

SOUTENANCE PROJET 7

IMPLEMENTER UN MODELE DE SCORING

Florian FOLLAIN Février 2024



A close-up, high-angle photograph of several rolled-up architectural blueprints. The blueprints are white with black lines and text, showing various technical drawings, dimensions, and annotations. The rolls are stacked and slightly unrolled, revealing different sections of the plans. The background is a light, textured surface.

SOMMAIRE

1 Présentation du projet

2 EDA et feature engineering

3 Modélisation avec mlflow

4 Dashboard et deployment

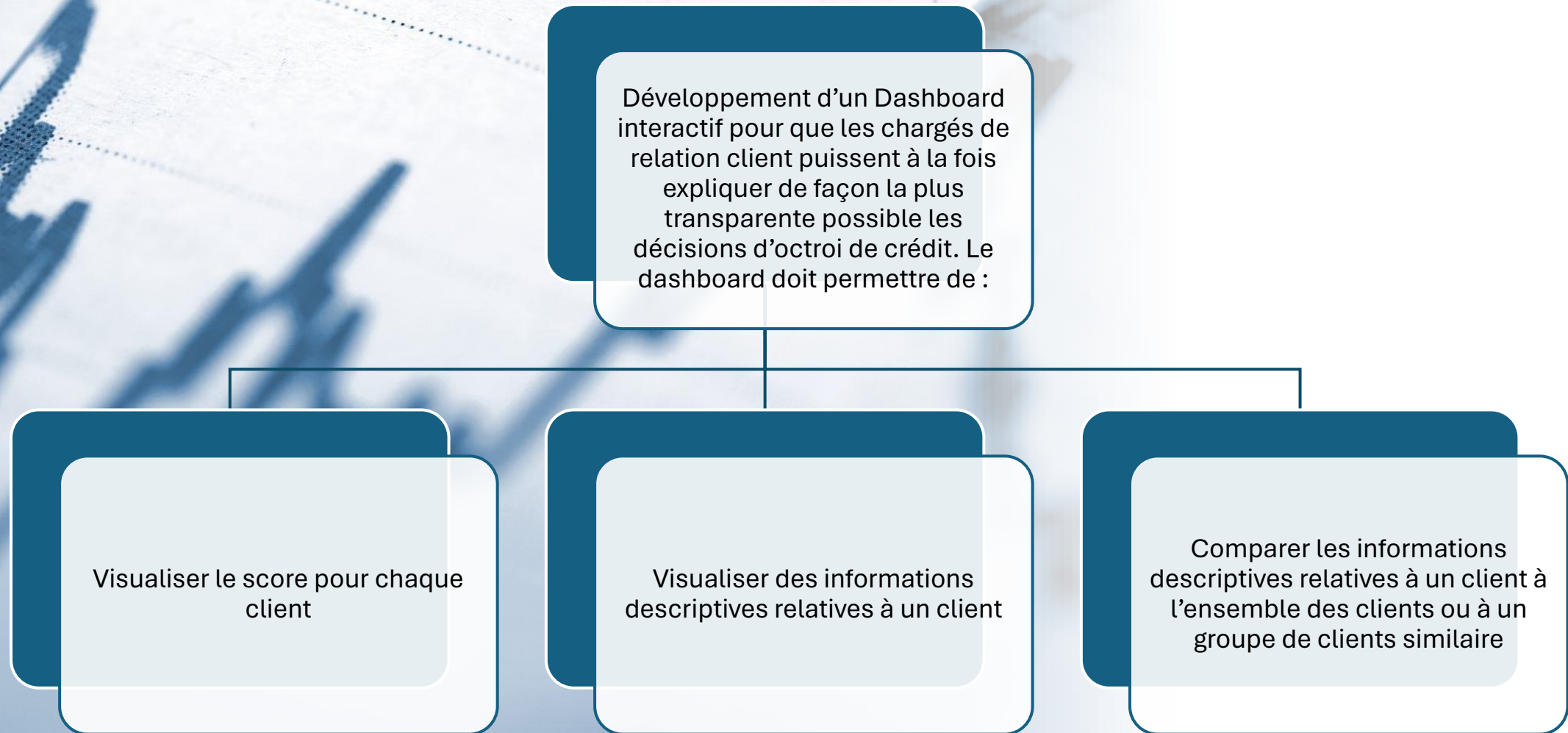
5 Limites et améliorations

1-PRESENTATION DU PROJET



DASHBOARD

Prêt à dépenser souhaite développer un modèle de Scoring de la probabilité de défaut de paiement client pour étayer la décision d'accorder ou non un prêt potentiel



