

Spécialité : Data Scientist

PROJET 7 : IMPLÉMENTEZ UN MODÈLE DE SCORING

Soutenance de :
Fatoumata Binta DIALLO

Data Scientist au sein de "**Prêt à dépenser**", société financière proposant des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt ;

➔ Problème de **classification supervisée binaire** à résoudre

Missions:

- ☐ Construire un modèle de scoring de prédiction de la probabilité de défaut de paiement d'un client.
- ☐ Développer un dashboard interactif pour l'aide à la prise de décision.

Données:

- ☐ Historiques des clients de la société financière disponible sur le lien suivant: <https://www.kaggle.com/c/home-credit-default-risk/data>

I/ Description des données et méthodologie utilisée

II/versionnage des codes avec git/github

III / Présentation du tableau de bord et de son fonctionnement

IV/ Conclusion

I. DESCRIPTION DES DONNÉES ET MÉTHODOLOGIE UTILISÉE

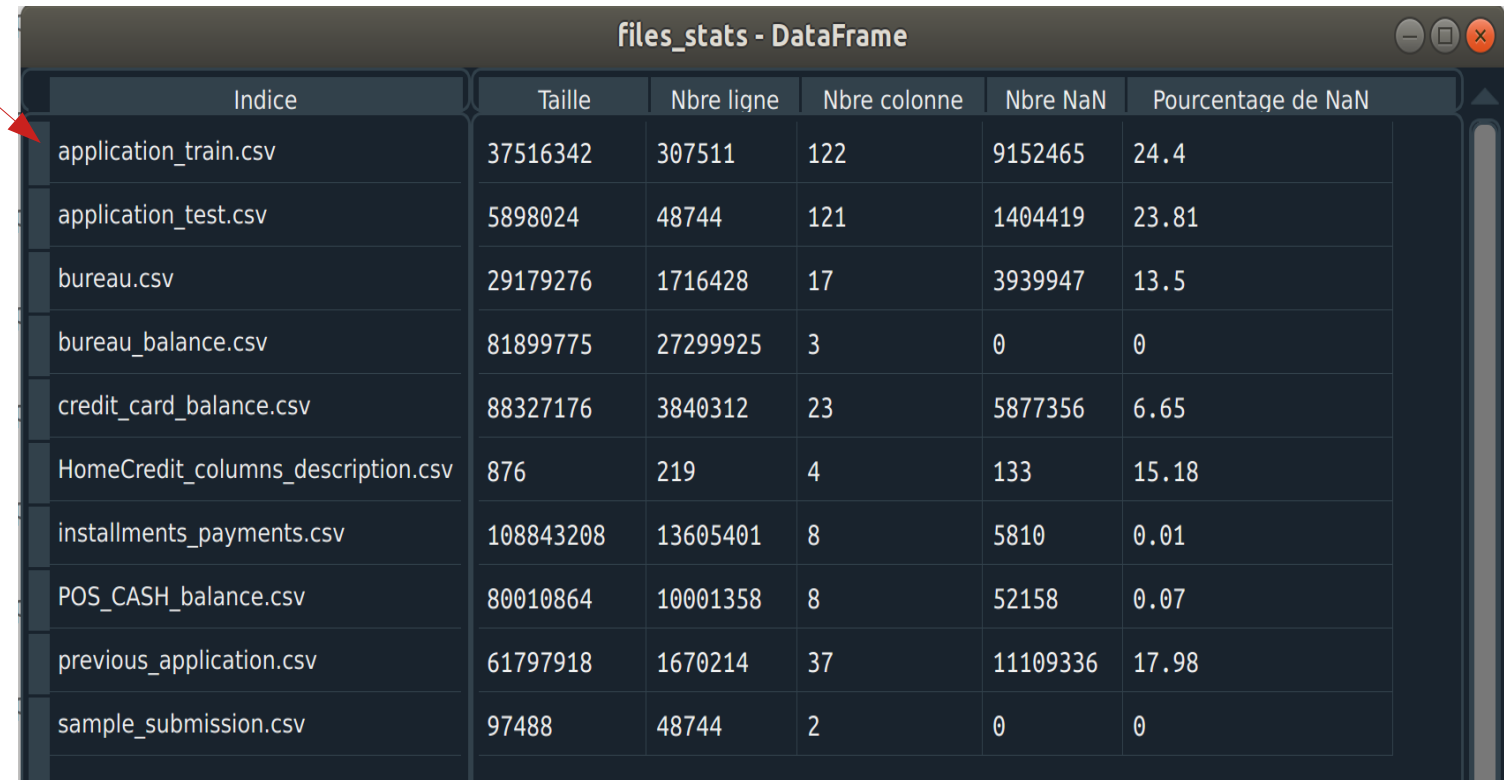
- 10 fichiers .csv hébergés sur **Kaggle**

