

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:

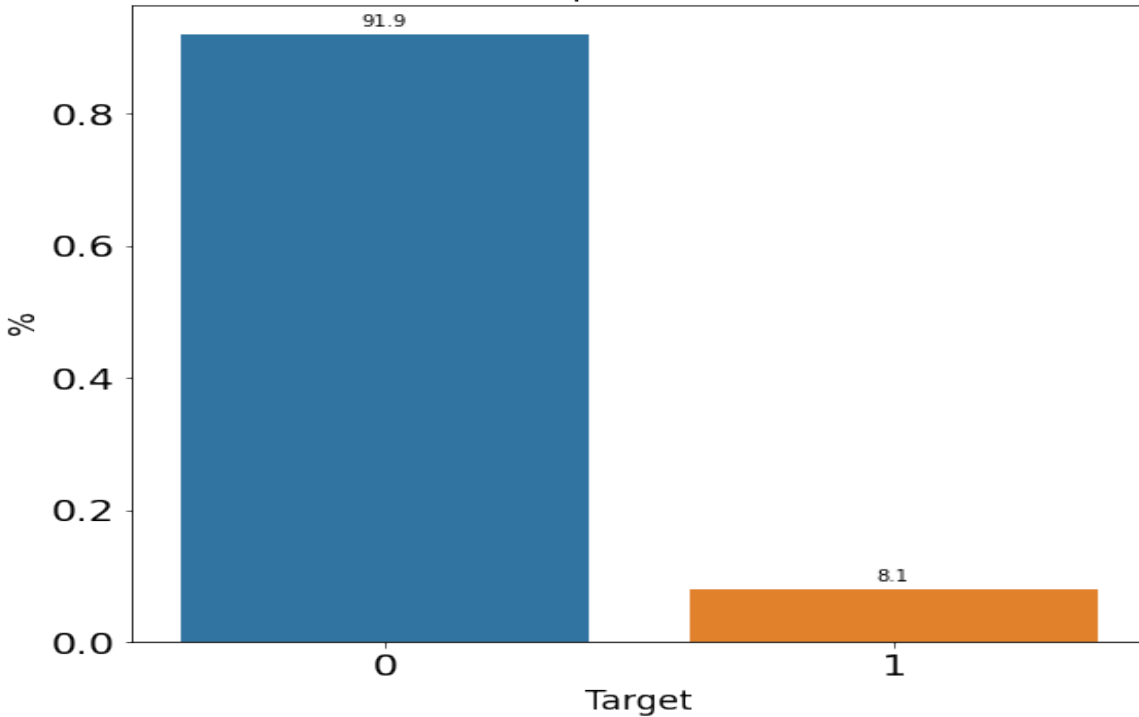
- ◆ Informations générales sur le client (Age, sexe, statut familiale,...)
- ◆ Informations relatives au crédit (crédit, annuité,..)

Travaux:

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

Analyse de la variable 'TARGET'