DEMANDES DU MANAGER

Tenir compte du déséquilibre des données

Tenir compte du déséquilibre du coût métier

Démarche MLOps

Détection du data drift

Déployer l'API (https://github.com/flystoneflorian/P7_api)

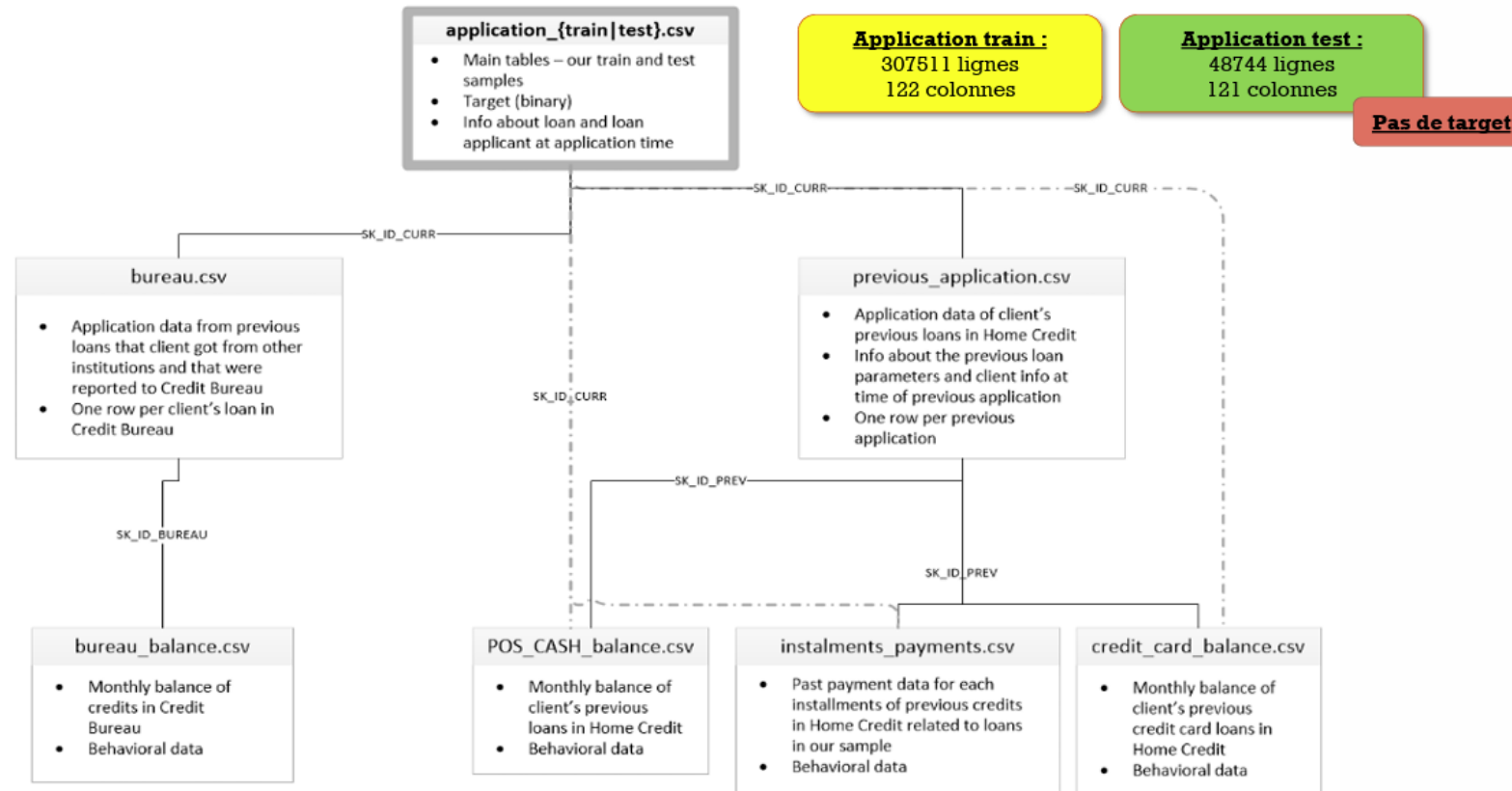
Déployer le Dashboard (https://github.com/flystoneflorian/P7_scoring)

