Projet 7 : Implémentez un modèle de scoring

Parcours Data Scientist – OpenClassrooms

Soutenance

Sommaire

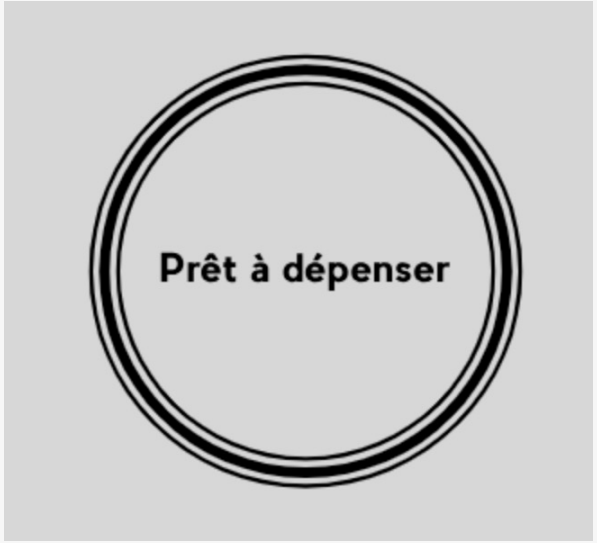
I - Présentation

Problématique

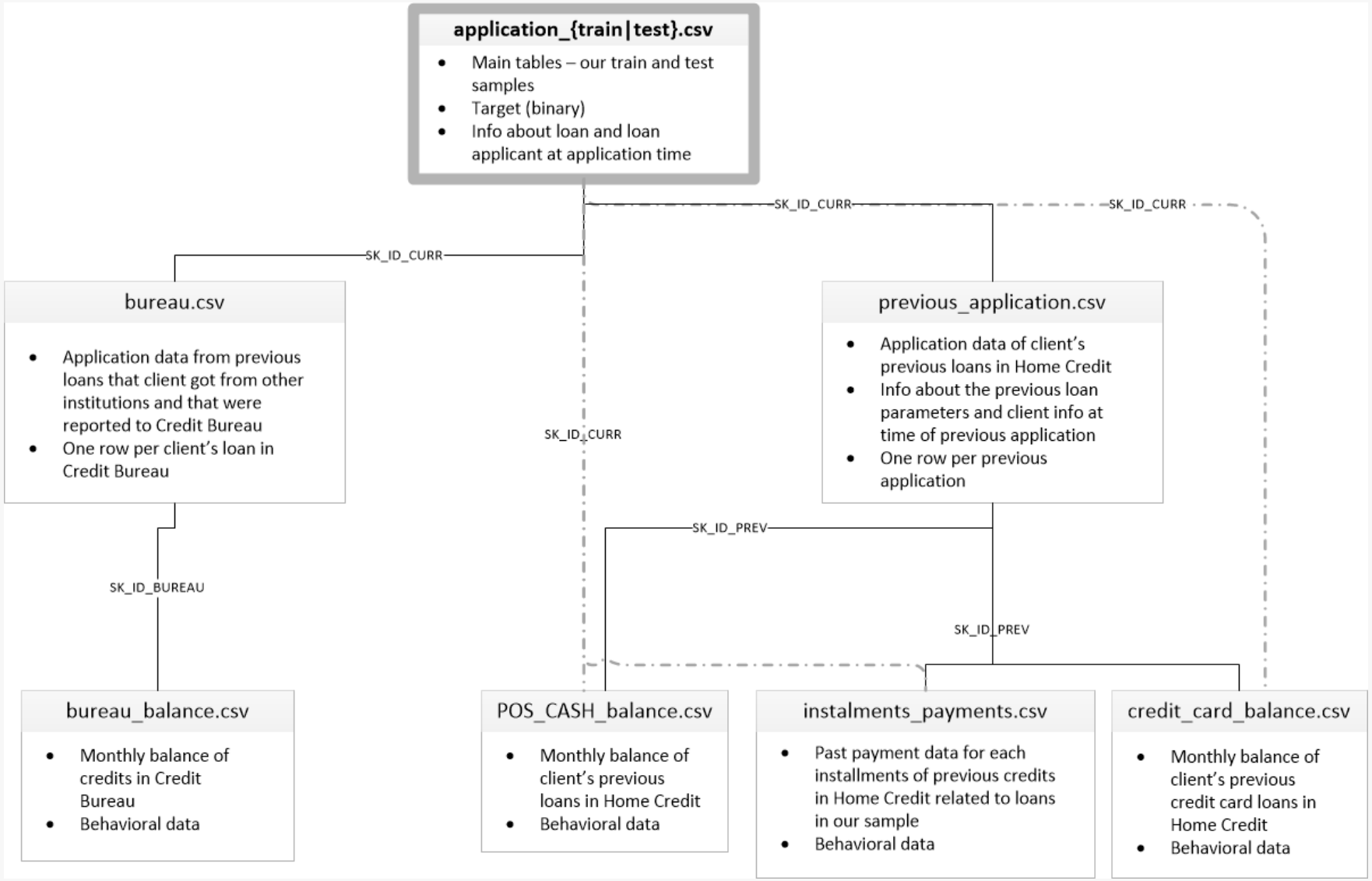
- société financière "Prêt à dépenser"
- outil de “scoring crédit”

