IMPLÉMENTEZ UN MODÈLE DE SCORING

PROJET N°7
PARCOURS « DATA SCIENTIST »

ETUDIANT : SAHEL TAHERIAN

SOUTENANCE DE PROJET

24 MAI 2022

Plan de la présentation

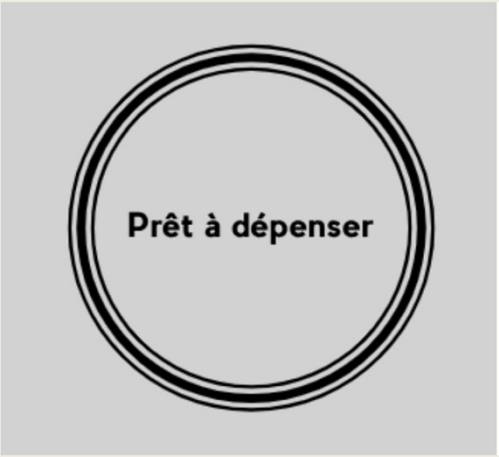
I. Présentation de la problématique

II. Préparation des données et exploration

III. Pistes de modélisations

IV. Présentation du Dashboard

Rappel de la problématique



ENTREPRISE "PRÊT À DÉPENSER"

- Crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt

BESOIN

- Modèle de scoring de la probabilité de défaut de paiement du client

OBJECTIFS

- Dashboard interactif à destination des chargés de relation client

II – PRÉPARATION DU JEU DE DONNÉES

Jeu de données:

- 10 sources de données avec 346 colonns
- Base de données principale : application_train.csv
  - 307 511 clients
  - 122 features : âge, sexe, emploi, logement, revenus, informations relatives au crédit, etc.
  - Taux de remplissage global du dataframe : 75.60 %
  - Cible : défaut de crédit / pas de défaut de crédit

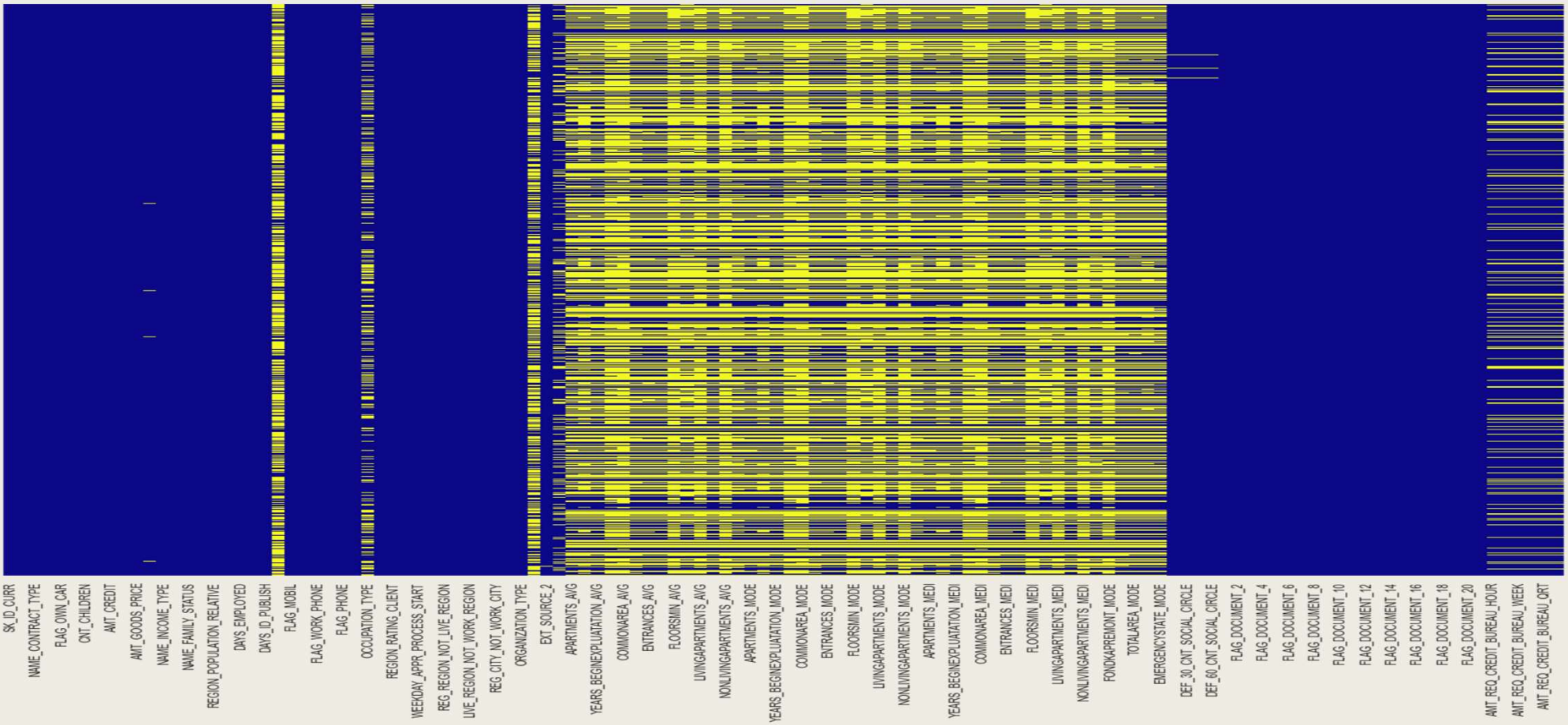
Analyse exploratoire des données et Feature engineering

