

SOUTENANCE DE MÉMOIRE DE FIN DE PROJET P7

Parcours: DATA SCIENTIST

Landry Didier GAMASSA
OPENCLASSROOMS

Soutenance du 20 janvier 2023

Sommaire

1-PRESENTATION DU PROJET

- Thématique du projet
- Problématiques du projet

2-ETUDE DES DONNEES

- Présentation des données
- Présentation du Notebook Kaggle

3-MODELISATION

- Workflow Mlops appliqué au projet
- Aperçu des Techniques ML appliquées au projet
- Choix des métriques d'évaluation des modèles ML testés
- Visualisation Courbe ROC Modèle Baseline avec data d'entraînement équilibré et non équilibré
- Analyse des objectifs métier
- Visualisation Courbes ROC Modèle ML Appli Web Scoring Client et modèle Objectifs métiers
- Selection des Features /Features importants

4-INTERPRETATION GLOBALE ET LOCALE DU MODELE AVEC SHAP

- Interprétation globale du modèle shap
- Interprétation locale du modèle avec shap

5-DASHBOARD

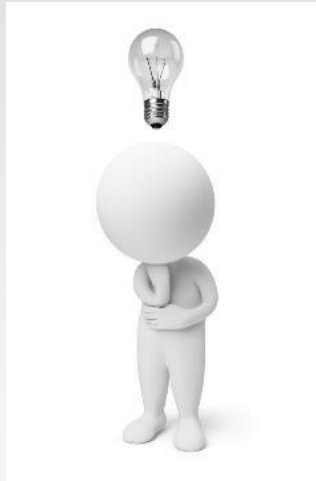
- Architecture Client-Serveur/Protocole HTTP
- Déploiement en local
- Déploiement sur le Cloud
- Workflow de déploiement sur le Cloud

6-BILAN DU PROJET

- Rappel sur les objectifs du projet
- Bilan Technique
- Bilan méthodologique
- Améliorations à réaliser
- Leçons apprises

I. PRESENTATION DU PROJET

I.1 Thématique du projet



👉 **Implémentez un modèle de scoring**

I.2_PROBLEMATIQUES DU PROJET

Etude d'un modèle de Scoring

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

Développement d'un dashboard interactif

Développement d'un Dashboard interactif

pour que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit.

Le dashboard doit permettre de:

1. Visualiser le score pour chaque client
2. Visualiser des informations descriptives relatives à un client
3. Comparer les informations descriptives relatives à un client à l'ensemble des clients ou à un groupe de clients similaires.

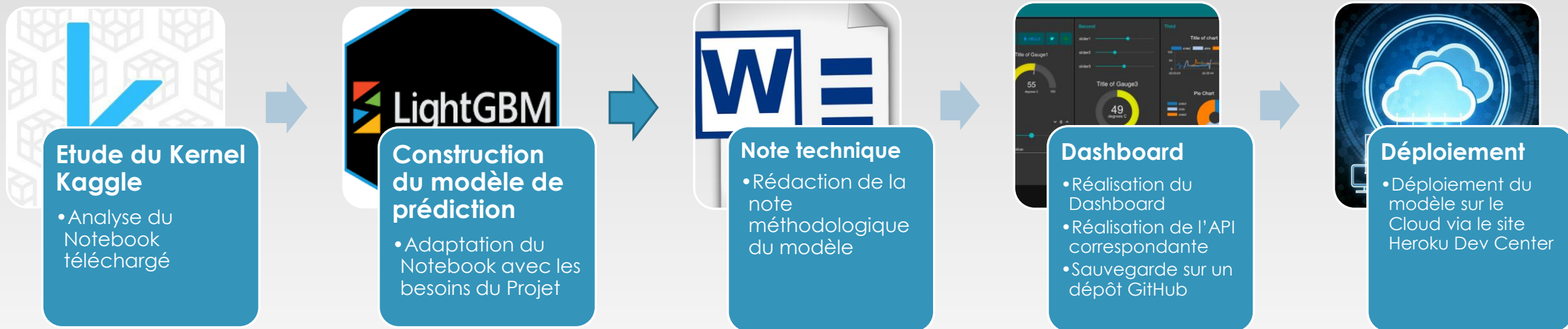
Demandes et suggestion du manager

- Partir d'un kernel Kaggle pour faciliter l'étude et la préparation des données.
- Réaliser une note méthodologique expliquant en détails la construction du modèle.
- Déploiement du dashboard sur le cloud

Lien données:

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

I.3_PLAN D'ACTIONS



II. ETUDE DES DONNEES

II.1_Présentation du jeux des données

Application train:

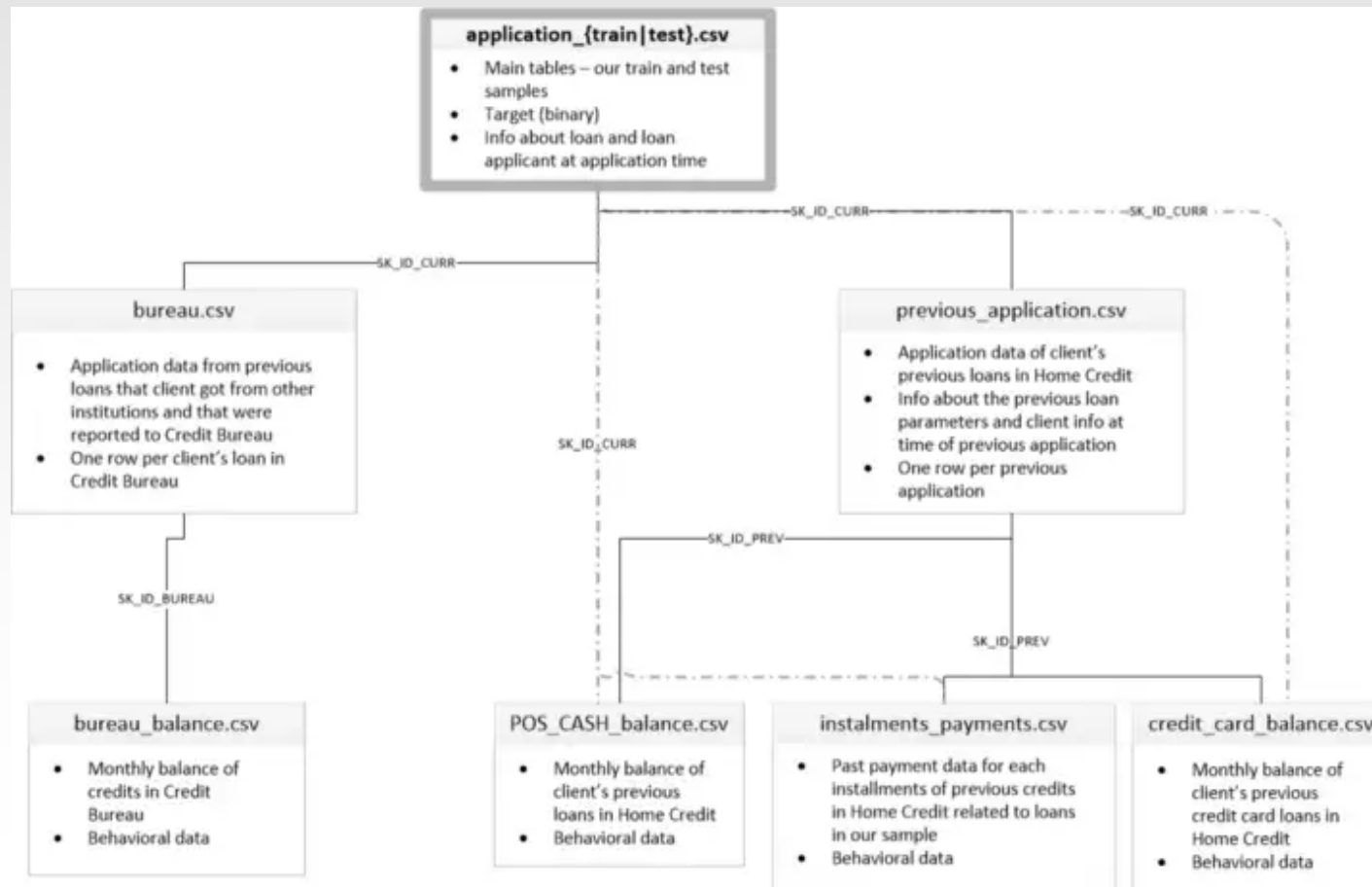
307511 lignes
122 colonnes

Il existe 8 sources de données différentes fournies avec une table principale divisée en deux fichiers pour **Train (avec TARGET)** et **Test (sans TARGET)**. :