Missions :

- Construire un modèle de scoring
- Construire un dashboard interactif à destination des gestionnaires de la relation client
- Mettre en production le modèle de scoring de prédiction à l’aide d’une API, ainsi que le dashboard interactif



Dataset


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

Traitement des données

Kernel
Kaggle

→ Données manquantes : suppression >50% données manquantes

↓

Imputation : 0, médiane, mode

→ Preprocessing : Standard Scaler et One Hot Encoder

↓

Séparation du dataset : train, validation, test

II - Modélisation

Métriques

Score métier :

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

$$gain = \frac{TP * TP_{coeff} + TN * TN_{coeff} + FP * FP_{coeff} + FN * FN_{coeff}}{(TN + FP + FN + TP)}$$

$$\text{Accuracy} = \frac{Vrai\ positif + Vrai\ négatif}{Total}$$

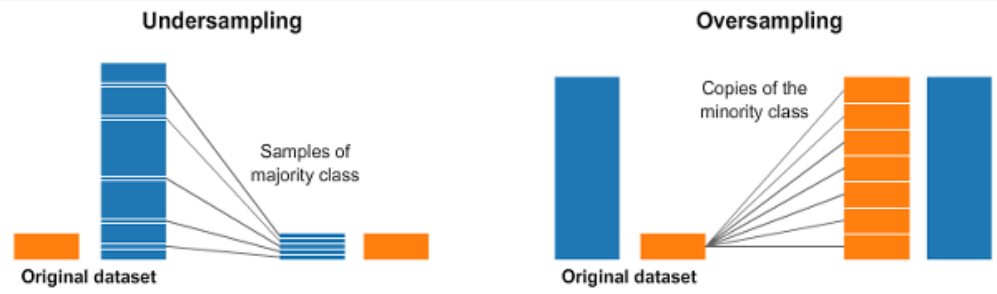
$$Precision(i) = \frac{nb\ d'individus\ correctement\ attribués\ à\ la\ classe\ i}{nb\ d'individus\ attribués\ à\ la\ classe\ i}$$

$$Recall(i) = \frac{nb\ d'individus\ correctement\ attribués\ à\ la\ classe\ i}{nb\ d'individus\ appartenant\ à\ la\ classe\ i}$$