Inspiré par le Kernel :

https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

- Identification / imputation des valeurs manquantes
- Analyse des outliers / valeurs atypiques.
- Création des nouvelles variables
- Visualisation des corrélations avec notre cible.
- Encodage des variables catégorielles: One Hot Encoding
- Standardisation des données : MinMaxScaler

Suppression des colonnes avec plus de 50 % de données manquantes



Preprocessing

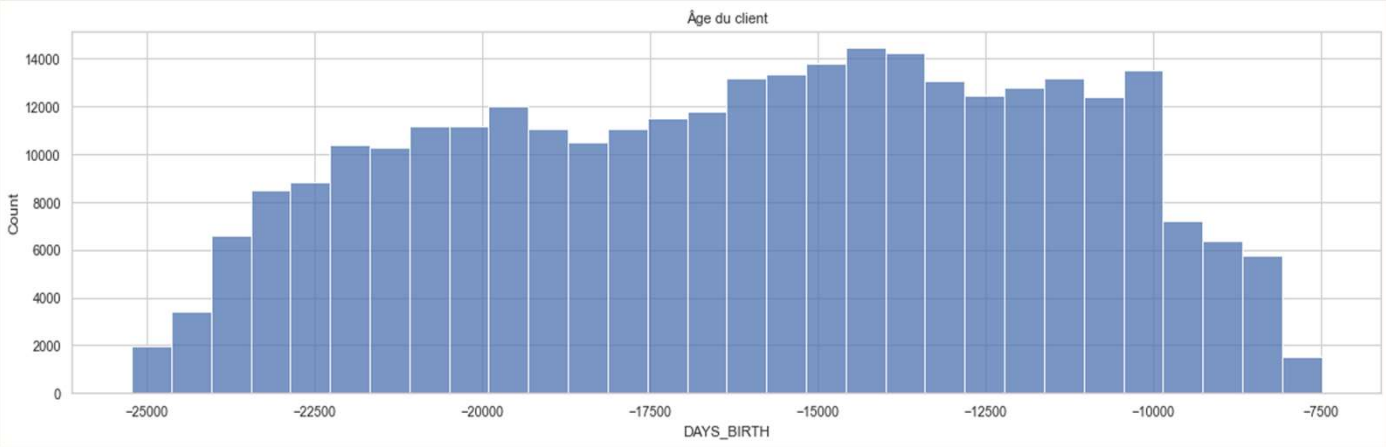
Les 20 premières variables avec plus de valeurs manquantes

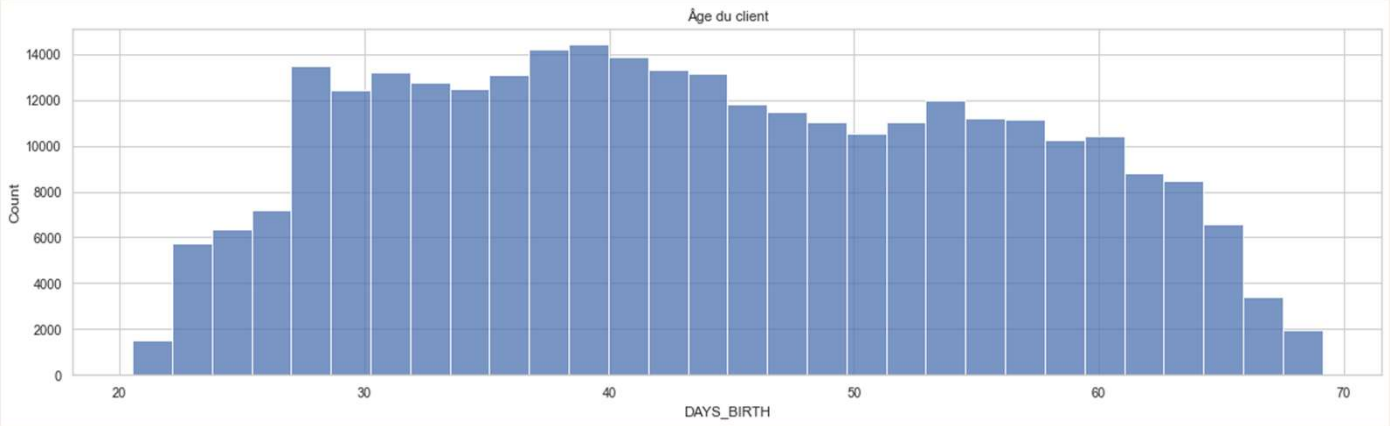
	Total	%
COMMONAREA_MEDI	214865	69.87
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_AVG	213514	69.43
NONLIVINGAPARTMENTS_MEDI	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_MODE	210199	68.35
LIVINGAPARTMENTS_AVG	210199	68.35
LIVINGAPARTMENTS_MEDI	210199	68.35
FLOORSMIN_AVG	208642	67.85
FLOORSMIN_MODE	208642	67.85
FLOORSMIN_MEDI	208642	67.85
YEARS_BUILD_MEDI	204488	66.50
YEARS_BUILD_MODE	204488	66.50
YEARS_BUILD_AVG	204488	66.50
OWN_CAR_AGE	202929	65.99
LANDAREA_MEDI	182590	59.38
LANDAREA_MODE	182590	59.38
LANDAREA_AVG	182590	59.38