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

➔ **Fichier d'entraînement :**

*variable cible **"TARGET"**

- [0 pour client en règle
1 sinon]




Indice	Taille	Nbre ligne	Nbre colonne	Nbre NaN	Pourcentage de NaN
application_train.csv	37516342	307511	122	9152465	24.4
application_test.csv	5898024	48744	121	1404419	23.81
bureau.csv	29179276	1716428	17	3939947	13.5
bureau_balance.csv	81899775	27299925	3	0	0
credit_card_balance.csv	88327176	3840312	23	5877356	6.65
HomeCredit_columns_description.csv	876	219	4	133	15.18
installments_payments.csv	108843208	13605401	8	5810	0.01
POS_CASH_balance.csv	80010864	10001358	8	52158	0.07
previous_application.csv	61797918	1670214	37	11109336	17.98
sample_submission.csv	97488	48744	2	0	0

- Contenu:

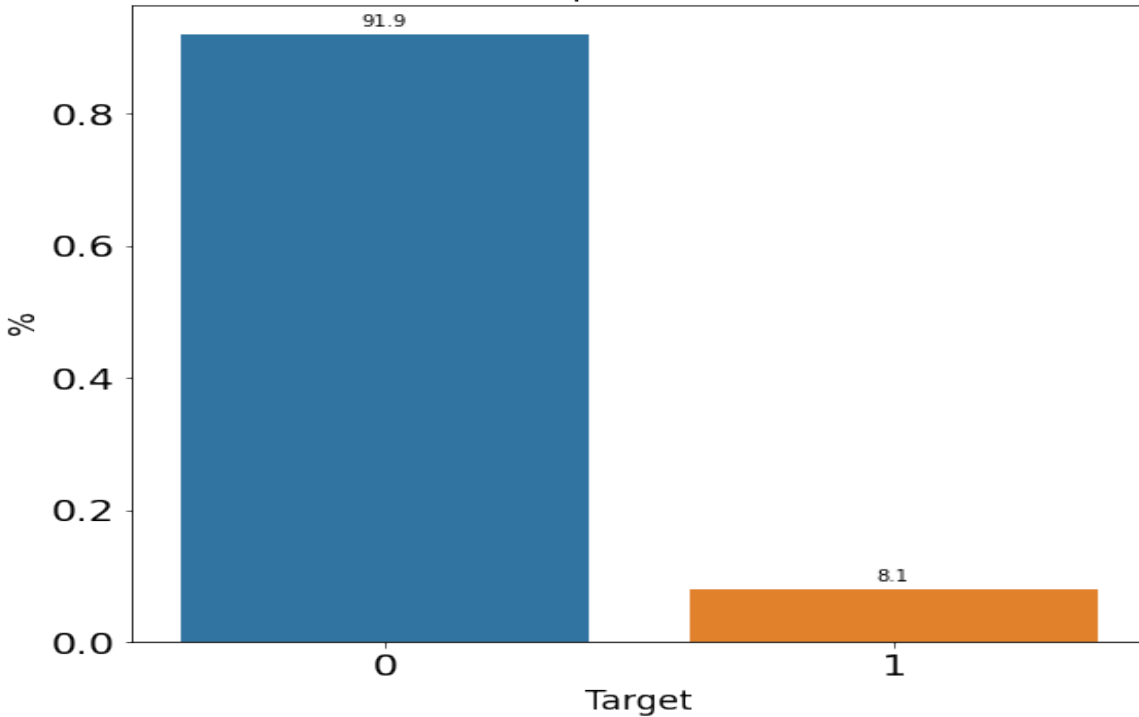
- ◆ Informations générales sur le client (Age, sexe, statut familiale,...)
- ◆ Informations relatives au crédit (montant du crédit, nombre 'annuité,...)

Travaux:

- Analyse succincte des données
- Prétraitement des données  Kernel Kaggle ([LightGBM with Simple Features](#))
 - ➔ Bon score dans la compétition (1900)
 - ➔ Feature engineering performant
 - ➔ Codage fonction pour entraîner des données avec LightGM
- Traitement des valeurs manquantes
 - ➔ Fichier finale : (307507, 608)

Analyse de la variable 'TARGET'