2-EDA ET FEATURE ENGINEERING



PRESENTATION DES DONNEES

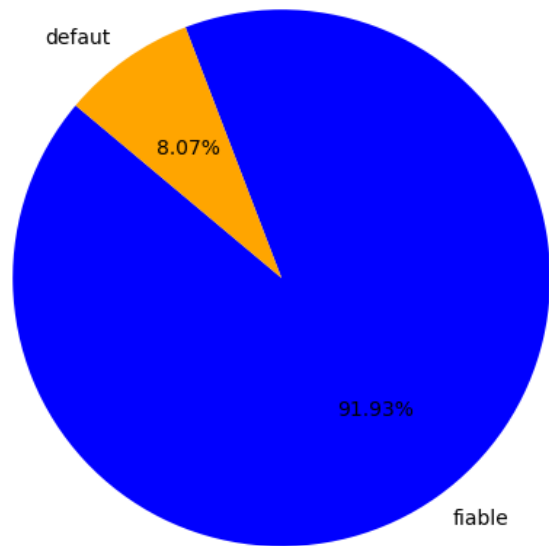
- Nous disposons d'un dataset de 307511 clients classés en 2 catégories et de 121 features par client



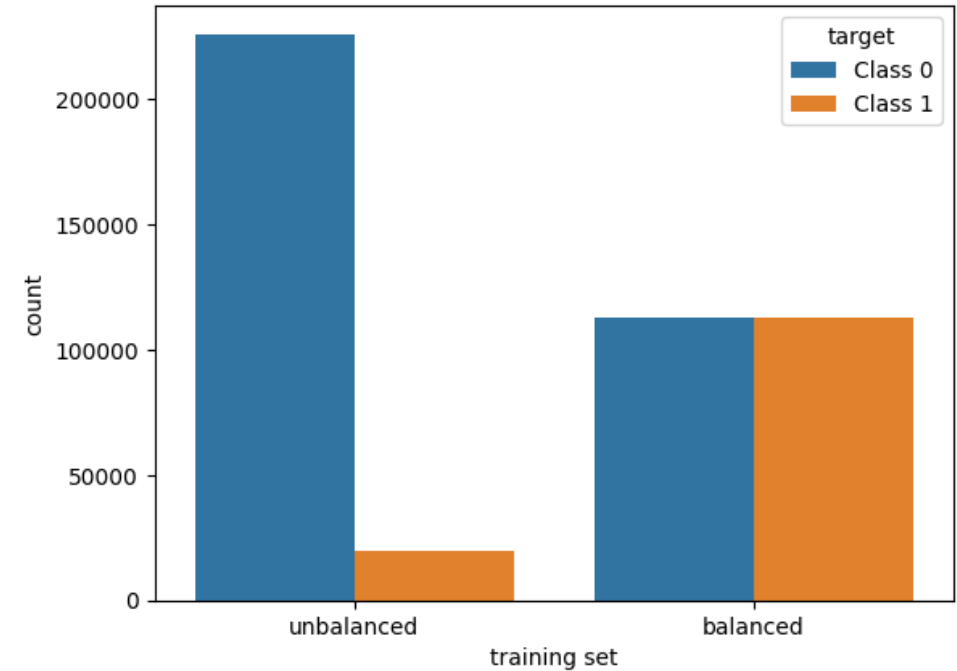
DISTRIBUTION DES DONNEES

- La répartition des clients est déséquilibrée, nous allons donc la rééquilibrer artificiellement avant de commencer l'entraînement du modèle

