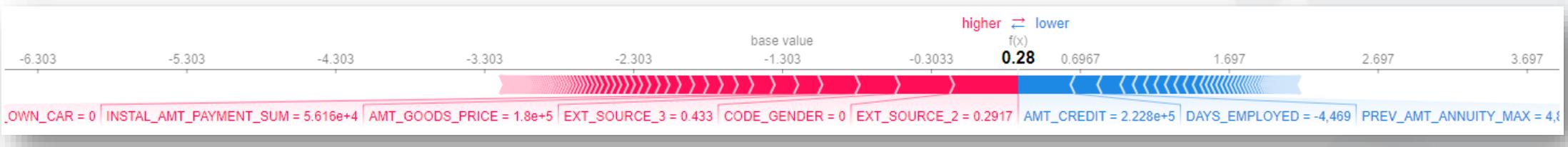
Pris en compte des différentes variables.

Variables personnelles, Variables bancaires, Variables externes

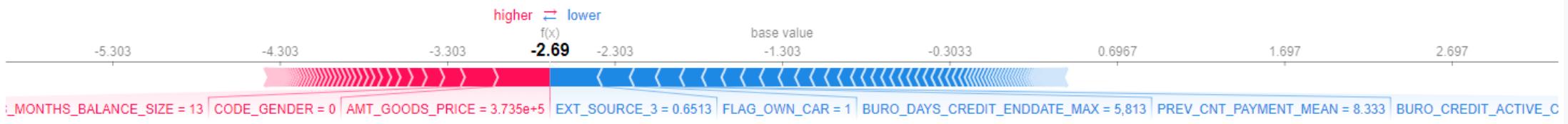
Interprétation locale



Une observation qui est en défaut



Une observation qui n'est pas en défaut



5.