Preprocessing

- **Outliers, valeurs atypiques, anormales**

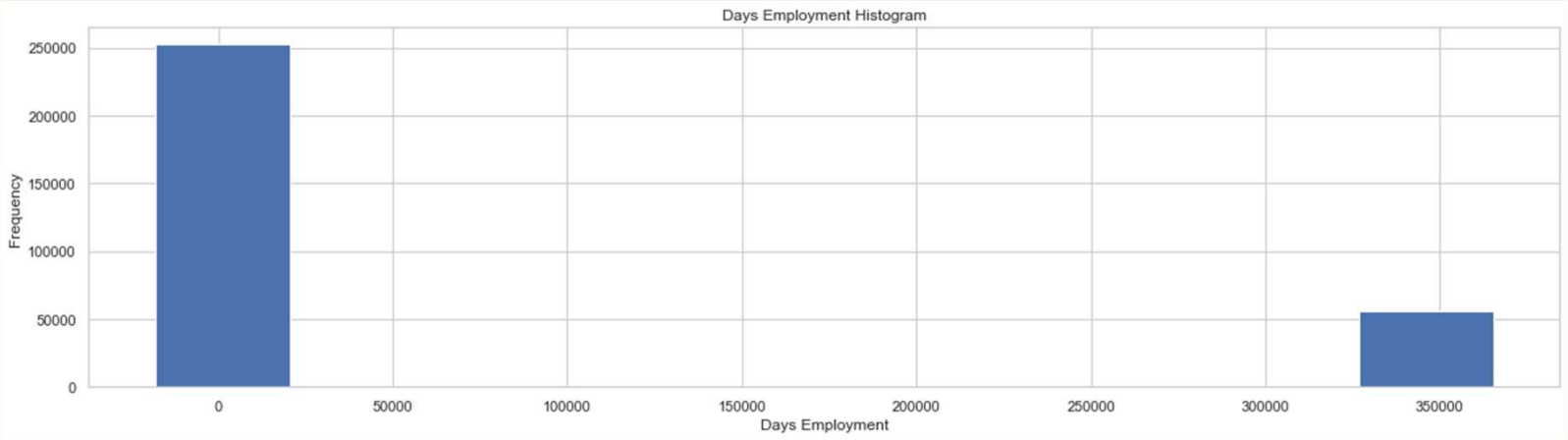
Âge du client

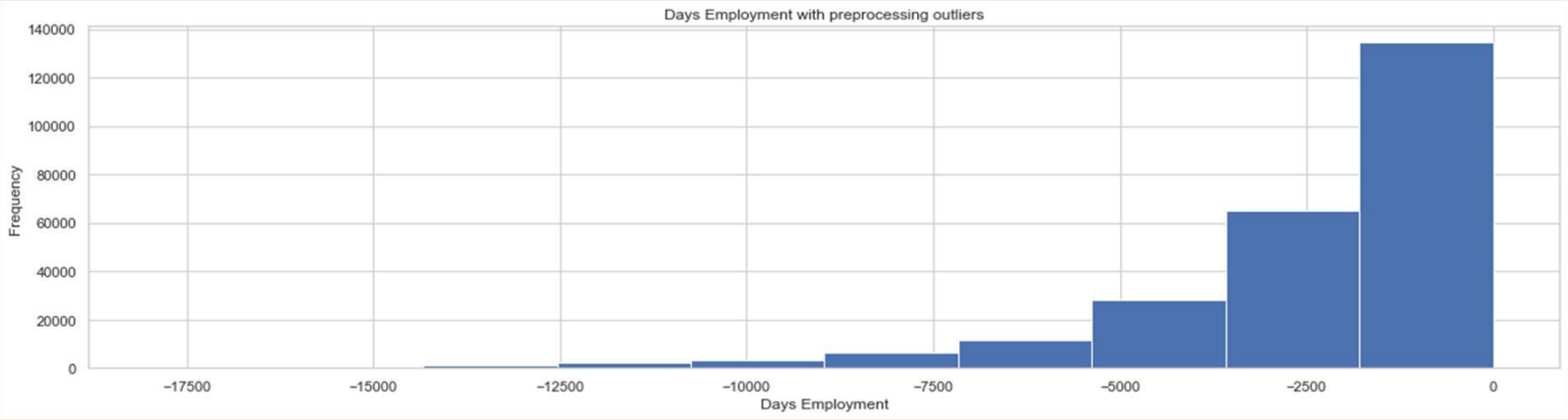




Preprocessing

DAYS_EMPLOYED



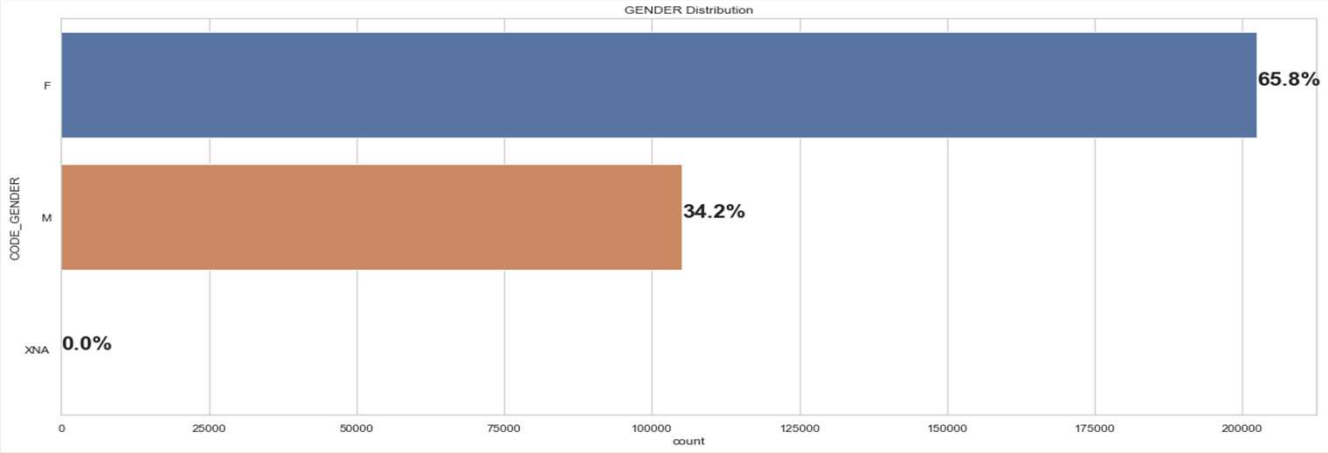


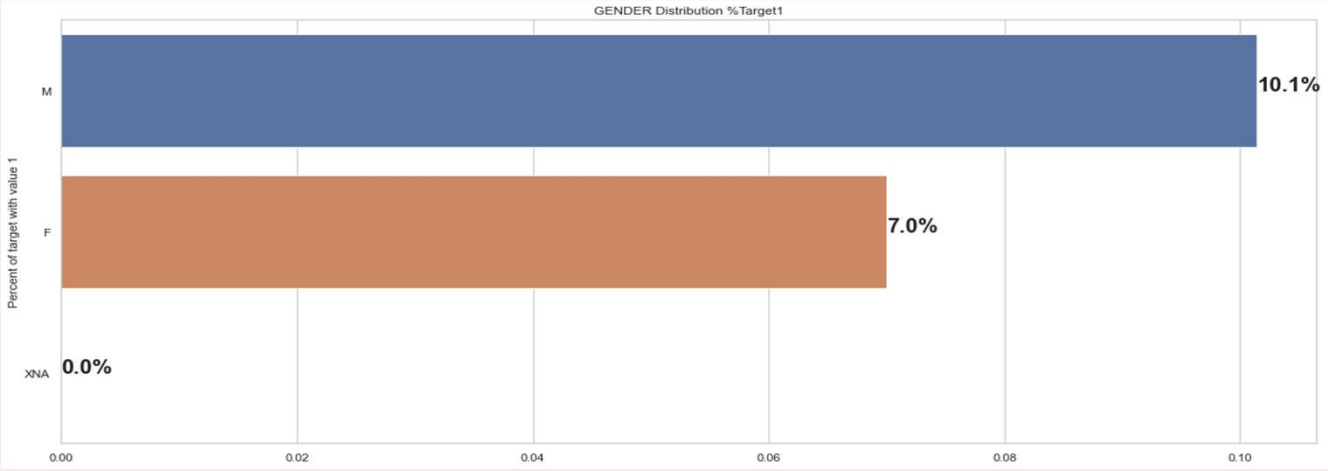
Création de 4 nouvelles variables métiers

- **CREDIT_INCOME_PERCENT: Pourcentage du montant du crédit par rapport au revenu d'un client**
- **ANNUITY_INCOME_PERCENT: Pourcentage de la rente de prêt par rapport au revenu d'un client**
- **CREDIT_TERM: Durée du paiement en mois**
- **DAYS_EMPLOYED_PERCENT: Pourcentage des jours employés par rapport à l'âge du client**

Preprocessing

Analyse des principales variables





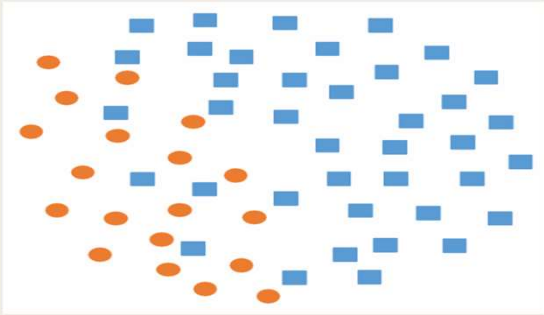
Essais des différentes approches de modélisation

Essais des différentes approches de modélisation

- Jeu de données déséquilibrés
- Métrique de performance
- Méthodologie
- Modèle retenu

Jeu de données désequilibrés

Un problème de classification binaire avec une classe sous représentée



- **92 % des clients sans défaut**
- **8 % des clients avec des défauts de paiement**
  - ***Le traitement suréchantillonnages pour avoir une répartition plus égalitaire :***

SMOTE (Synthetic Minority Oversampling Technique); la classe minoritaire est suréchantillonnée

```
Label 1, Before using SMOTE: 17412
Label 0, Before using SMOTE: 197845
```

```
Label 1, After using SMOTE: 197845
Label 0, After using SMOTE: 197845
```

Choix de la métrique

- **Eviter de mal catégoriser un client avec un fort risque de défaut:**
  - *Minimiser le pourcentage de faux négatifs et à maximiser le pourcentage de vrais positifs;*
    - **Faux négatifs**: Perte réelle si le crédit client accepté se transforme en défaut de paiement
    - **Vrais positifs :** Les cas d'acceptation, le crédit client sera remboursé
    - **Faux positifs:** Perte d'opportunité si le crédit client est refusé à tort, alors qu'il aurait été en mesure d'être remboursé.

les pertes d'un crédit en raison d'une mauvaise classification, dépendent des probabilités Faux Positifs et Faux Négatifs