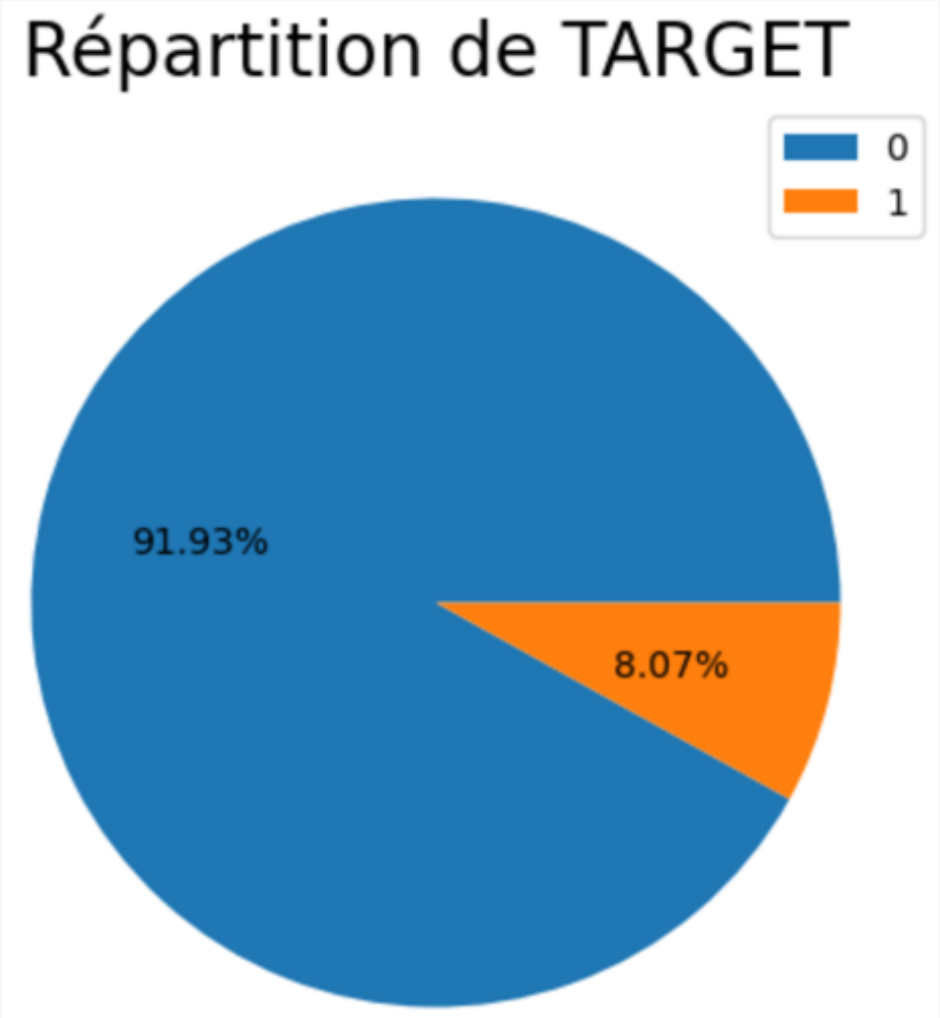
$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

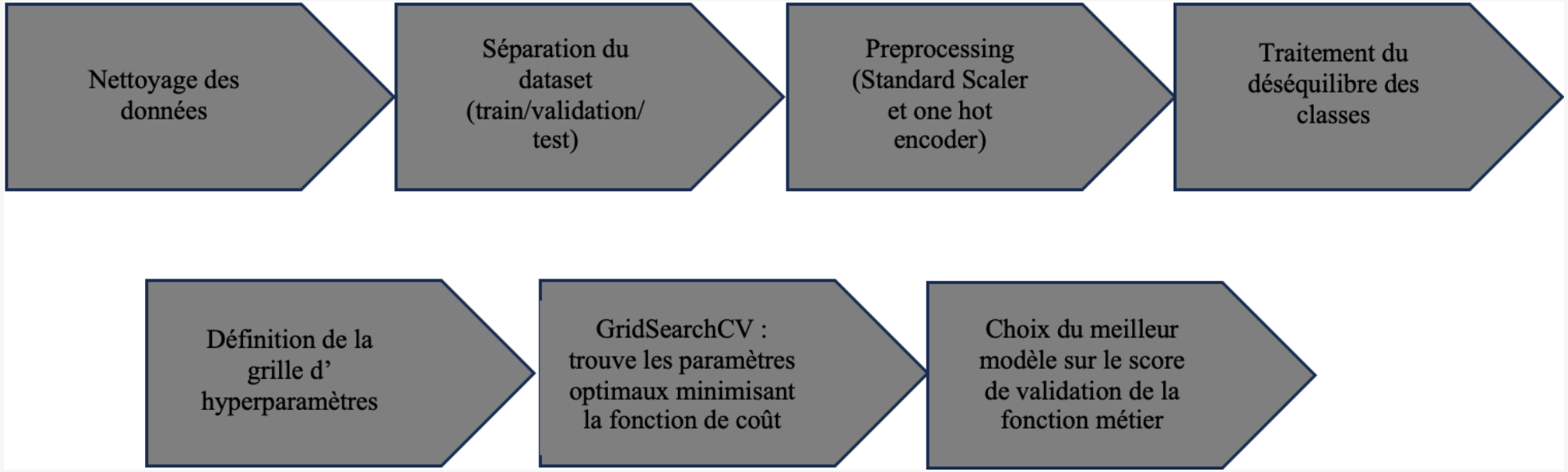
Déséquilibre des classes



- Aucun rééquilibrage
- Class_weight
- SMOTE
- Tomek link
- SMOTETomek
- RandomUnderSampler