Répartition des états de paiements



Target distribution



3-MODELISATION AVEC ML FLOW



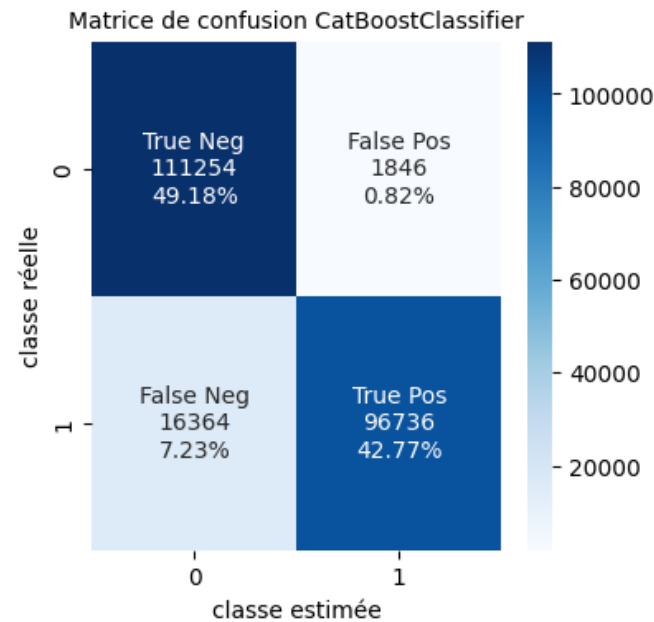
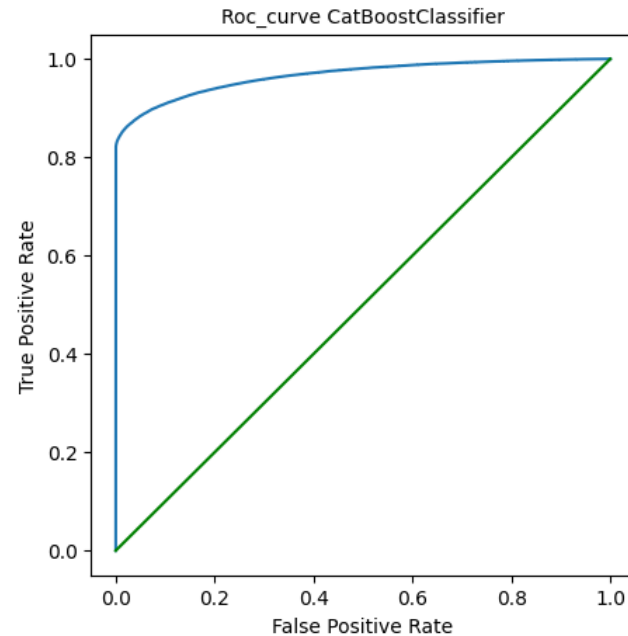
CHOIX DES ALGORITHMES

Nous avons retenu 5 algorithmes dont nous allons comparer les résultats

- Dummy qui est notre modèle naïf
- LogisticRegression qui est un modèle linéaire
- DecisionTreeClassifier qui est un arbre de décision
- LGBMClassifier qui est une méthode ensembliste
- CatBoost qui est aussi une méthode ensembliste plus performante sur les données catégorielles



CHOIX DES METRICS



Pour évaluer les modèles nous allons produire pour chacun une matrice de confusion

Puis calculer l'AUC (aire sous la courbe ROC)

Et enfin définir un score grâce à notre fonction de coût métier selon les consignes managériales