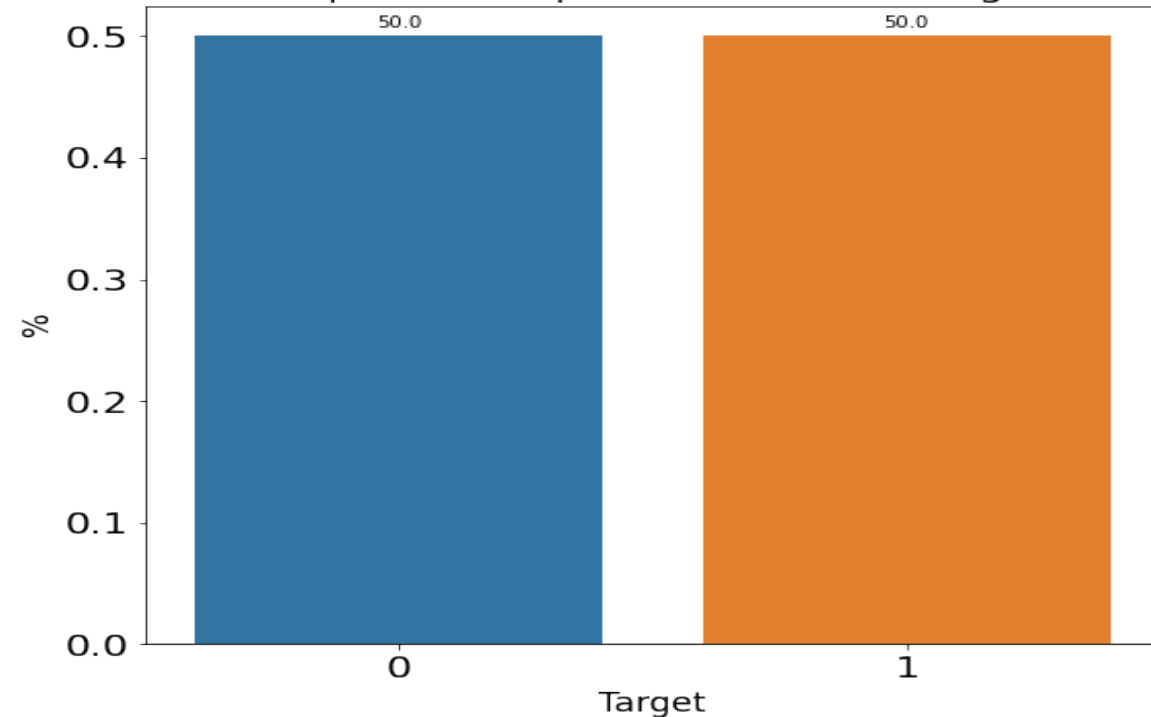
Répartition



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

Algorithme de re-échantillonnage
SMOTE

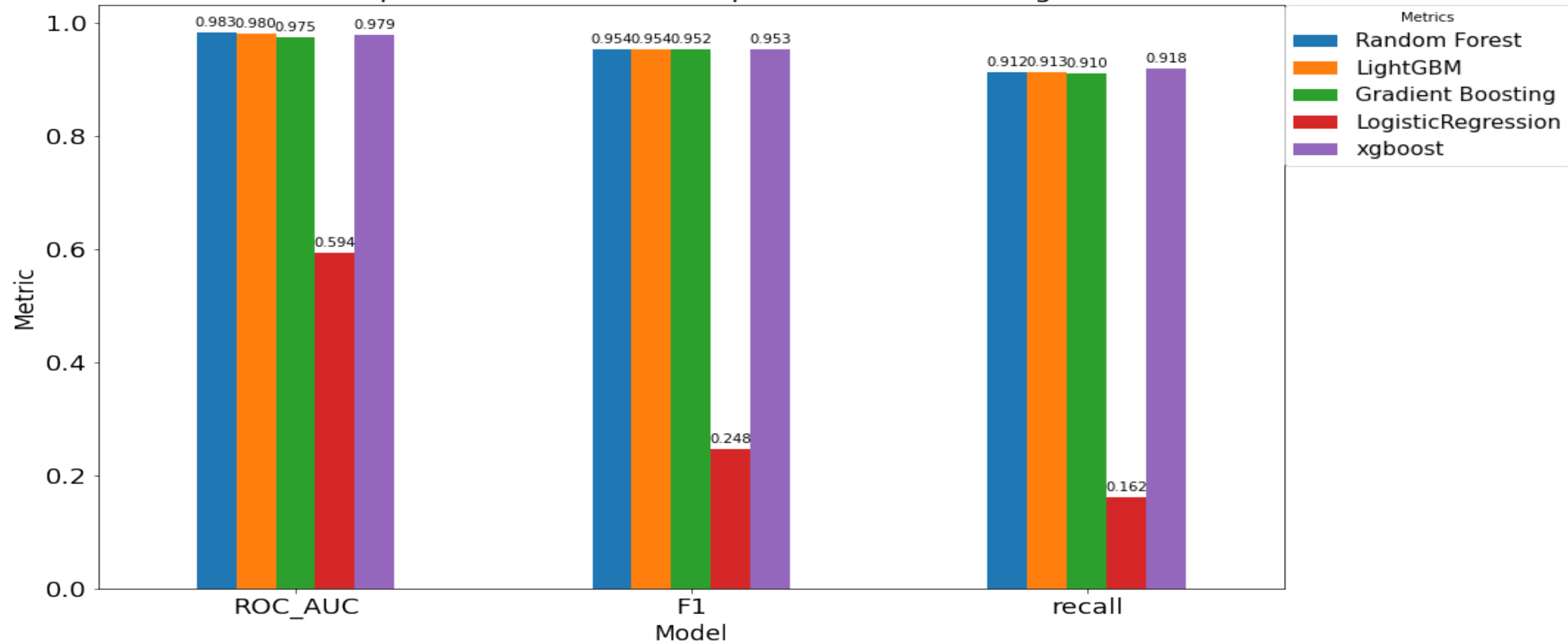
Répartition après rééchantillonnage



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

Recherche du meilleur modèle :