Processus de modélisation



Modèles

Run Name:	Logistic regression optimise	Random forest optimise	XGBoost optimise	Light GBM optimise	Decision Tree optimise	SVC optimise
Start Time:	2023-05-31 17:33:51	2023-05-31 19:05:11	2023-06-02 20:36:28	2023-06-05 20:30:04	2023-07-20 19:17:15	2023-07-24 13:
End Time:	2023-05-31 17:33:52	2023-05-31 19:05:25	2023-06-02 20:36:29	2023-06-05 20:30:05	2023-07-20 19:17:16	2023-07-24 13:
Duration:	0.5s	13.9s	0.5s	0.6s	0.9s	3.6min

> Parameters

v Metrics

Show diff only

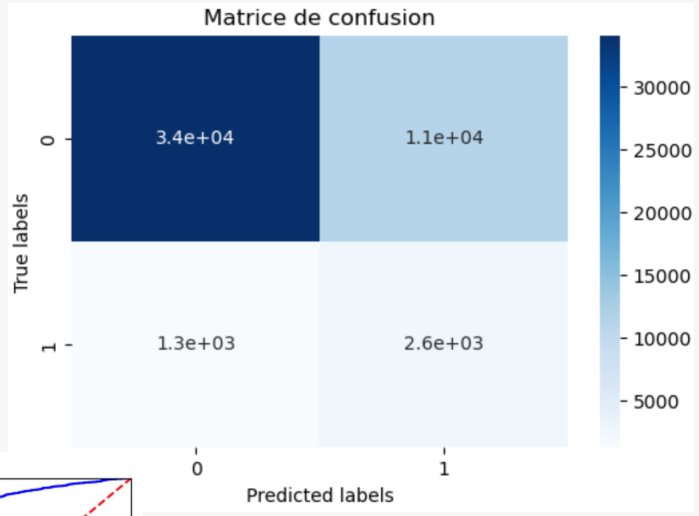
val_accuracy	0.7	0.696	0.709	0.708	0.654	0.735
val_f1_score	0.273	0.268	0.279	0.281	0.226	0.236
val_fbeta_score	0.43	0.424	0.436	0.44	0.367	0.348
val_precision	0.17	0.167	0.174	0.175	0.138	0.154
val_recall	0.699	0.69	0.697	0.707	0.626	0.508
val_rocauc	0.699	0.693	0.703	0.708	0.641	0.632
val_score_métier	0.4	0.39	0.408	0.415	0.301	0.297

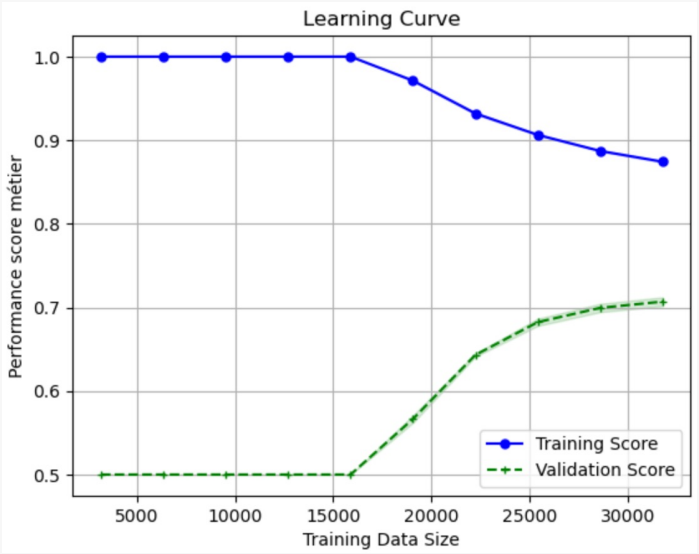
- **Dummy Classifier (baseline)** : renvoie l'étiquette de classe la plus fréquente.
- **Logistic Regression** : relation mathématique entre les variables d'entrée et la variable de sortie.
- **SVC :** trouve un hyperplan optimal qui sépare les données d'entraînement en différentes classes. L'hyperplan est déterminé de manière à maximiser la marge entre les points de données de chaque classe.
- **Decision Tree**: prend un ensemble de données en entrée et construit un modèle prédictif sous forme d'arbre hiérarchique. Chaque nœud de l'arbre représente une caractéristique de l'ensemble de données, chaque branche représente une règle de décision basée sur cette caractéristique, et chaque feuille représente une classe ou une valeur prédite.
- **Random Forest** : combine plusieurs arbres de décision pour effectuer des prédictions ; classe prédite déterminée par un vote majoritaire.
- **XG Boost** : utilise un ensemble de modèles d'arbres de décision pour effectuer des prédictions ; itérations successives pour minimiser une fonction de perte.
- **Light GBM** : algorithme d'apprentissage automatique basé sur le gradient boosting ; technique d'échantillonnage basée sur le gradient.