FONCTION DE COUT

- Grâce à une fonction customisée nous allons pénaliser le modèle en l'entraînant avec une contrainte particulière
- En effet dans la cadre d'un prêt bancaire le poids des erreurs n'est pas le même selon que l'on refuse un prêt qui aurait pu être accordé ou nous accordions un prêt qui n'aurait pas dû l'être
- Ici le rapport nous est fourni, il est de 1 pour 10
- Notre score se lit de 0 à 10, 0 étant le score parfait



MLFLOW

Experiments



- ☐ Default
- ☒ CatBoost_Classifier_Evaluation
- ☒ LGBM_Classifier_Evaluation
- ☒ Dummy_Classifier_Evaluation
- ☒ Decision_Tree_Evaluation
- ☒ Logistic_Regression_Evaluation

Displaying Runs from 5 Experiments



Time created ▾

State: Active ▾

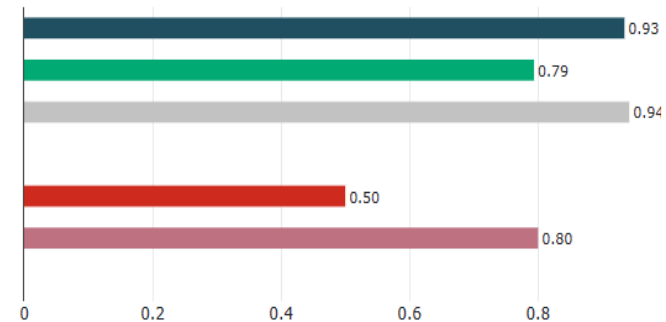
Sort: Created ▾

[Share](#)Table **Chart** Evaluation **Experimental**

Run Name
● CatBoost Classifier
● Logistic Regression
● LGBM Classifier
● Logistic Regression
● Dummy Classifier
● Decision Tree
● Logistic Regression