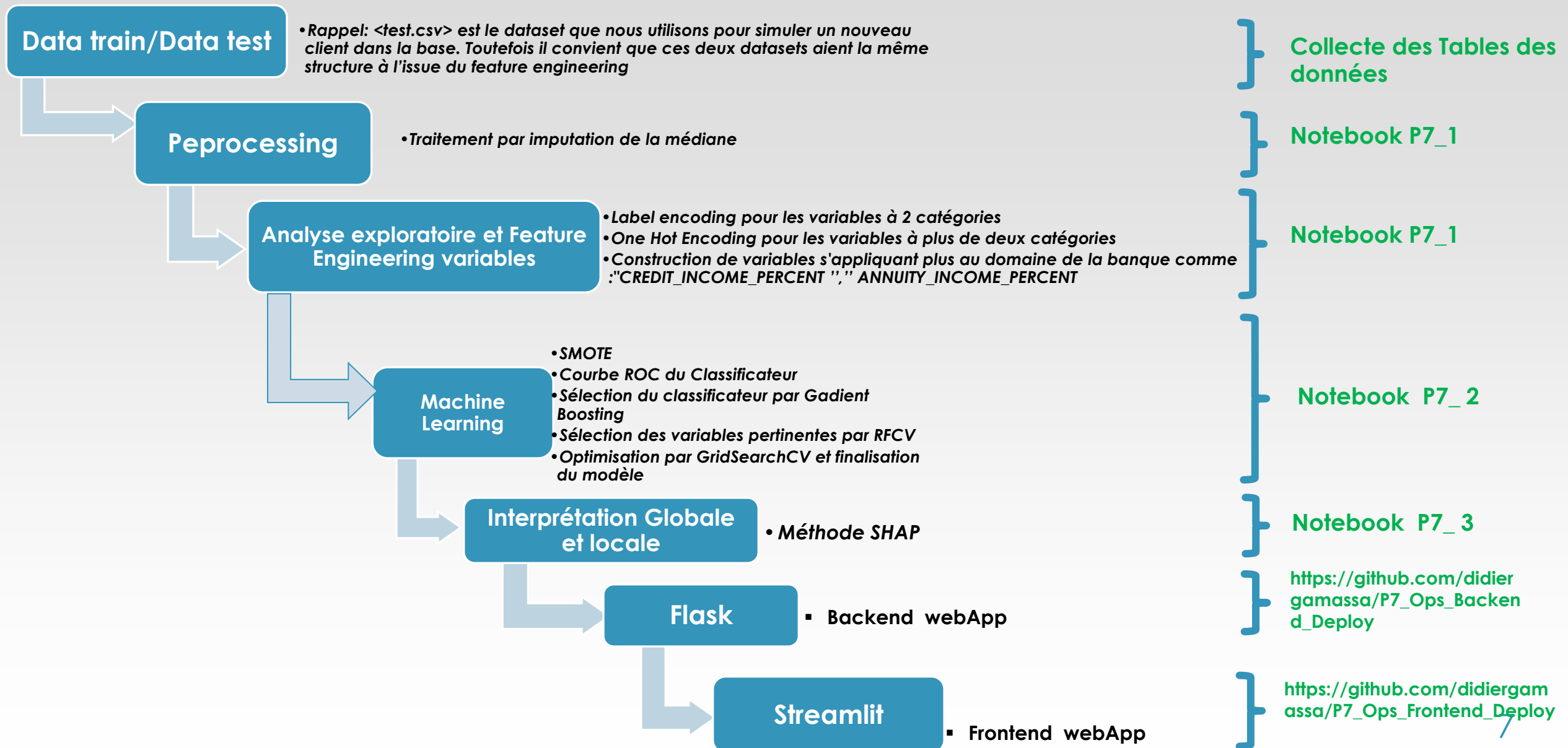
Application test

48744 lignes
121 colonnes

Les données sont fournies par Home Crédit, un service dédié aux lignes de crédit (prêts) fournies à la population non bancarisée. Prédire si un client remboursera ou non un prêt ou aura des difficultés est un besoin commercial essentiel, pour toute banque..