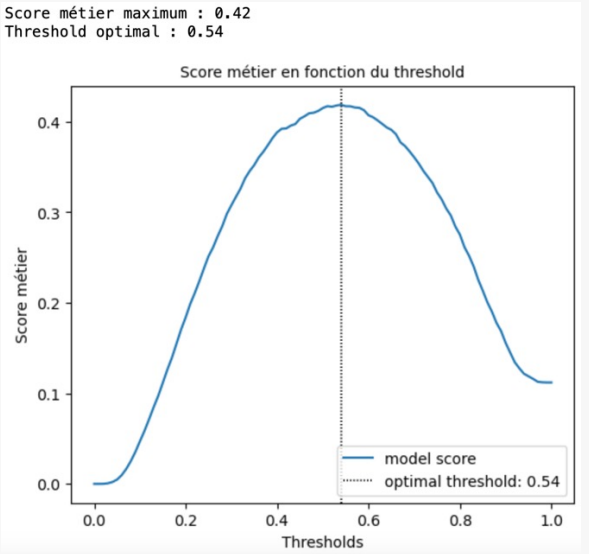
Modèle final

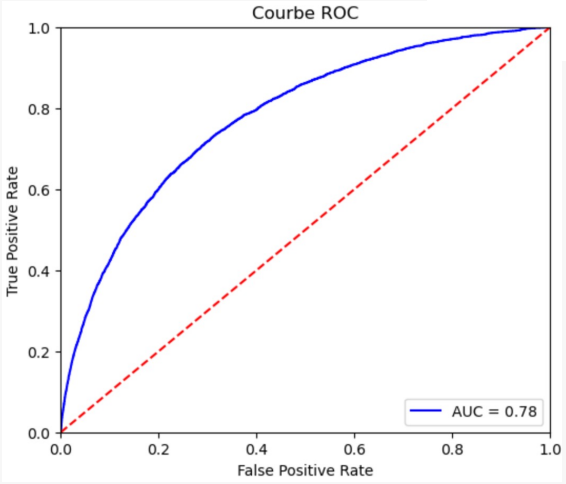
Light GBM

```
Score métier :0.42
Accuracy score : 0.74
Precision score : 0.19
Recall score : 0.66
F1 score : 0.30
Fbeta score : 0.44
ROC AUC score : 0.71
```