Tableau de bord



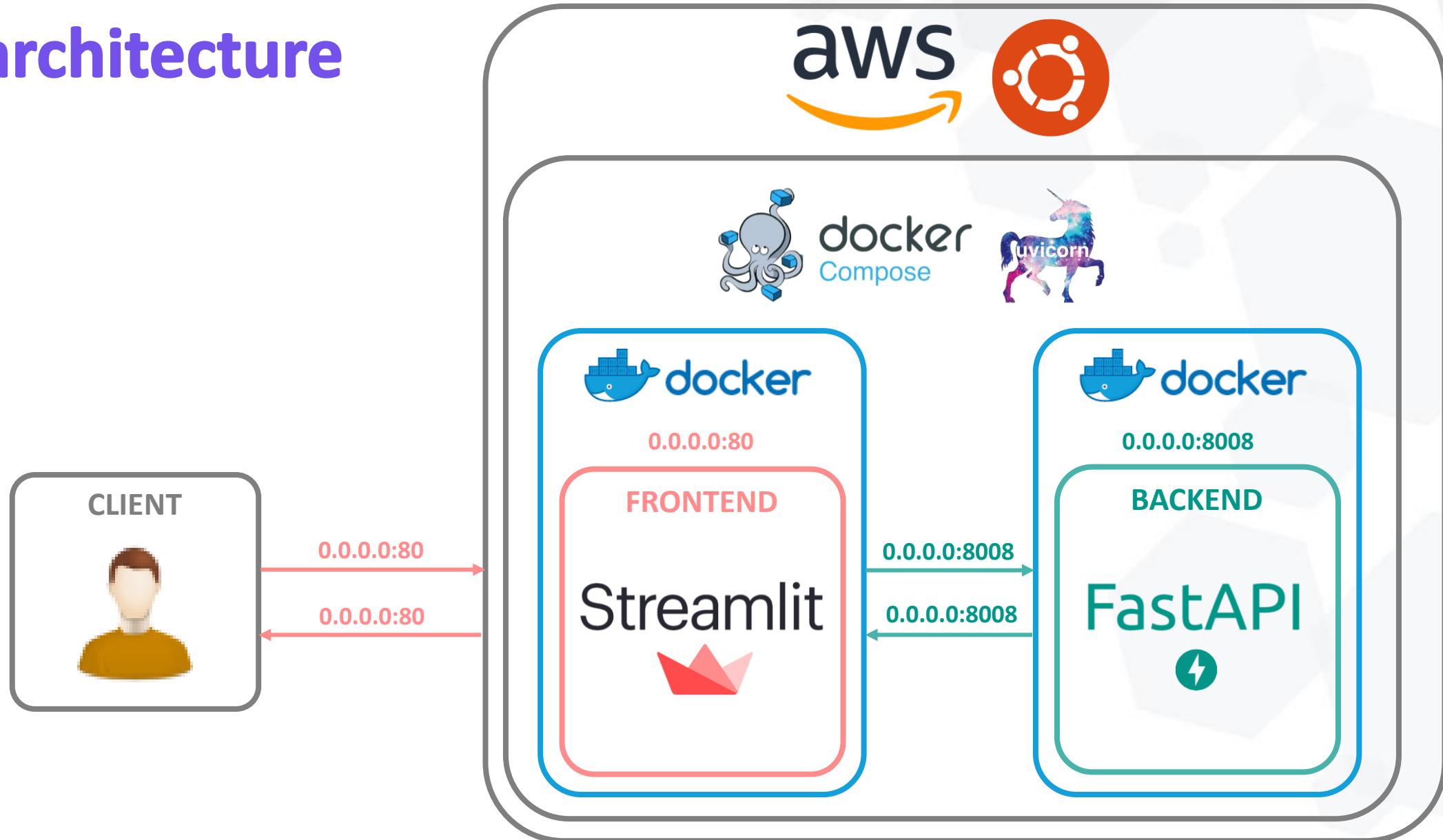
Outils utilisés



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}` Client Details

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

`GET /api/predictions/clients/shap/{id}` Client Shap Df

`GET /api/statistics/ages` Statistical Age

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

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

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

`GET /api/statistics/extSource2` Statistical Ext Source 2

`GET /api/statistics/extSource3` Statistical Ext Source 3

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

Tableau de bord

Prêt à dépenser - Default Risk

Client selection

Client Id list
100057

See local interpretation
 See stats

Option(s) will take more time.

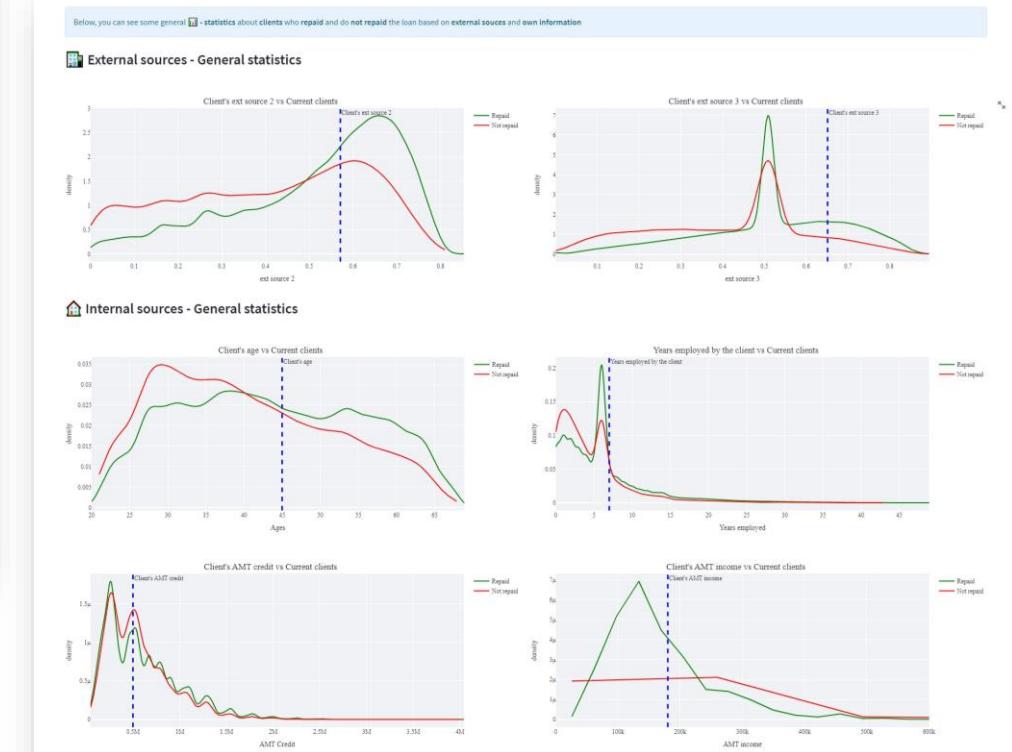
Predict

Select a client to obtain information related to probability of a client not paying the loan. In addition, you can analyze some stats.