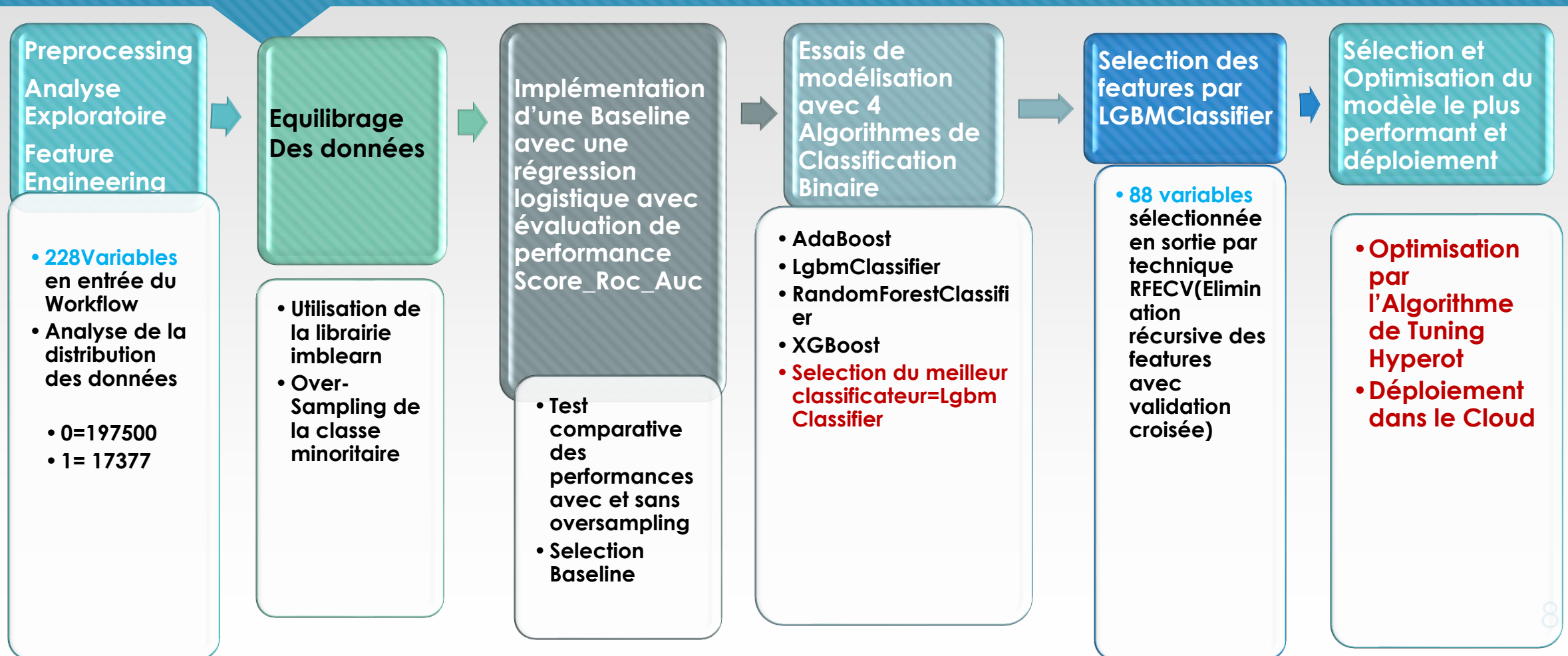


II.2 Présentation du Notebook Kaggle



III MODELISATION

III.1_Workflow MOps (Notebook 2)



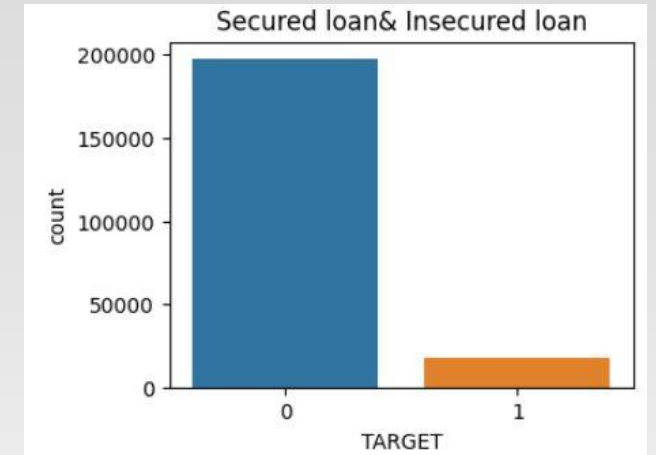
III.2 Aperçu sur les techniques ML appliquées à la modélisation

1_Analyse de la distribution des données: `X_train, X_val, y_train, y_val = train_test_split(data_train, TARGET, test_size=0.3, random_state=42, stratify=y)`

Détection d'une classification binaire:



TARGET	
0	197880
1	17377

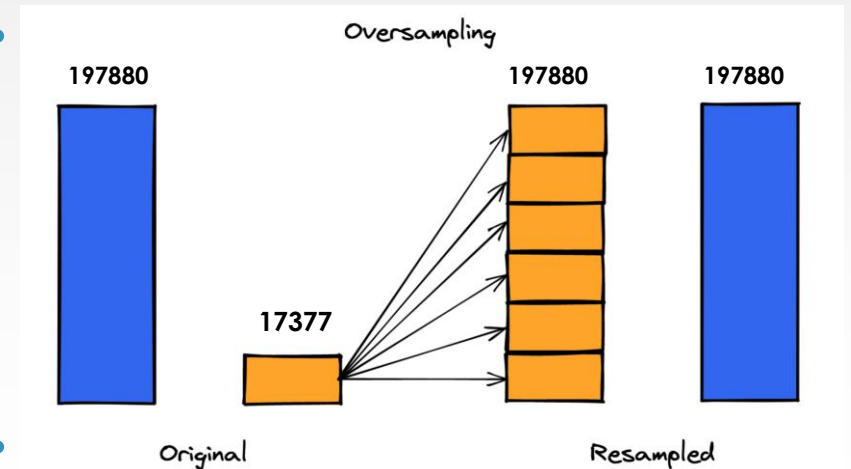


2_L'algorithme Smote :Nécessaire pour équilibrer les classes

Before sampling class distribution:- `Counter({0: 197880, 1: 17377})`

After sampling class distribution:- `Counter({0: 197880, 1: 197880})`

Smote



III.3_Aperçu sur les techniques ML appliquées à la modélisation

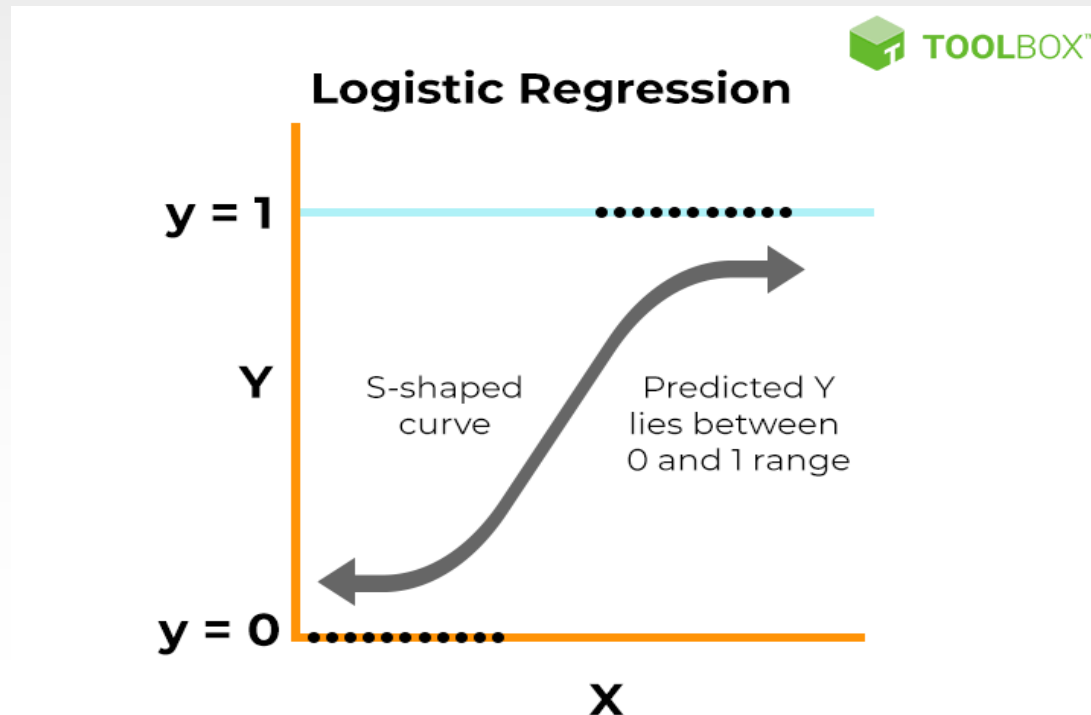
3_Selection d'un algorithme de référence à l'exploration du meilleur algorithme de classification à sélectionner :

3.1 Usage général:

✓ **DummyClassifier** → Ce classificateur sert de référence de base à comparer avec d'autres classificateurs plus complexes.

✓ 3.2 Usage particulier de type Classification Binaire: from sklearn.linear_model import Logistic_Regression