Répartition

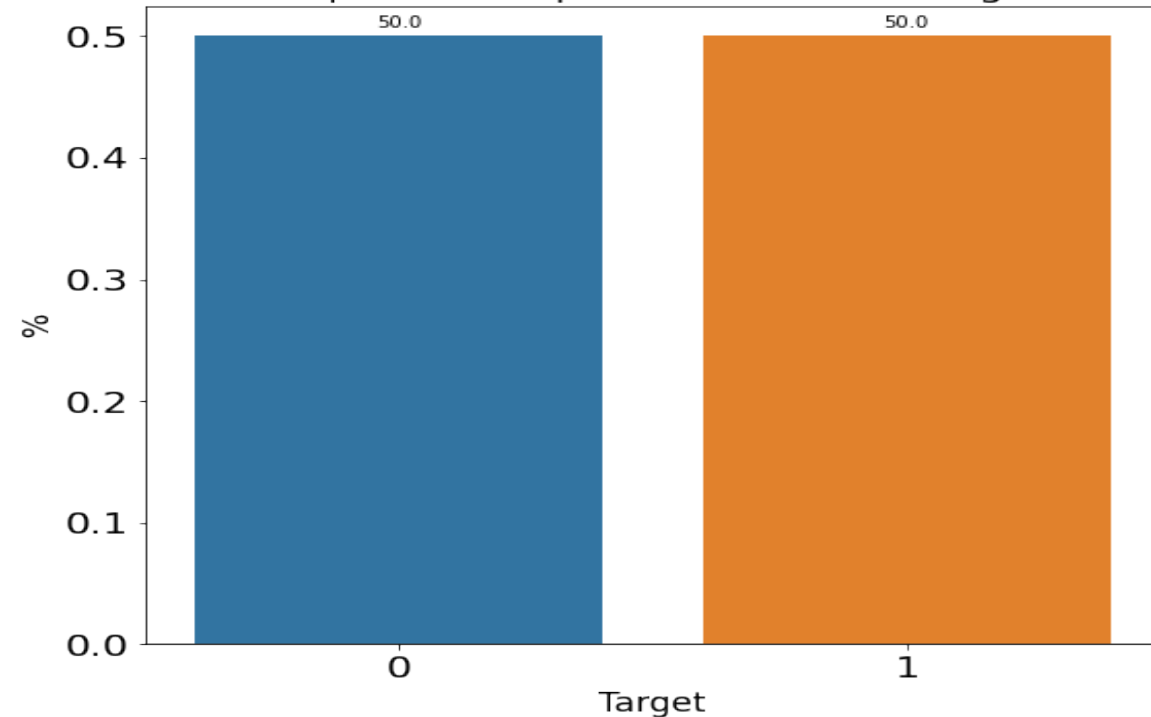


- Jeu de données **Déséquilibré**
 - Erreur dans la prédiction

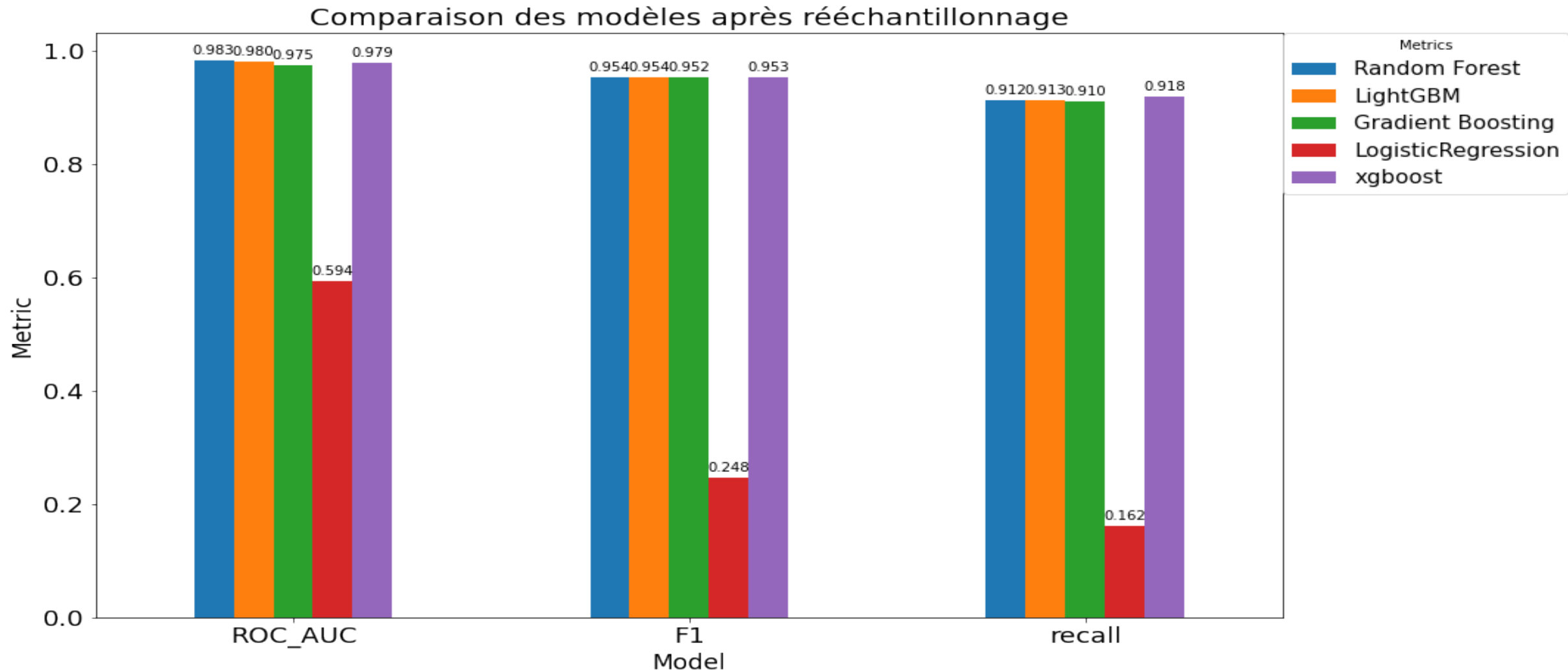


Algorithme de re-échantillonnage
SMOTE

Répartition après rééchantillonnage

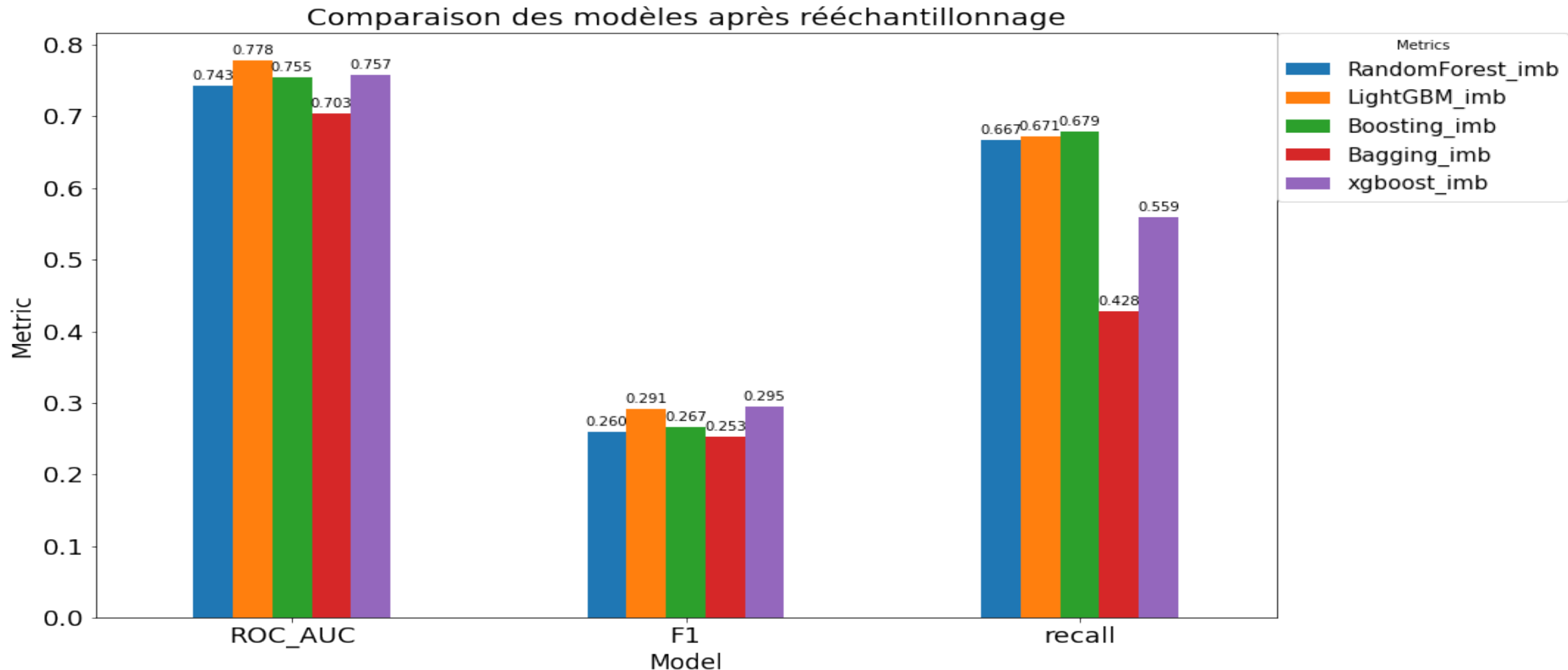


Recherche du meilleur modèle :



- ♦ La méthode de rééchantillonnage avant l'entraînement créer un overfitting.
Idée : Utilisation de modèles avec échantillonneurs d'équilibrage intérieurs

Recherche du meilleur modèle :



- ◆ Les résultats montrent que les modèles ne sont plus en surapprentissage
- ◆ LightGBM avec équilibrage interne automatique de rééchantillonnage est celui qui présente le meilleur résultat

II. VERSIONNING DES CODES AVEC Git/GitHub



https://github.com/DIALLOFatoumataBinta/Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO

- ☐ Création compte sur GitHub : **DIALLOFatoumataBinta**
- ☐ Création du projet sur 'repository': **Projet_7_Openclassroom**
- ☐ Initialisation de Git (configuration d'identité) et du dépôt Git
- ☐ Indexer et commiter vos fichiers
- ☐ Envoie du commit sur le dépôt distant par **commande ssh**
- ☐ Création de plusieurs branches dont une pour le déploiement

https://github.com/DIALLOFatoumataBinta/Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO

← → ↺

https://github.com/DIALLOFatoumataBinta/Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO/tree/fichierPDF

☆

🔍 Rechercher

⚙️ Les plus visités

🌐 /home/bdiallo/Docum...