- **Pénaliser les faux positif et négatifs**
- **Maximiser deux critères Recall et Précision.**
  - *Perte plus importante si un prêt n'est pas remboursé que si on ne prend pas 1 client*

permet d'identifier un compromis entre les 2 métriques

$$\text{Fscore} = \frac{precision.\,recall}{\beta^2.\,precision + recall}$$

Beta : importance relative du recall par rapport à la précision

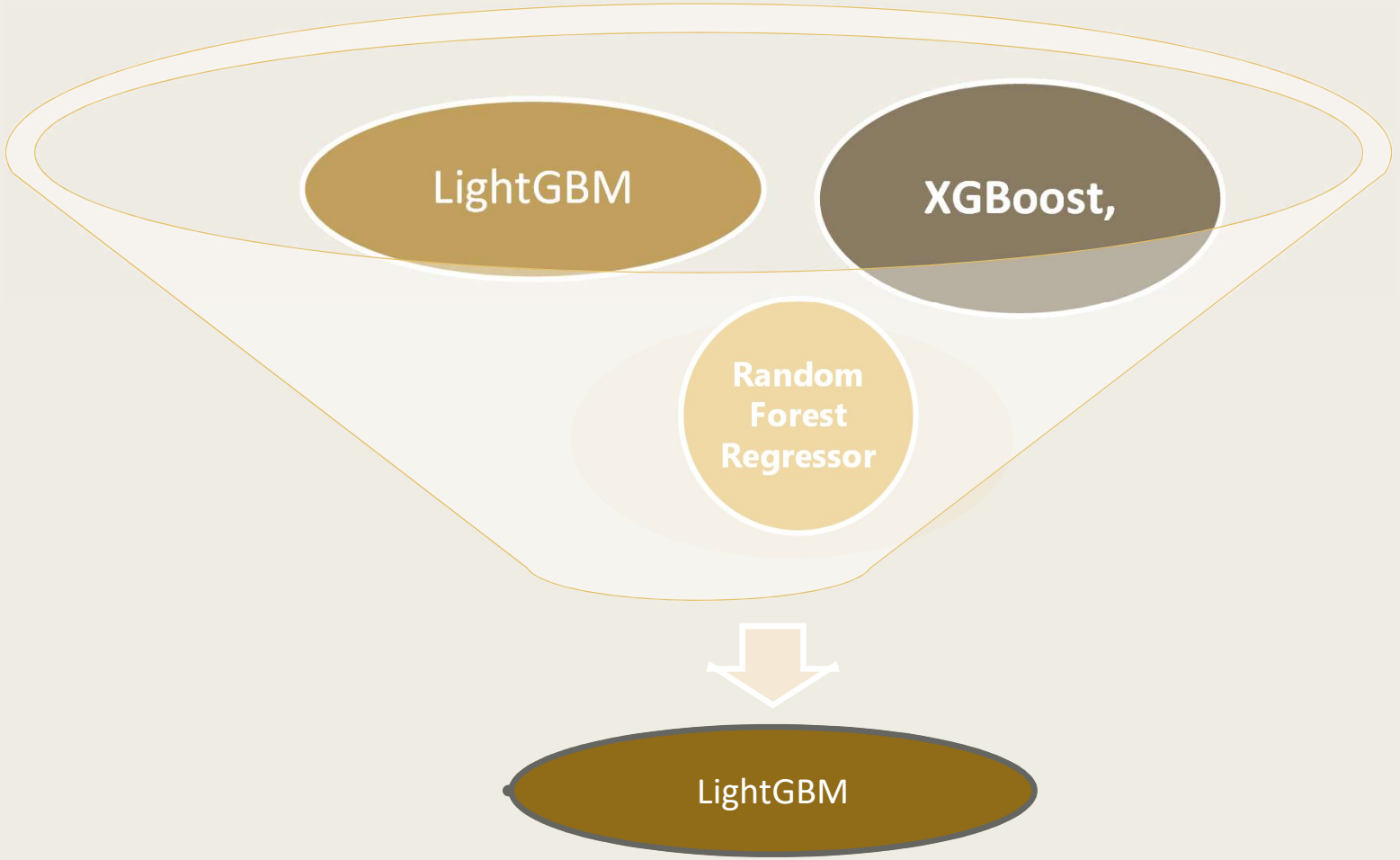
Méthodologie

Processus

- **Mettre en place un modèle baseline pour évaluer les performances de nos futurs modèles**
  - ***Baseline: Régression Logistique:***
    - **Données déséquilibrés (Simple Régression logistique)**
    - **Données équilibrés (Utilisation d'un Grid Search)**
- ***Modèles entrainés:***
  - ***XGBOOST : Simple et avec un Grid Search***
  - ***LightGBM : Simple et avec un Grid Search***
  - ***Random Forest : Simple et avec un Grid Search***

- **Interprétabilité de Modèle:**
  - ***Feauture importance***
  - ***Lime***
  - ***Shap***