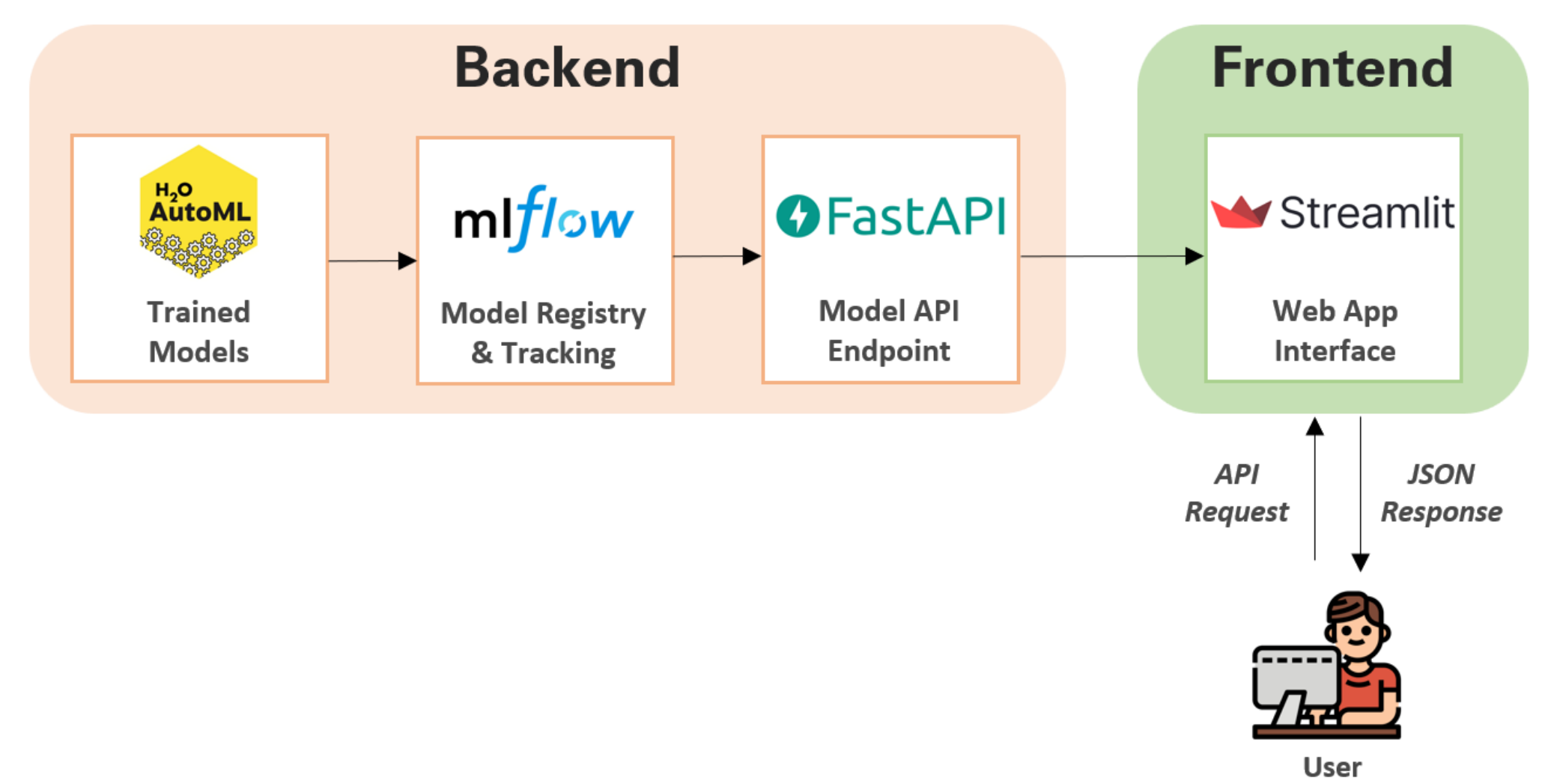





```
params = {'n_estimators' : [100, 300, 500, 800],
          'max_depth': [-1, 2, 5, 7],
          'num_leaves': [7, 15, 31, 63, 127],
          'learning_rate': [0.05, 0.1, 0.2, 0.4]
         }
```

Run Name:	pip final	pip jeu test
Start Time:	2023-06-26 19:38:03	2023-06-26 19:49:23
End Time:	2023-06-26 19:38:26	2023-06-26 19:49:24
Duration:	22.7s	434ms
val_accuracy	0.745	0.742
val_f1_score	0.295	0.293
val_fbeta_score	0.443	0.44
val_precision	0.19	0.188
val_recall	0.663	0.663
val_rocauc	0.707	0.706
val_score_métier	0.419	0.416

III – Pipeline de déploiement


Backend
Frontend
H2O AutoML
Trained Models
mlflow
Model Registry & Tracking
FastAPI
Model API Endpoint
Streamlit
Web App Interface
API Request
JSON Response
User

```
(venv) (base) macbook-pro:GitHub oceaneyouyoutte$ pytest test_api.py
============================= test session starts ==============================
platform darwin -- Python 3.9.6, pytest-7.4.0, pluggy-1.2.0
rootdir: /Users/oceaneyouyoutte/Desktop/Data Science/OCR/Projet 7/GitHub
plugins: anyio-3.7.0
collected 4 items

test_api.py ....                                                         [100%]

============================== 4 passed in 0.24s ===============================
```

```
(venv) (base) macbook-pro:GitHub oceaneyouyoutte$ pytest test_dashboard.py
============================= test session starts ==============================
platform darwin -- Python 3.9.6, pytest-7.4.0, pluggy-1.2.0
rootdir: /Users/oceaneyouyoutte/Desktop/Data Science/OCR/Projet 7/GitHub
plugins: anyio-3.7.0
collected 2 items

test_dashboard.py ..
```