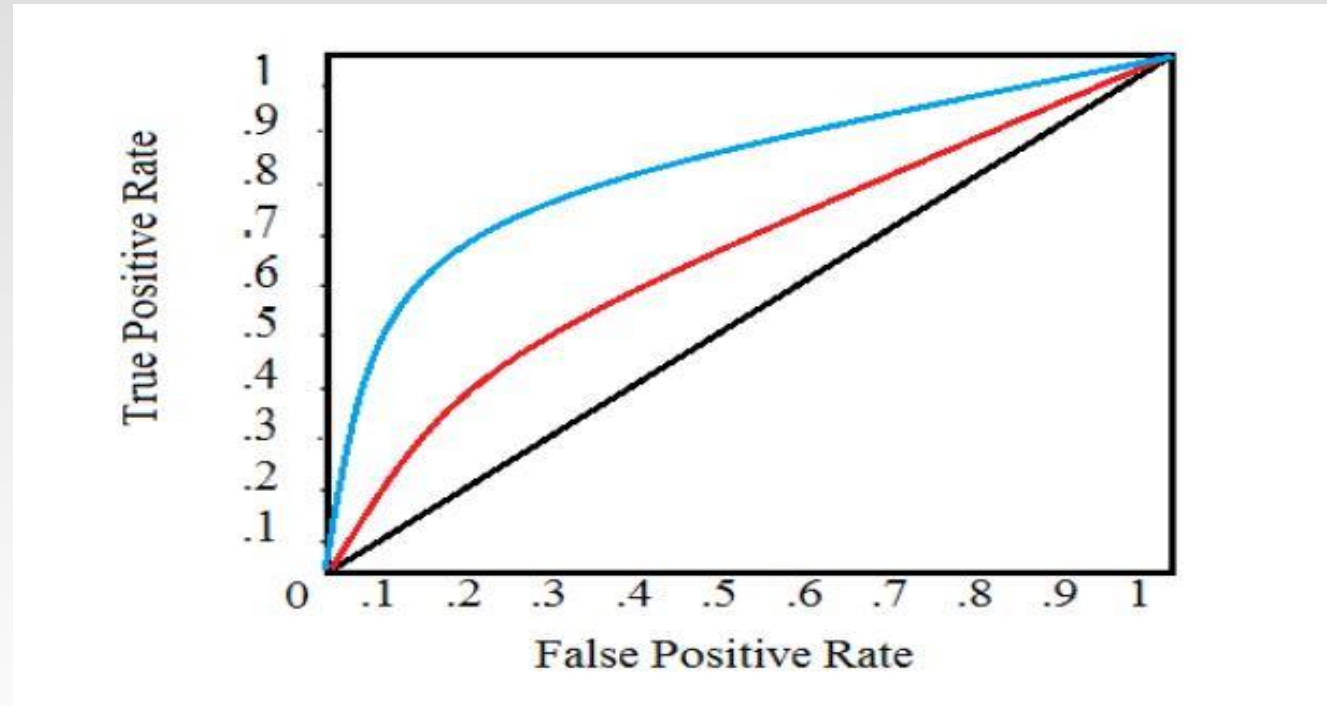
Logistic_Regression → Ce classificateur est un algorithme de classification ,utilisé lorsque les données de la dataset ont une sortie binaire : 0 ou 1



III.4_Aperçu sur les techniques ML appliquées à la modélisation

4_Choix d'une méthode d'évaluation qualité du modèle: Courbe ROC du modèle

La courbe ROC (Receiver Operating Characteristic) représente le taux de vrais positifs par rapport au taux de faux positifs : $TPR=F(FPR)$



L'Area Under the Curve (AUC) 'est l'aire sous la courbe ROC. (Il s'agit de l'intégrale de la courbe.) Cette métrique est comprise entre 0 et 1.

Lorsque nous mesurons un classificateur selon le ROC AUC, nous ne générons pas de prédictions 0 ou 1, mais plutôt une probabilité entre 0 et 1.

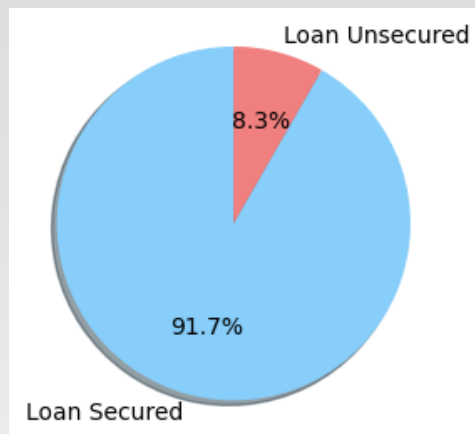
Dans le cas de notre étude, nous utilisons le ROC AUC ou le score F1 pour refléter plus précisément les performances d'un classificateur.

III.5_Métriques d'évaluation en Classification Binaire

Notebook2:

Projet =Classification binaire:

L'objectif de notre étude est de répondre par une classification binaire à la problématique présentée par la société Prêt à Dépenser. Il conviendra de répartir les individus d'un ensemble dans deux groupes disjoints selon que l'élément est solvable ou insolvable.



Matrice de confusion:

En apprentissage automatique supervisé, la matrice de confusion est une matrice qui mesure la qualité d'un système de classification. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TN + FP + FP + FN)}$

Les métriques d'évaluation des Modèle de Classification binaire sont judicieusement sélectionnés:

$$\text{Recall} = \text{Sensitivity} = \text{TPR} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}} = \frac{TP}{TP + FN}$$

Il permet de savoir le **pourcentage de positifs bien prédit** par notre modèle.