Modèle retenu



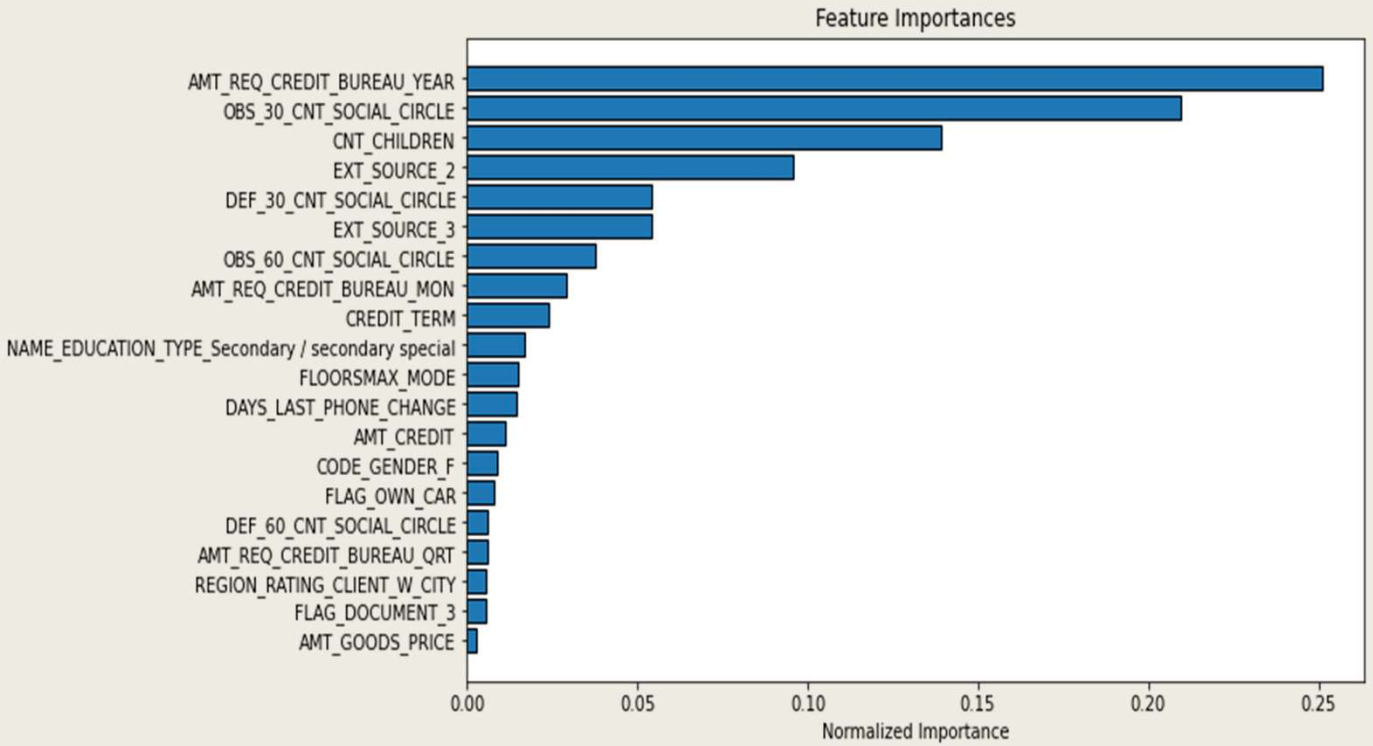
Modèle retenu

Résultat des métriques

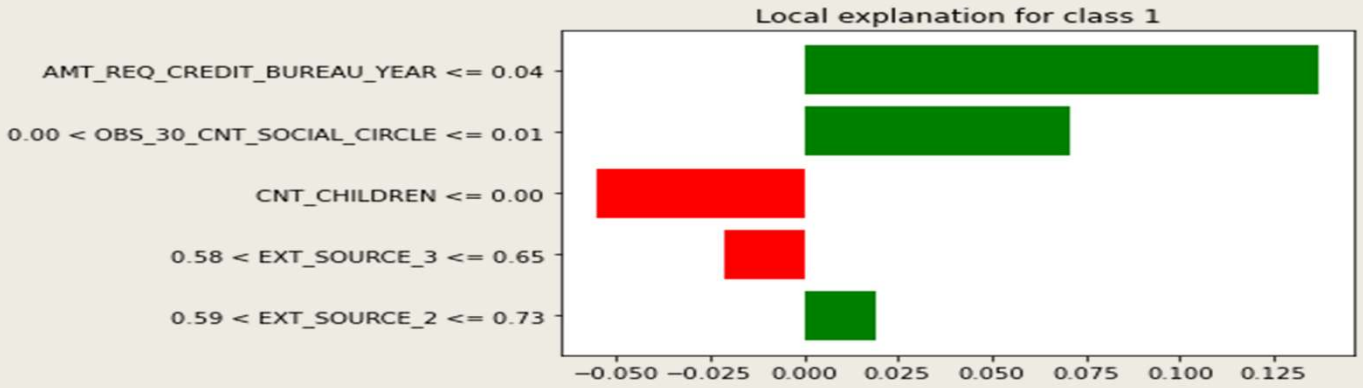
Model	AUC	Accurancy	Score custom	Time
LGBMClassifier	0.741376	0.919581	0.602977	43.0968
LogisticRegression	0.73633	0.692729	0.579606	22.4765
RandomForestClassifier	0.704171	0.91789	0.602196	273.956
XGBClassifier	0.703975	0.917185	0.603659	236.391