Lien GitHub : https://github.com/OceaneYYT/P7-OCR
Lien API : https://p7-ocr-fastapi-95768180a01f.herokuapp.com/
Lien Dashboard : https://p7-ocr-dashboard-a790d1a0f622.herokuapp.com/

IV – Data Drift

[redacted]

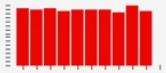
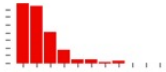
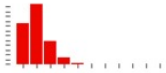
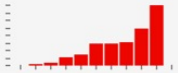
Dataset Drift

Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5

121	10	0.0826
Columns	Drifted Columns	Share of Drifted Columns

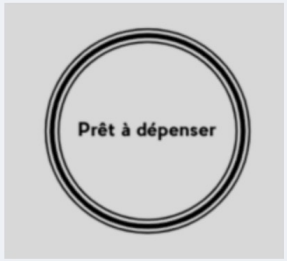
Drift is detected for 8.264% of columns (10 out of 121).

Search

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
SK_ID_CURR	num			Detected	Wasserstein distance (normed)	9.18745
AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.47302
AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.280117
AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.212161
AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.209213
AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.160231
AMT_REQ_CREDIT_BUREAU_WEEK	num			Detected	Wasserstein distance (normed)	0.152966
NAME_CONTRACT_TYPE	cat			Detected	Jensen-Shannon distance	0.14678
DAYS_LAST_PHONE_CHANGE	num			Detected	Wasserstein distance (normed)	0.137271
FLAG_EMAIL	num			Detected	Jensen-Shannon distance	0.123534