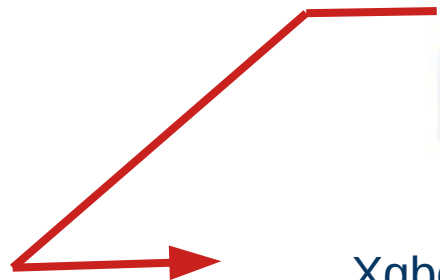
Comparaison des modèles après rééchantillonnage



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

Optimisation par l'adaptation des hyperparamètres :

	Models	Training_set	Validation_set_f1	Validation_set_roc_auc	Validation_set_recall
0	LightGBM	0.975941	0.949833	0.956190	0.999341
1	xgboost	0.983879	0.956049	0.960144	0.994346
2	RandomForest	0.979573	0.953769	0.958282	0.994059



Xgboost est le modèle qui présente le mieux résultats



Xgboost sera utilisé comme modèle de prédiction

II. VERSIONNING DES CODES AVEC Git/GitHub



https://github.com/DIALLOFatoumataBinta/Projet_7_Openclassroom

- ☐ 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

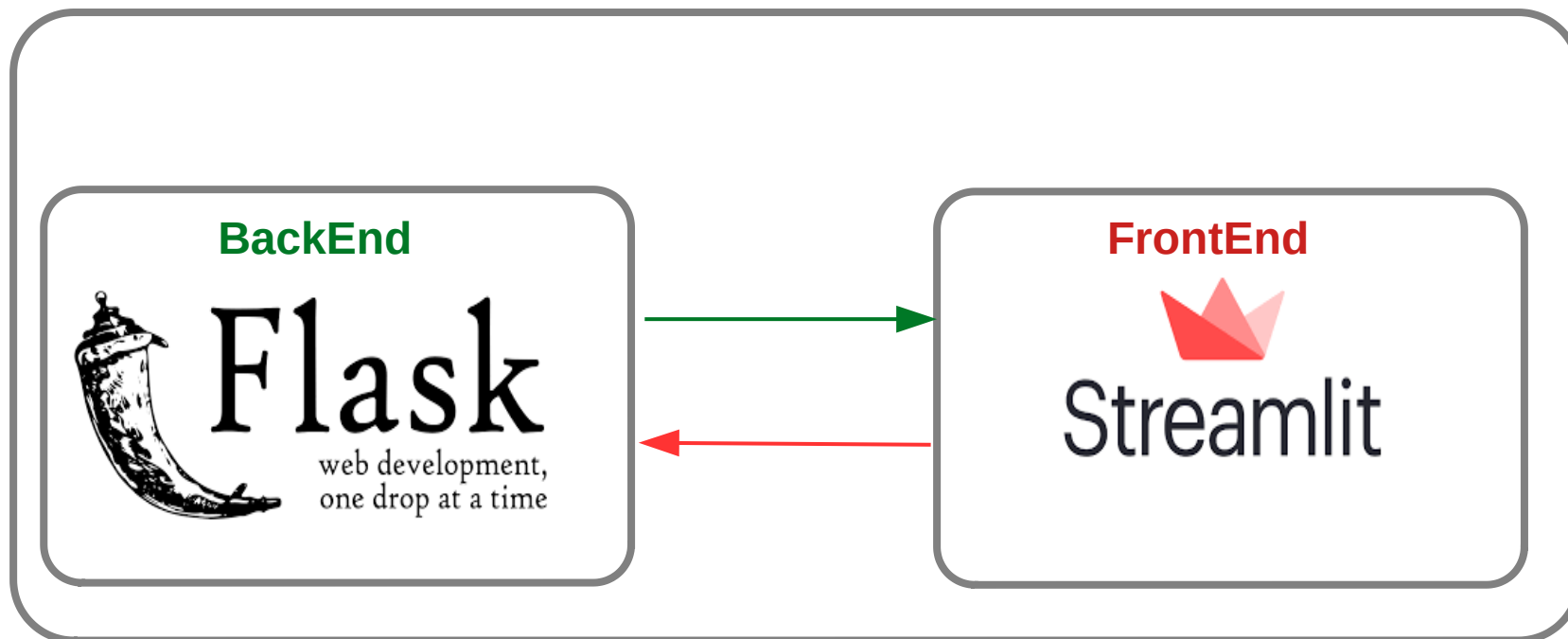
III. PRÉSENTATION DU TABLEAU DE BORD ET DE SON FONCTIONNEMENT



API de prédiction du score



Tableau de bord du projet (dashboard)