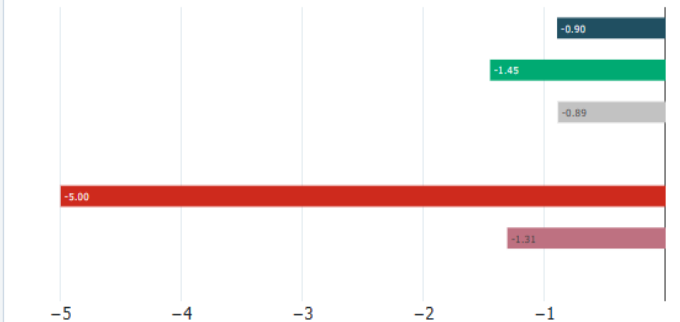
best_score_auc

Comparing first 7 runs










best_score_metier

Comparing first 7 runs



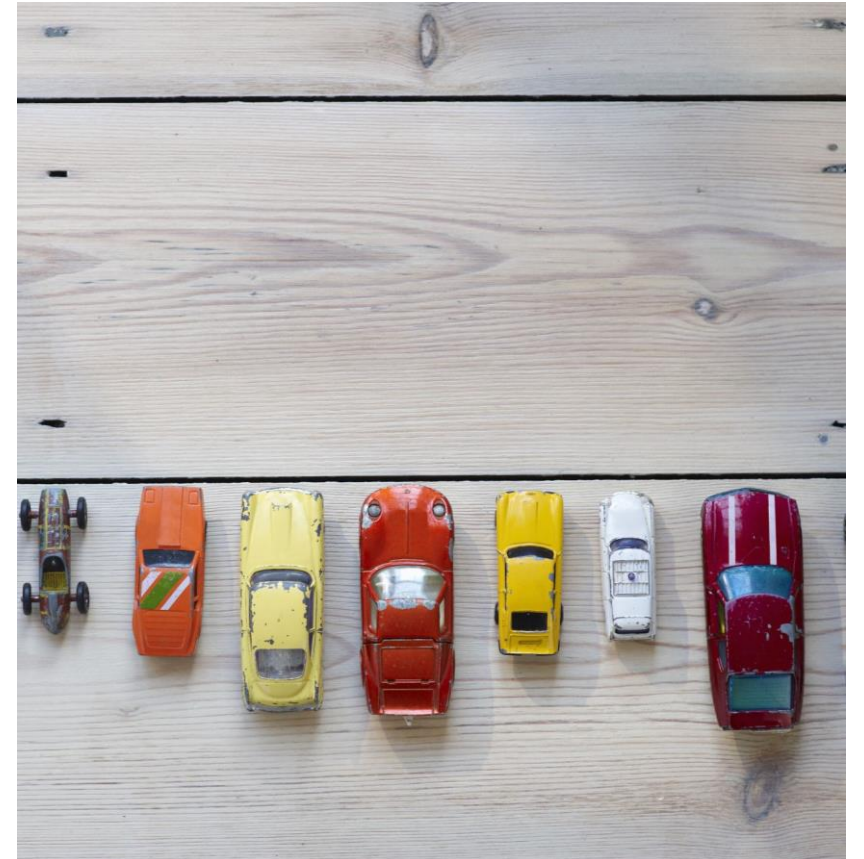
DATA DRIFT

Drift is detected for 7.438% of columns (9 out of 121).

<div><div></div><div>Search</div><div>×</div></div>					
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test
> AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)
> AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)
> AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)
> AMT_CREDIT	num			Detected	Wasserstein distance (normed)
> AMT_ANNUITY	num			Detected	Wasserstein distance (normed)

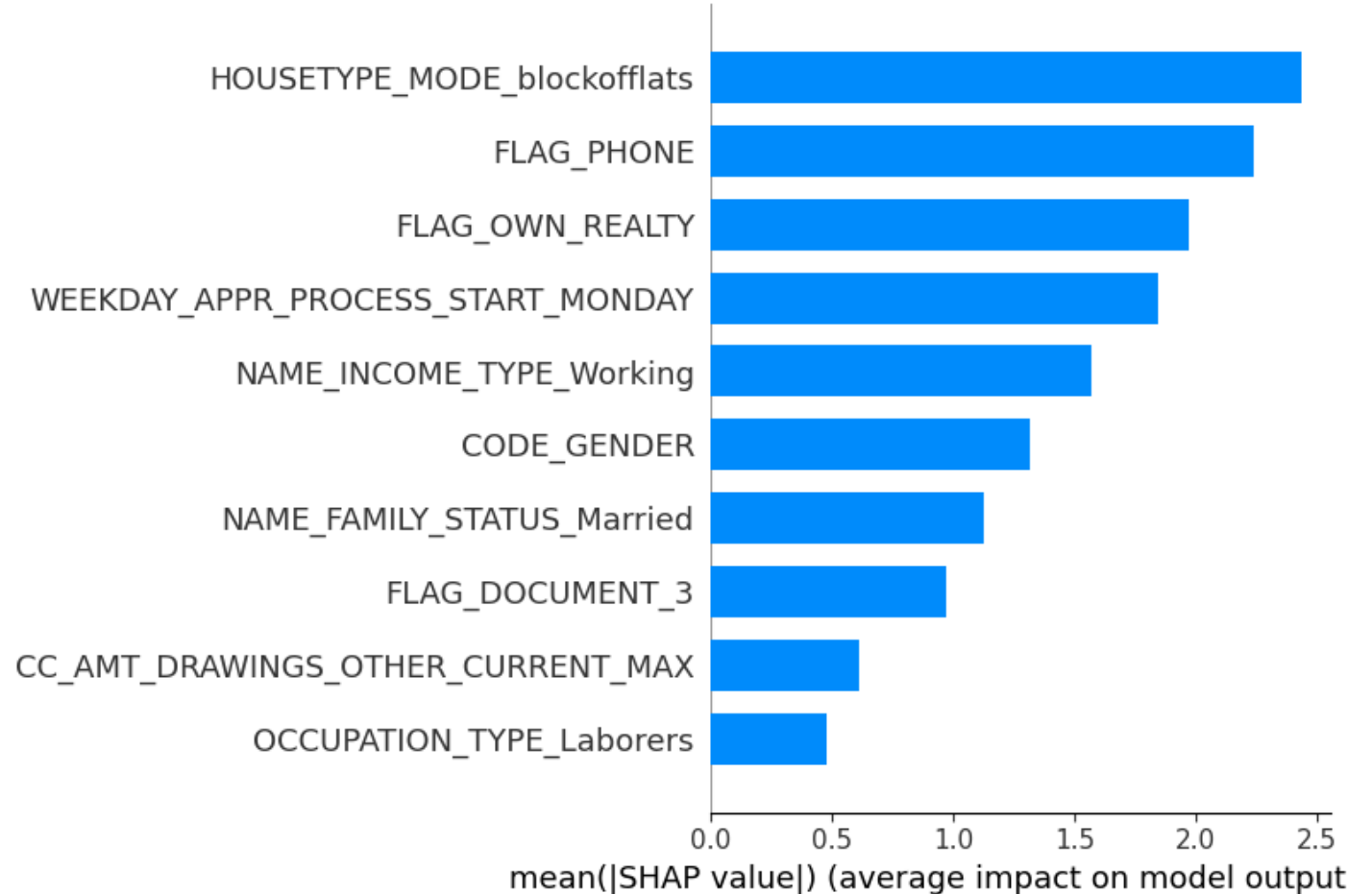
CHOIX DU MODELE

```
{'best_model_name': 'CatBoostClassifier',  
'best_score': 0.7315915119363395,  
'best_params': {'CatBoost__max_depth': 3,  
'CatBoost__n_estimators': 300},  
'second_best_model_name': 'LGBMClassifier',  
'second_best_score': 0.7519451812555261,  
'second_best_params':  
{'LGBM__colsample_bytree': 0.99,  
'LGBM__reg_lambda': 2, 'LGBM__subsample':  
0.5}}
```