V – Dashboard

p7-ocr-dashboard-a790d1a0f622.herokuapp.com



Navigation

Home

ID Client

<Select>

Vous avez choisi le client ID : <Select>

Created by Océane Youyoutte

Dashboard Prêt à dépenser - Home Page

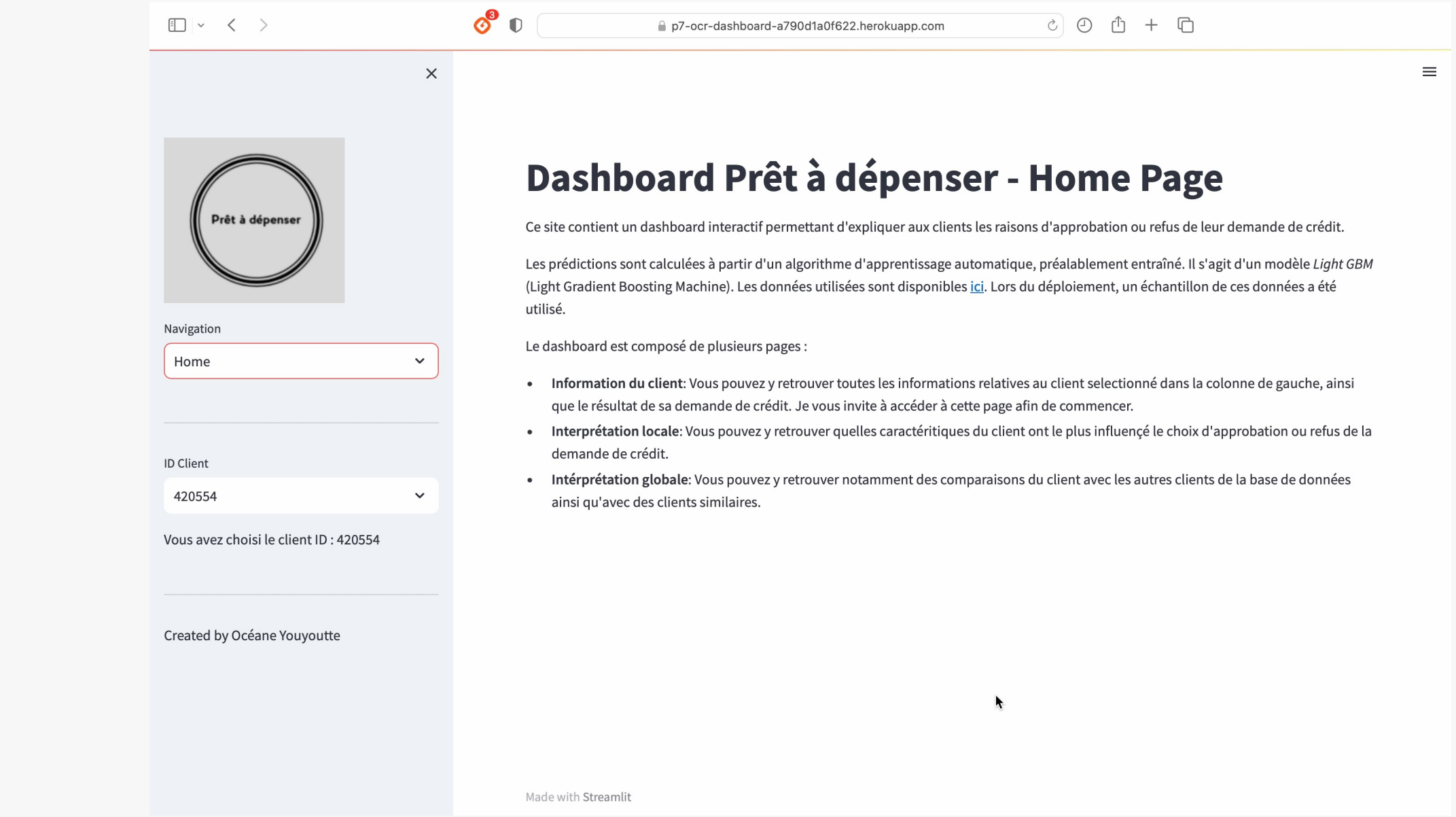
Ce site contient un dashboard interactif permettant d'expliquer aux clients les raisons d'approbation ou refus de leur demande de crédit.

Les prédictions sont calculées à partir d'un algorithme d'apprentissage automatique, préalablement entraîné. Il s'agit d'un modèle *Light GBM* (Light Gradient Boosting Machine). Les données utilisées sont disponibles ici. Lors du déploiement, un échantillon de ces données a été utilisé.

Le dashboard est composé de plusieurs pages :

- **Information du client**: Vous pouvez y retrouver toutes les informations relatives au client selectionné dans la colonne de gauche, ainsi que le résultat de sa demande de crédit. Je vous invite à accéder à cette page afin de commencer.
- **Interprétation locale**: Vous pouvez y retrouver quelles caractéritiques du client ont le plus influencé le choix d'approbation ou refus de la demande de crédit.
- **Intérprétation globale**: Vous pouvez y retrouver notamment des comparaisons du client avec les autres clients de la base de données ainsi qu'avec des clients similaires.

Made with Streamlit


p7-ocr-dashboard-a790d1a0f622.herokuapp.com
Prêt à dépenser
Navigation
Home
ID Client
420554
Vous avez choisi le client ID : 420554
Created by Océane Youyoutte
Dashboard Prêt à dépenser - Home Page
Ce site contient un dashboard interactif permettant d'expliquer aux clients les raisons d'approbation ou refus de leur demande de crédit.
Les prédictions sont calculées à partir d'un algorithme d'apprentissage automatique, préalablement entraîné. Il s'agit d'un modèle Light GBM (Light Gradient Boosting Machine). Les données utilisées sont disponibles ici. Lors du déploiement, un échantillon de ces données a été utilisé.
Le dashboard est composé de plusieurs pages :
Information du client: Vous pouvez y retrouver toutes les informations relatives au client selectionné dans la colonne de gauche, ainsi que le résultat de sa demande de crédit. Je vous invite à accéder à cette page afin de commencer.
Interprétation locale: Vous pouvez y retrouver quelles caractéritiques du client ont le plus influençé le choix d'approbation ou refus de la demande de crédit.
Intérprétation globale: Vous pouvez y retrouver notamment des comparaisons du client avec les autres clients de la base de données ainsi qu'avec des clients similaires.
Made with Streamlit