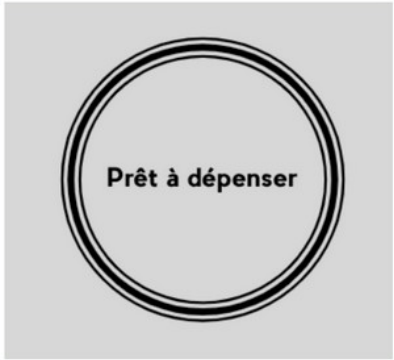
Navigation Web Firefox 11 juil. 16:31

P7_DIALLO_Fatoumata/DIALLO_Fatoumata_dash

localhost:8501

Rechercher

×



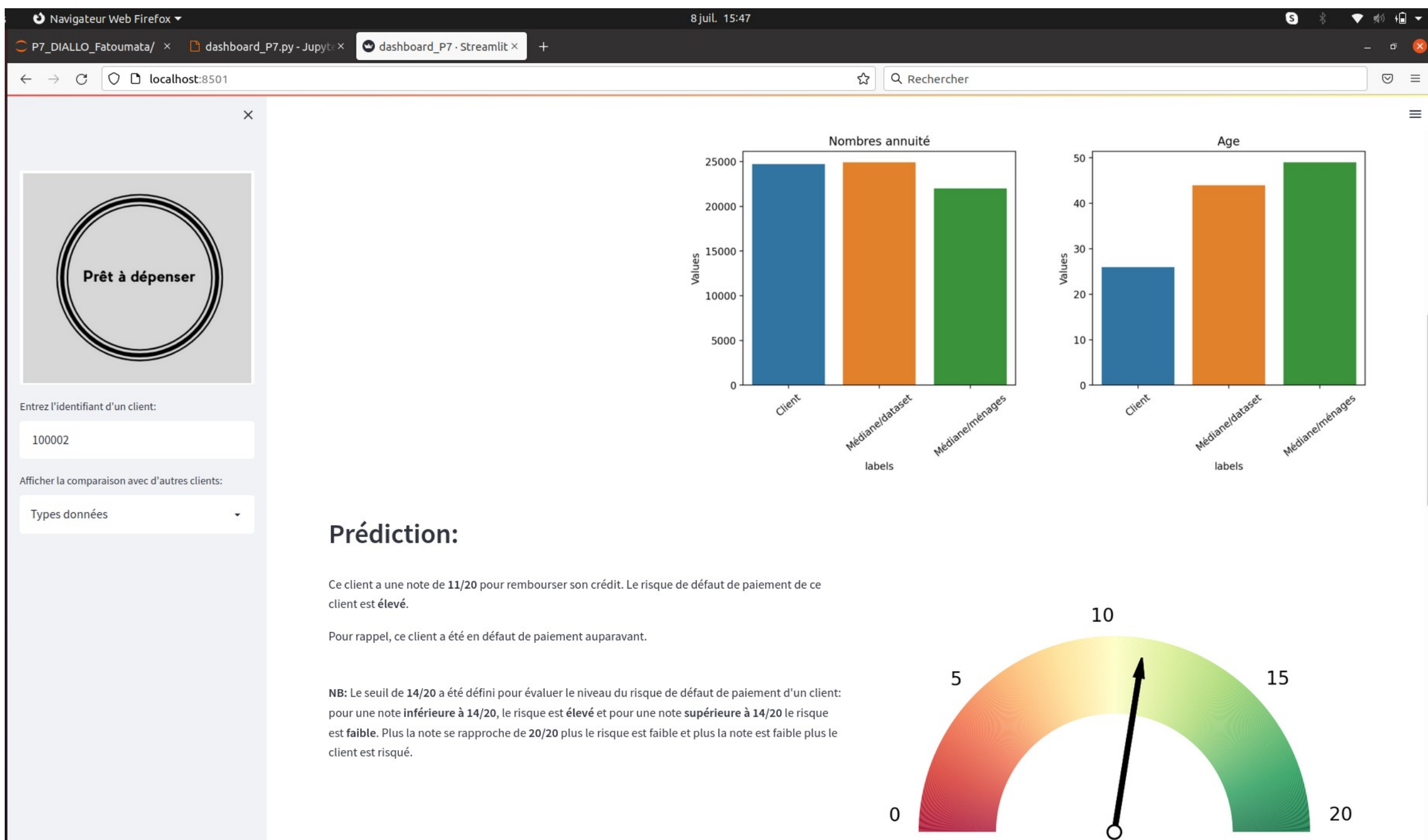
Entrez l'identifiant d'un client:

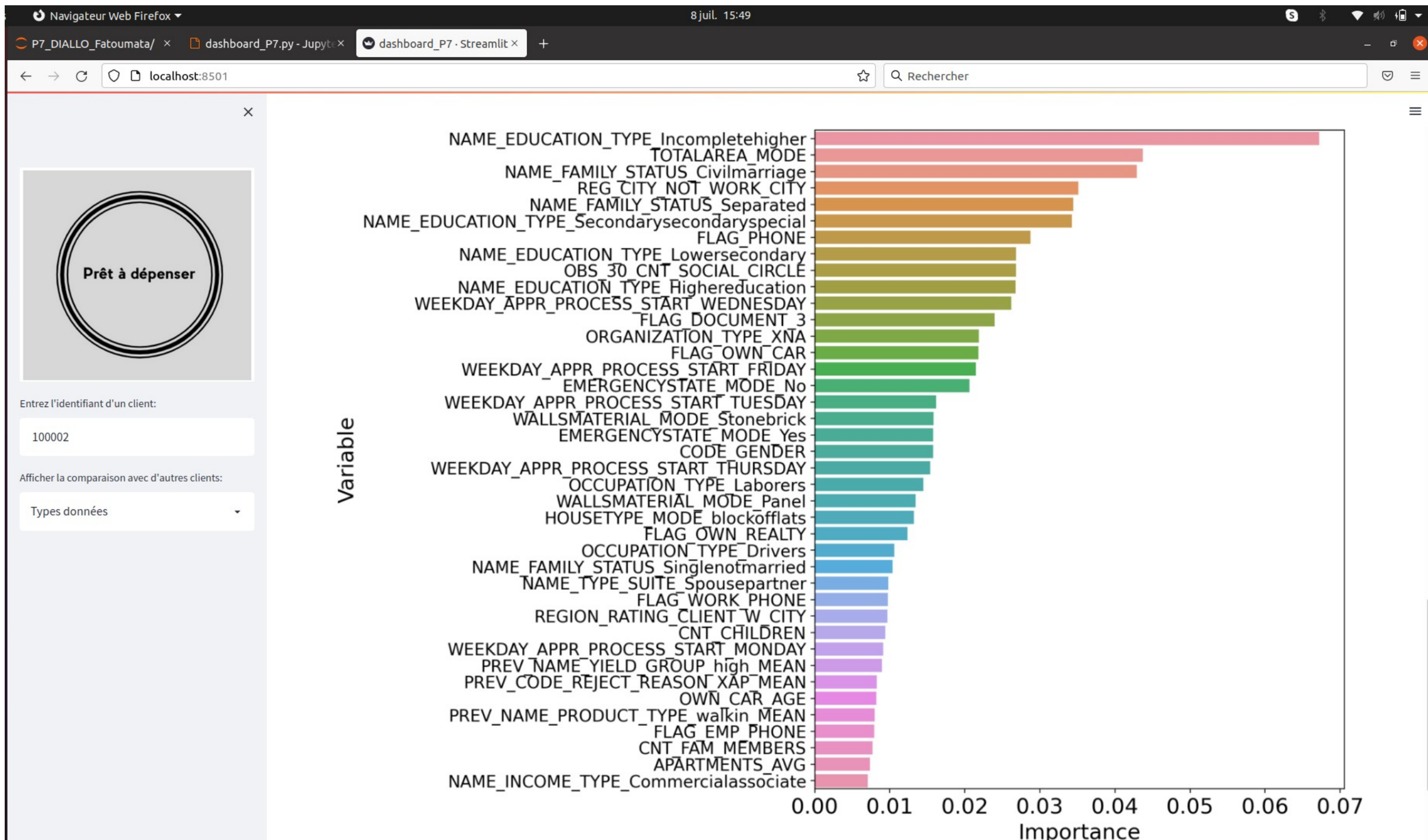
Bienvenue sur votre portail de scoring client

S'il vous plait entrez un identifiant correct.

Made with Streamlit







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
- d'utiliser Flask et Streamlit pour créer une web application (dashboard)



MERCI POUR VOTRE ATTENTION