Interprétabilité de Modèle: LGBM

Feature importance

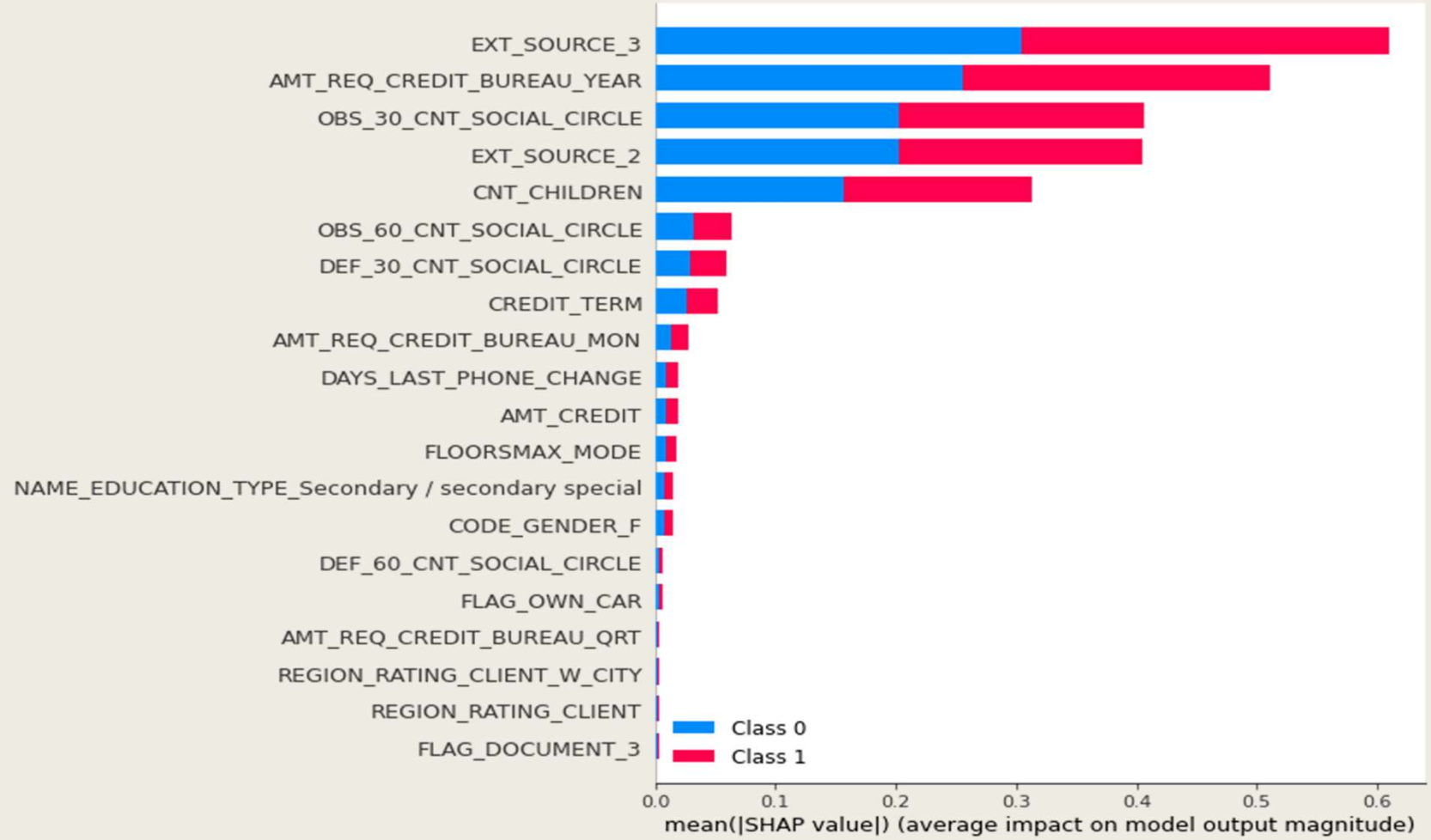


LIME



Interprétabilité de Modèle: LGBM

SHAP



Dashboard et le déploiement de modèle

Les étapes de déploiement du modèle

- **Entraînement d'un modèle d'apprentissage automatique sur un système local;**
- **Création de l'interface à l'aide de Streamlit, pour rendre le modèle accessible;**
- **Envelopper l'inférence avec un framework backend de FastAPI;**
- **Utilisation d'un docker pour conteneuriser l'application;**
- **Hébergement de l'API (FastApi) sur le site heroku**