Plus il est **élevé**, plus le modèle de Machine Learning **maximise le nombre de Vrai Positif**.

Mais attention, cela ne veut pas dire que le modèle ne se trompe pas.

Quand le recall est **haut**, cela veut plutôt dire qu'il **ne ratera aucun positif**. Néanmoins cela ne donne aucune information sur sa qualité de prédiction sur les négatifs

$$\text{Precision} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}} = \frac{TP}{TP + FP}$$

Il permet de connaître le **nombre de prédictions positifs bien effectuées**. En d'autres termes c'est le **nombre de positifs bien prédit** (Vrai Positif) **divisé par l'ensemble des positifs prédit** (Vrai Positif + Faux Positif).

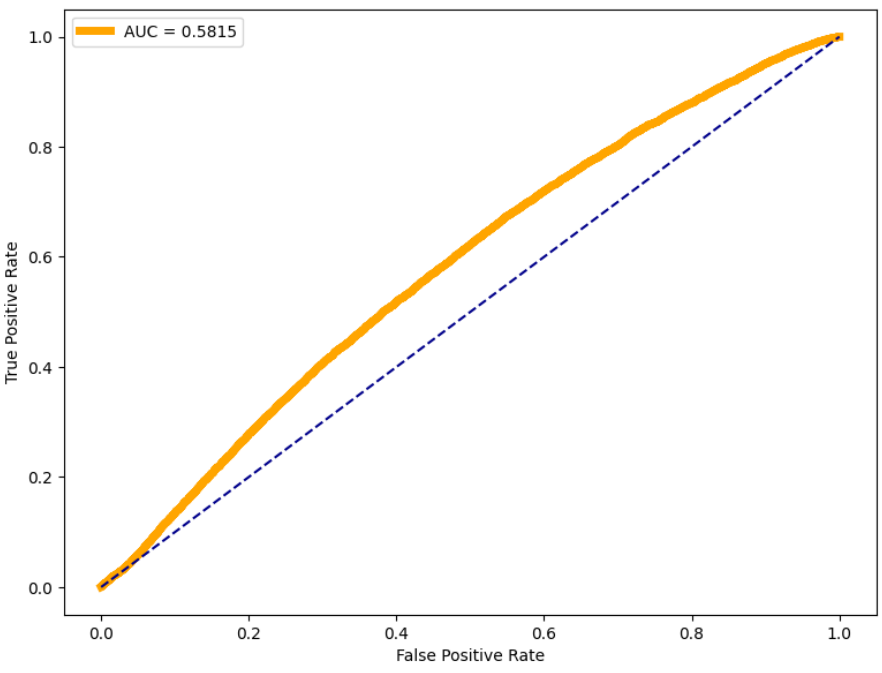
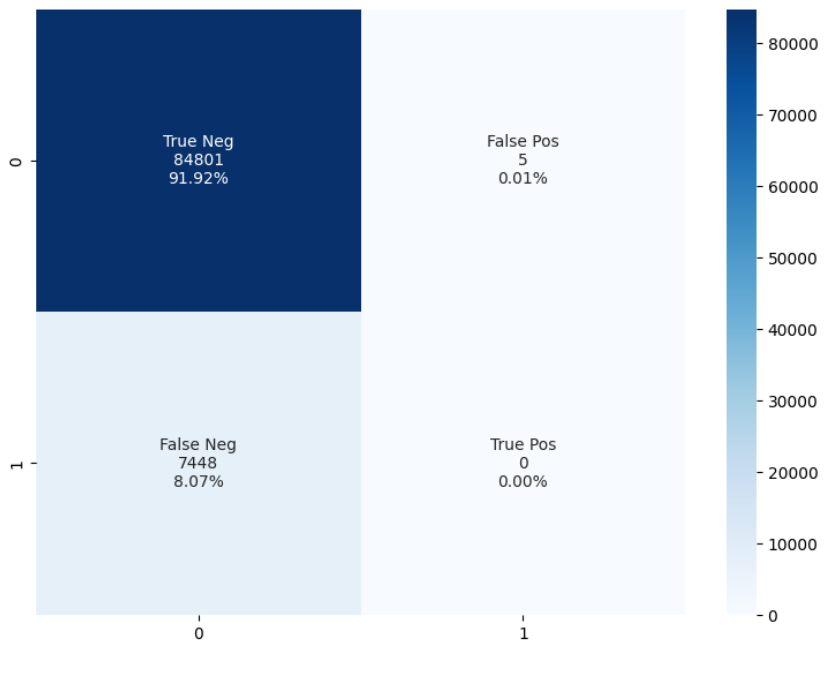
Plus elle est **élevé**, plus le modèle de Machine Learning **minimise le nombre de Faux Positif**.