🗨️ Qu'est-ce que le méti...

📖 Getting Started

🌐 carte consulaire - Rec...

🌐 no way

🌐 Images correspondan...

📄 Brouillon - Jupyter No...

👤 2

...

🐙

Search or jump to...

/

Pull requests

Issues

Marketplace

Explore

DIALLOFatoumataBinta / **Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO** Public

📌 Pin

👁️ Unwatch 1

🔗 Fork

<> Code

🔍 Issues

🔗 Pull requests

🔄 Actions

📁 Projects

📖 Wiki

🔒 Security

📈 Insights

⚙️ Settings

🔔 fichierPDF had recent pushes less than a minute ago

Compare & pull request

🔗 fichierPDF ▾

🌿 2 branches

🏷️ 0 tags

Go to file

Add file ▾

Code ▾

This branch is 1 commit ahead, 7 commits behind main.

🔗 Contribute ▾

🐙 DIALLOFatoumataBinta note

e564730 1 minute ago 🕒 2 commits

📁 DATA	deploiement rattrapage	3 hours ago
📁 app	deploiement rattrapage	3 hours ago
📄 DIALLO_Fatoumata_note_methodol...	note	1 minute ago
📄 LightGBM_imb.pkl	deploiement rattrapage	3 hours ago
📄 Procfile	deploiement rattrapage	3 hours ago
📄 logo.png	deploiement rattrapage	3 hours ago
📄 requirements.txt	deploiement rattrapage	3 hours ago
📄 runtime.txt	deploiement rattrapage	3 hours ago
📄 setup.sh	deploiement rattrapage	3 hours ago
📄 wsgi.py	deploiement rattrapage	3 hours ago

🔔 Help people interested in this repository understand your project by adding a README.

Add a README

About

P7 definitif

☆ 0 stars

👁️ 1 watching

🔗 0 forks

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Environments 1

🚀 diallo-p7-heroku-oc-streamlit Active

Languages

Python 98.4%

Shell 1.3%

Procfile 0.3%

III. PRÉSENTATION DU TABLEAU DE BORD ET DE SON FONCTIONNEMENT

**HEROKU**

Rajout des fichiers sur **GitHub** :
Procfile ligne de commande lance **setpu.sh** et **streamlit**
requirements.txt – librairie utilisée
runtime.txt – ma version de python
setup.sh – **setup des credentials** puis le port
wsgi.py **appel le main**

- ☐ Création de son compte sur **heroku**
- ☐ Création du nom de l'application sur '**heroku**': **diallo-p7-heroku-oc-streamlit**
- ☐ Connexion avec Github
- ☐ Choix de la branche à déployer
- ☐ Déploiement et visualisation : <https://diallo-p7-heroku-oc-streamlit.herokuapp.com/>



HEROKU

Salesforce Platform



HEROKU

Jump to Favorites, Apps, Pipelines, Spaces...



Personal > diallo-p7-heroku-oc-streamlit



Open app

More ▾

GitHub DIALLOFatoumataBinta/Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO

[Overview](#) [Resources](#) [Deploy](#) [Metrics](#) [Activity](#) [Access](#) [Settings](#)

Add this app to a pipeline

Create a new pipeline or choose an existing one and add this app to a stage in it.

Add this app to a stage in a pipeline to enable additional features



Pipelines let you connect multiple apps together and **promote code** between them. [Learn more.](#)



Pipelines connected to GitHub can enable **review apps**, and create apps for new pull requests. [Learn more.](#)

Choose a pipeline ▾

Deployment method

Heroku Git
Use Heroku CLIGitHub
ConnectedContainer Registry
Use Heroku CLI

App connected to GitHub

Code diffs, manual and auto deploys are available for this app.

Connected to

[DIALLOFatoumataBinta/Openclassrooms_P7_deploiement_heroku_streamlit_DIALLO](#) by [DIALLOFatoumataBinta](#)

Disconnect...

Releases in the [activity feed](#) link to GitHub to view commit diffs

Automatic deploys

Enables a chosen branch to be automatically deployed to this app.



You can now change your main deploy branch from "master" to "main" for both manual and automatic deploys, please follow the instructions [here](#).