Démonstration

- **Dépôt github :**

https://github.com/Sahel129/Implementez-un-modele-de-scoring

- **Dashboard (En local) :**

https://localhost:8501/

- **L'API (À distance):**

https://my-fastapi.herokuapp.com/docs

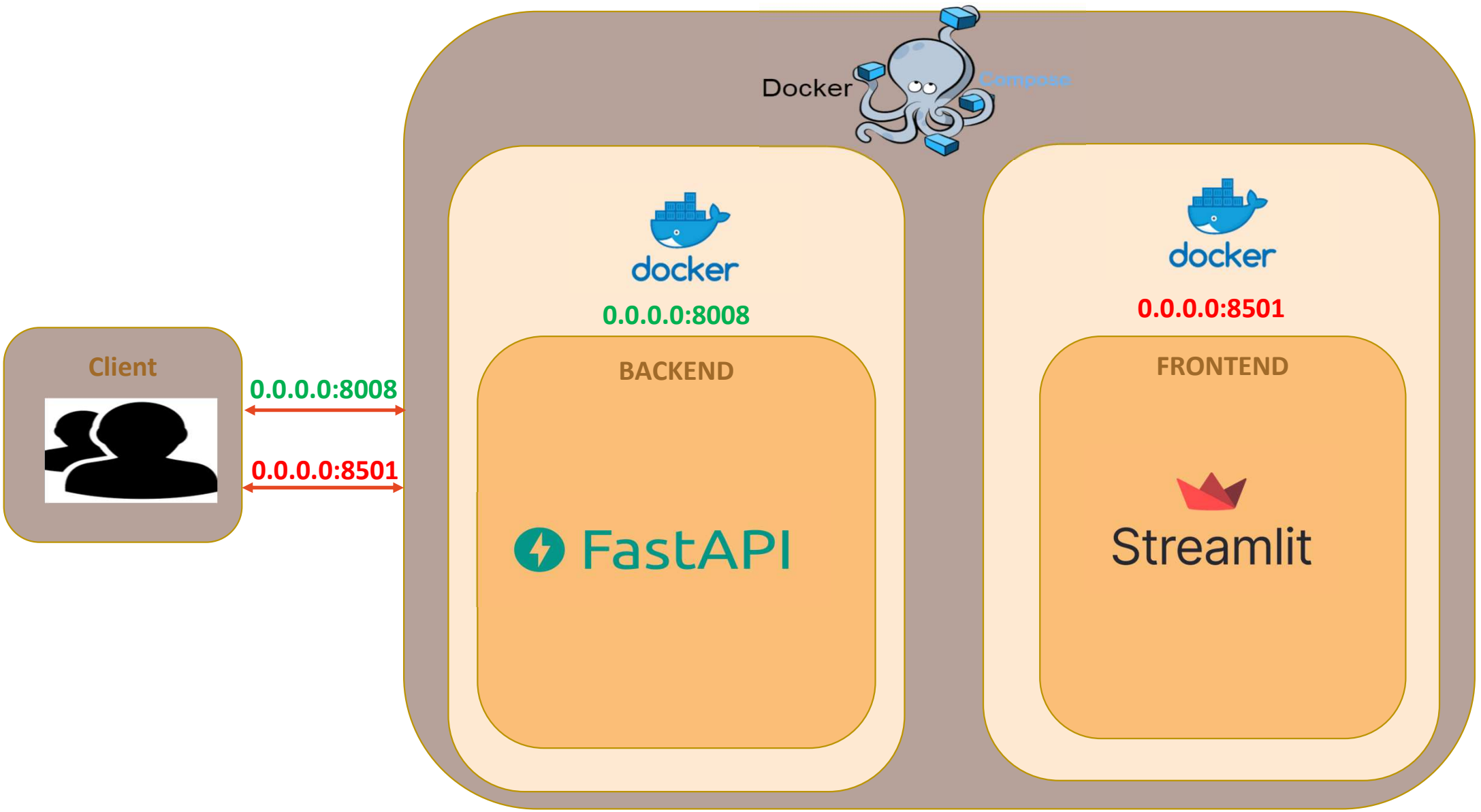
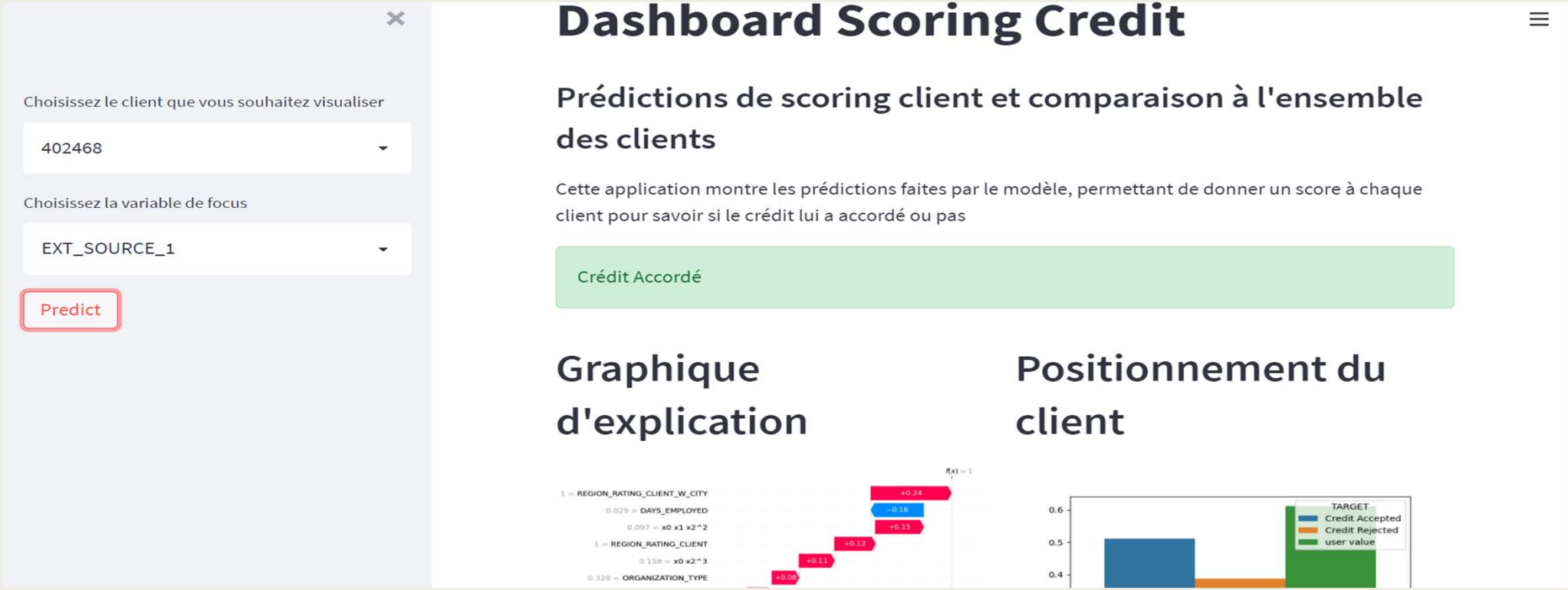

Docker
Compose
docker
0.0.0.0:8008
BACKEND
FastAPI
docker
0.0.0.0:8501
FRONTEND
Streamlit
Client
0.0.0.0:8008
0.0.0.0:8501

Tableau de bord interactif



Commencer la démonstration

MERCI DE
VOTRE ATTENTION