INTERPRETATION

- La méthode SHAP consiste à calculer la valeur de Shapley pour toutes les variables de tous les individus c'est-à-dire la moyenne de l'impact d'une variable (sur la sortie, donc la prédiction) pour toutes les combinaisons de variables possibles. La somme des effets de chaque variable explique alors la prédiction.



4-DASHBOARD ET DEPLOIEMENT



DEPLOYMENT

- Pour déployer notre application nous avons utilisé Fast API pour créer une API de prédiction, hébergée sur GITHUB et mis en réseau via Heroku
- Pour déployer le Dashboard nous avons utilisé la plateforme Streamlit sur laquelle il a été conçu
- Présentation du Dashboard :

<https://flystoneflorian-p7-scoring-pythonapp7-dashboard-zvpdjo.streamlit.app/>



5-LIMITES ET AMELIORATIONS



LIMITES

La réalisation du modèle a nécessité la conception de nombreux blocs de transformation et de traitement des données. Chaque bloc fait appel à des méthodes paramétrables. De fait, les résultats sont dépendants des paramètres choisis. L'architecture du code permet d'optimiser les blocs indépendamment.

Sélection des variables

Les informations disponibles relatives à l'importance des variables sont débattues avec les experts métier en vue de définir les stratégies techniques à tester dans les différents blocs concernés :

- Valeurs manquantes
- Corrélations entre variables
- Seuil de variance
- Réduction de dimensions (RFE)

AMELIORATIONS

Equilibrage des données

L'équilibrage des données introduit des données artificielles donc la possibilité d'incohérences. Des tests peuvent être réalisés en variant certains paramètres (ratios classes).

Fonction d'évaluation du gain

Les règles métier et les critères financiers relatifs aux pertes et profits doivent être communiqués en vue d'établir une fonction d'évaluation du gain adaptée.

Choix de l'algorithme

Nous avons testé plusieurs classifieurs et retenu LightGBM pour sa rapidité d'exécution. D'autres classifieurs comme XGBoost peuvent potentiellement apporter de meilleures performances techniques. Il s'agit de les tester dans le pipeline complet si les contraintes de temps en production le permettent

MERCI !