Enable automatic deploys from GitHub



HEROKU



← → ↻ 🔒 🔍 https://dashboard.heroku.com/apps/diallo-p7-heroku-oc-streamlit/deploy/github



Rechercher



⚙️ Les plus visités 🌐 /home/bdiallo/Docum... 🗣️ Qu'est-ce que le mét... 🎯 Getting Started 🌐 carte consulaire - Rec... 🌐 no way 🌐 Images correspon... 📄 Brouillon - Jupyter No... 📄 2 ... 📁 Autres marque-pages

☁️ Salesforce Platform



HEROKU

Jump to Favorites, Apps, Pipelines, Spaces...



Automatic deploys

Enables a chosen branch to be automatically deployed to this app.



You can now change your main deploy branch from "master" to "main" for both manual and automatic deploys, please follow the instructions [here](#).

Enable automatic deploys from GitHub

Every push to the branch you specify here will deploy a new version of this app. **Deploys happen automatically:** be sure that this branch is always in a deployable state and any tests have passed before you push. [Learn more](#).

Choose a branch to deploy

🔗 main

☐ Wait for CI to pass before deploy

Only enable this option if you have a Continuous Integration service configured on your repo.

Enable Automatic Deploys

Manual deploy

Deploy the current state of a branch to this app.

Deploy a GitHub branch

This will deploy the current state of the branch you specify below. [Learn more](#).

Choose a branch to deploy

🔗 main

Deploy Branch

Receive code from GitHub



Build main 6d5b752c



Release phase



Deploy to Heroku



Your app was successfully deployed.


📄 View

https://diallo-p7-heroku-oc-streamlit.herokuapp.com

Rechercher

visités /home/bdiallo/Docum... Qu'est-ce que le méti... Getting Started carte consulaire - Rec... no way Images correspondant... Brouillon - Jupyter No... 2 ... Autres marque-pages