Client information



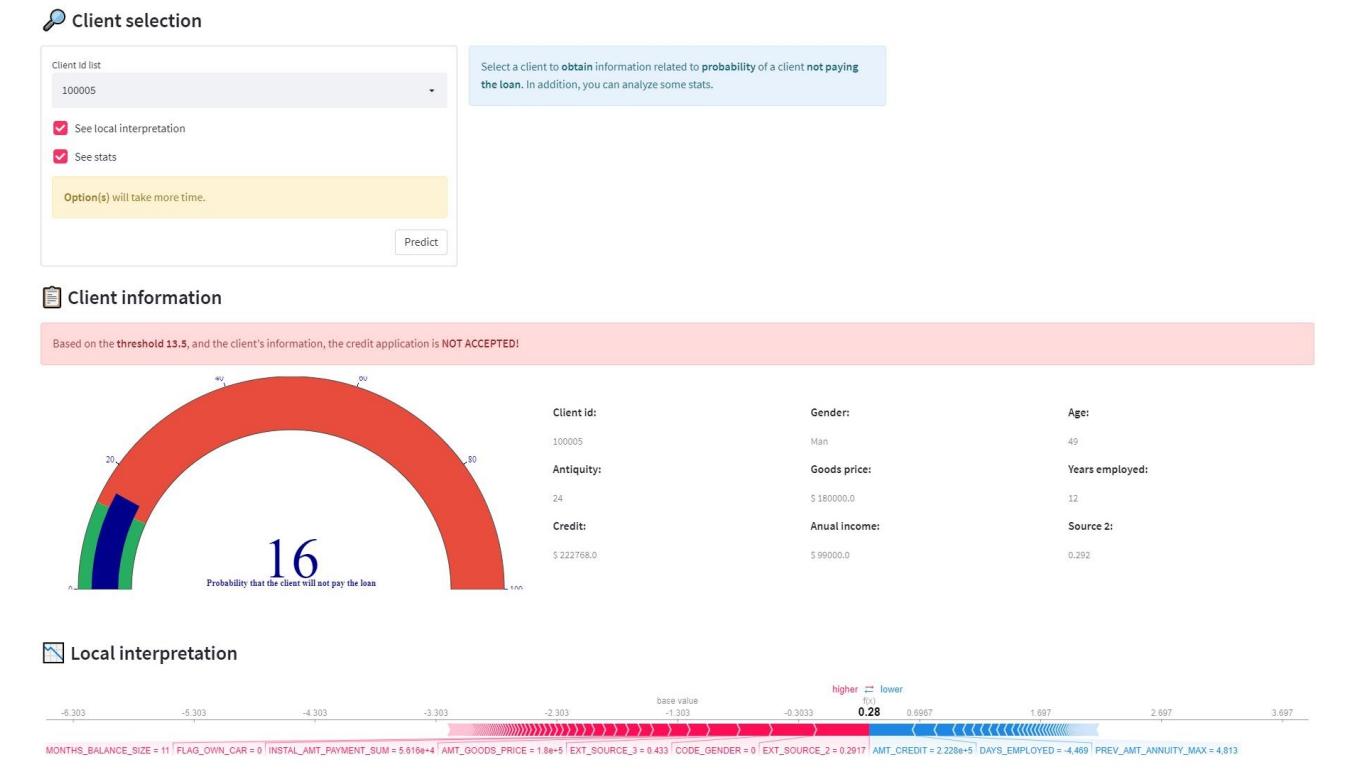
Client avec crédit accepté



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

Tableau de bord

Prêt à dépenser - Default Risk



Client avec crédit rejeté



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

6.

Conclusion



Les conclusions de la mission



Le modèle

- Faire une analyse exploratoire au début pour bien comprendre les données.
- Faire un 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'hyperparamétrisation du modèle
- Essayer d'autres algorithmes comme OneClassSVM, etc.
- Essayer plusieurs coefficients dans le custom score pour avoir un modèle plus ou moins strict



Tableau du bord

- Permettre la sélection de diverses variables au moment de la réalisation du graphique

Avez-vous des questions ?



MERCI

Soutenance de Projet
Samir HINOJOSA

10 février 2022



OPENCLASSROOMS