×



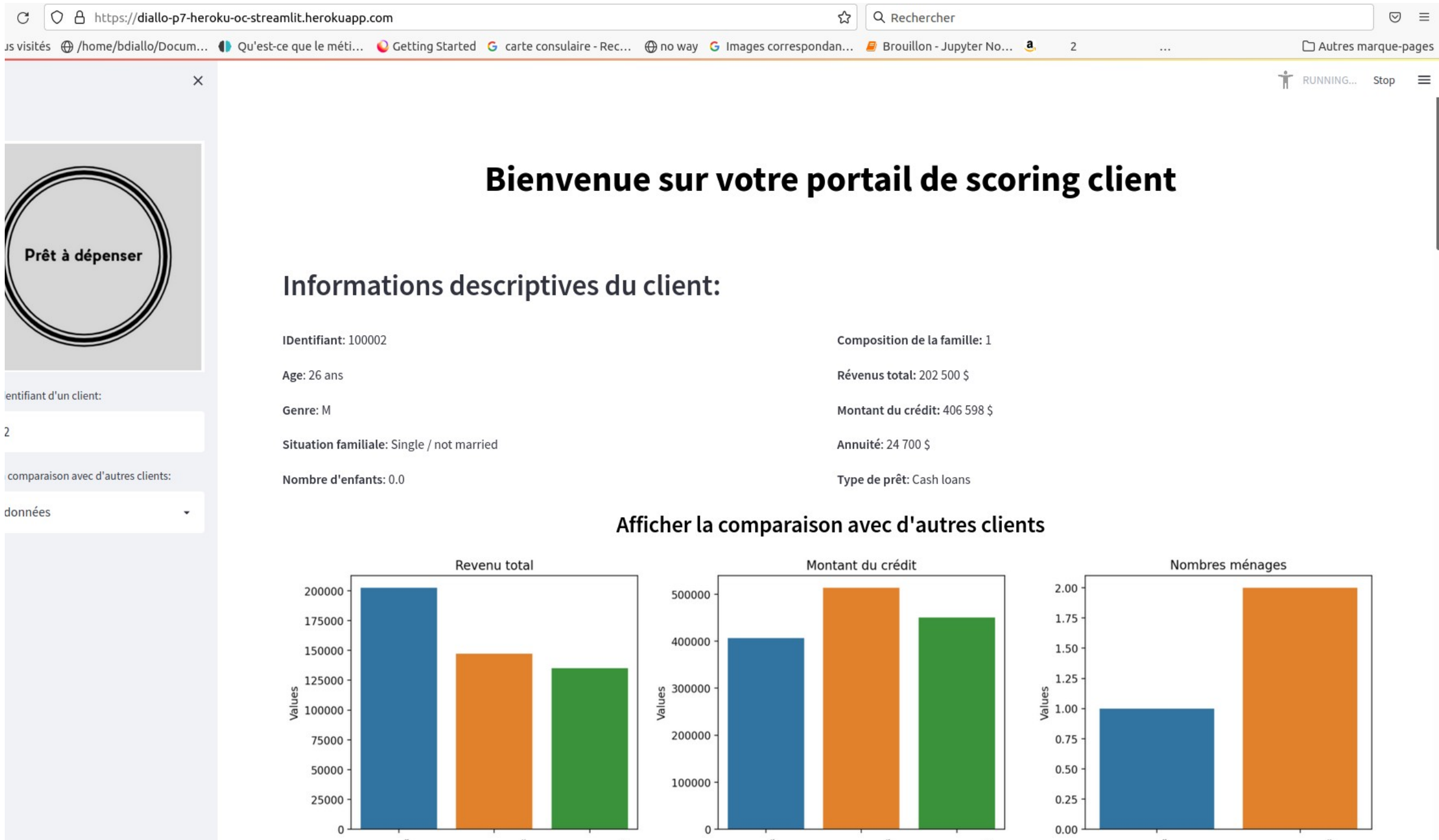
Prêt à dépenser

Bienvenue sur votre portail de scoring client

S'il vous plait entrez un identifiant correct.

ntifiant d'un client:

Made with Streamlit

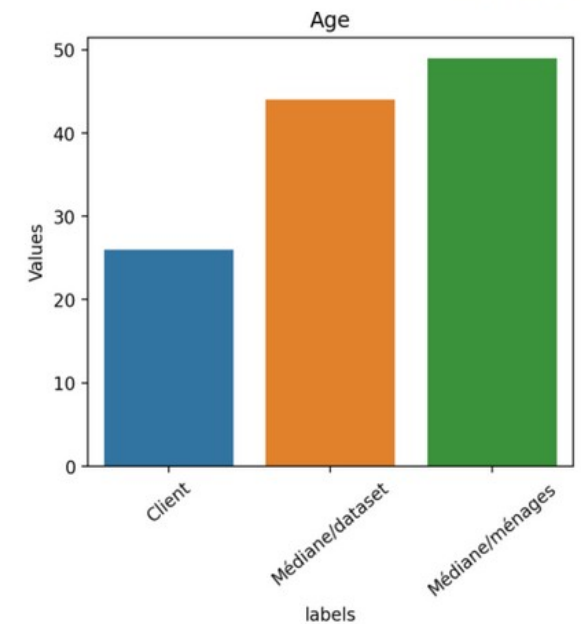
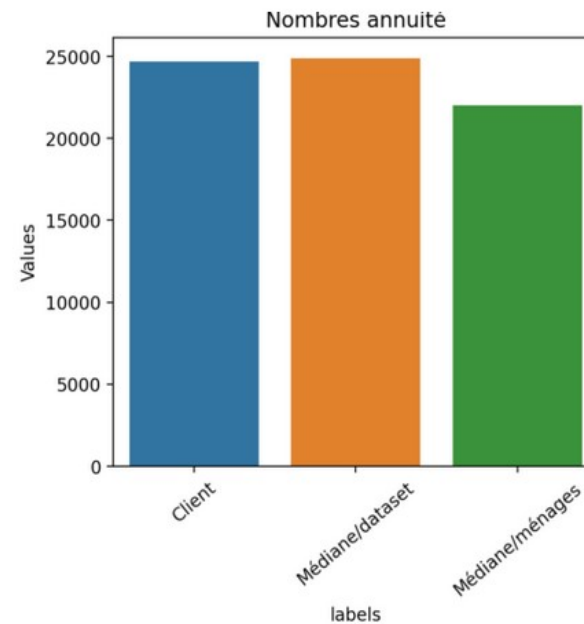


Prêt à dépenser

entifiant d'un client:

comparaison avec d'autres clients:

données

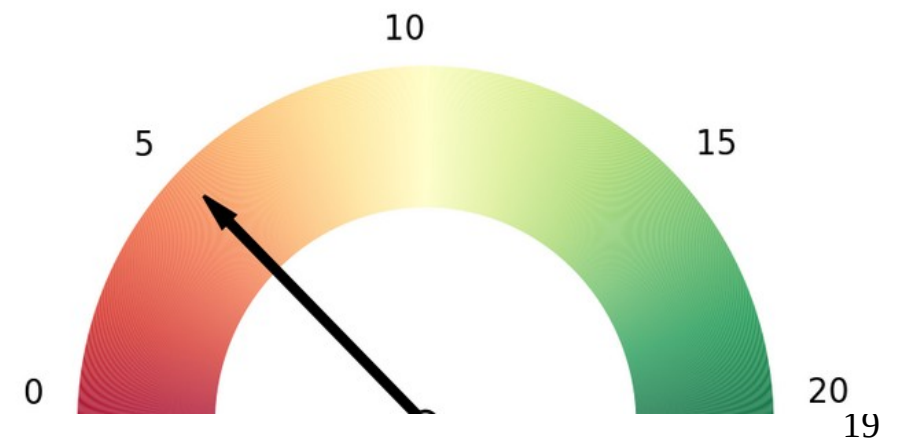


Prédiction:

Ce client a une note de 05/20 pour rembourser son crédit. Le risque de défaut de paiement de ce client est **élevé**.

Pour rappel, ce client a été en défaut de paiement auparavant.

NB: Le seuil de 14/20 a été défini pour évaluer le niveau du risque de défaut de paiement d'un client: pour une note **inférieure à 14/20**, le risque est **élevé** et pour une note **supérieure à 14/20** le risque est **faible**. Plus la note se rapproche de 20/20 plus le risque est faible et plus la note est faible plus le client est risqué.



IV. CONCLUSION

Travailler sur le projet 7 m'a permis :

- d'étudier un problème de classification binaire et de créer un modèle de 'scoring'
- de faire du versionning de code avec Git/GitHub
- de comprendre le concept d'API et le déploiement de modèle
- de comprendre comment faire une interprétabilité locale et globale avec LIME et 'Features importances'
- d'utiliser Flask et Streamlit pour créer une web application (dashboard)
- d'utiliser heroku connecté à GitHub pour le déploiement de mon application

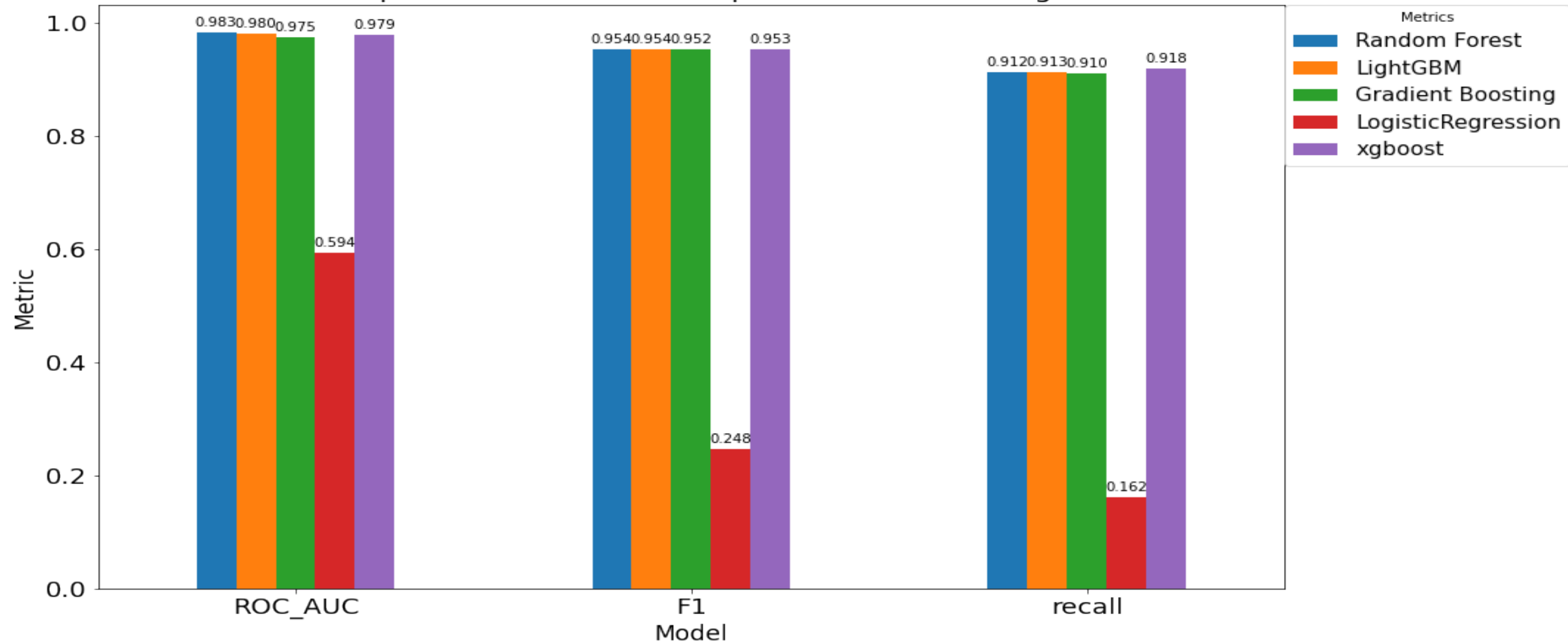


MERCI POUR VOTRE ATTENTION

Échantillonnage des données :
80 % ~> training & 20 % ~> testing

Recherche du meilleur modèle :

Comparaison des modèles après rééchantillonnage



- ◆ Métriques **ROC_AUC** & **F1** : **RandomForest**
- ◆ Métrique **recall** : **Xgboost**