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

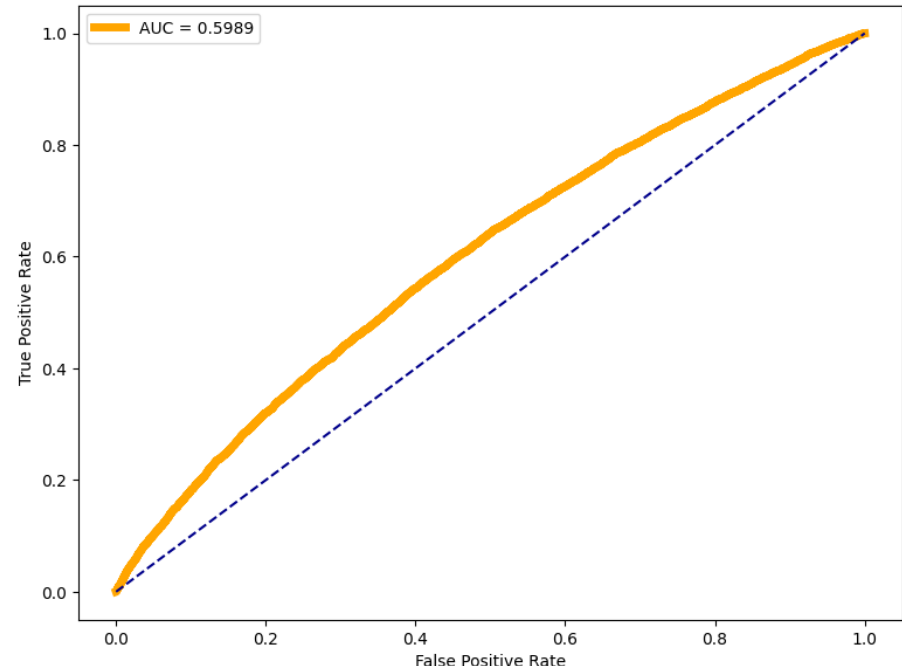
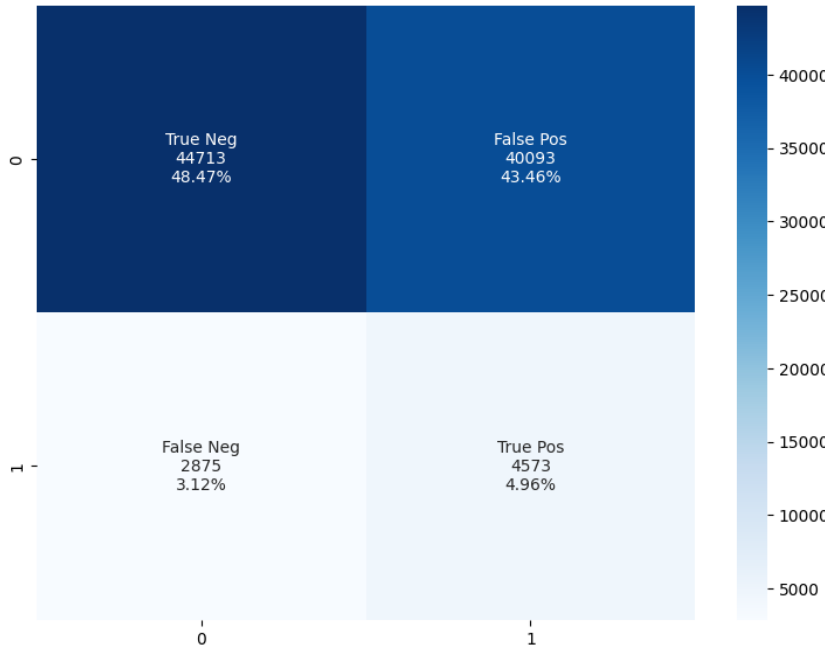
La mesure F peut être considérée comme un compromis entre rappel et précision. Il n'est élevé que lorsque le rappel et la précision sont élevés. Cela équivaut à rappel quand $\alpha = 0$ et précision quand $\alpha = 1$. La F-mesure suppose des valeurs dans l'intervalle [0,1].

III.6_Analyse et visualisation Courbe Roc modèles baseline

Baseline model => Logistic
Régression=0.5815
Mesure de qualité inférieure à
l'oversampling mais un Recall de
bonne qualité. Donc Modèle de
classification comme modèle
Baseline à notre projet ML.



Oversampling et mesure de
performance=0.5989
Mesure de bonne qualité mais un
Recall de mauvais qualité. Pas
exploitable comme Baseline



III.7_Analyse des objectifs Business

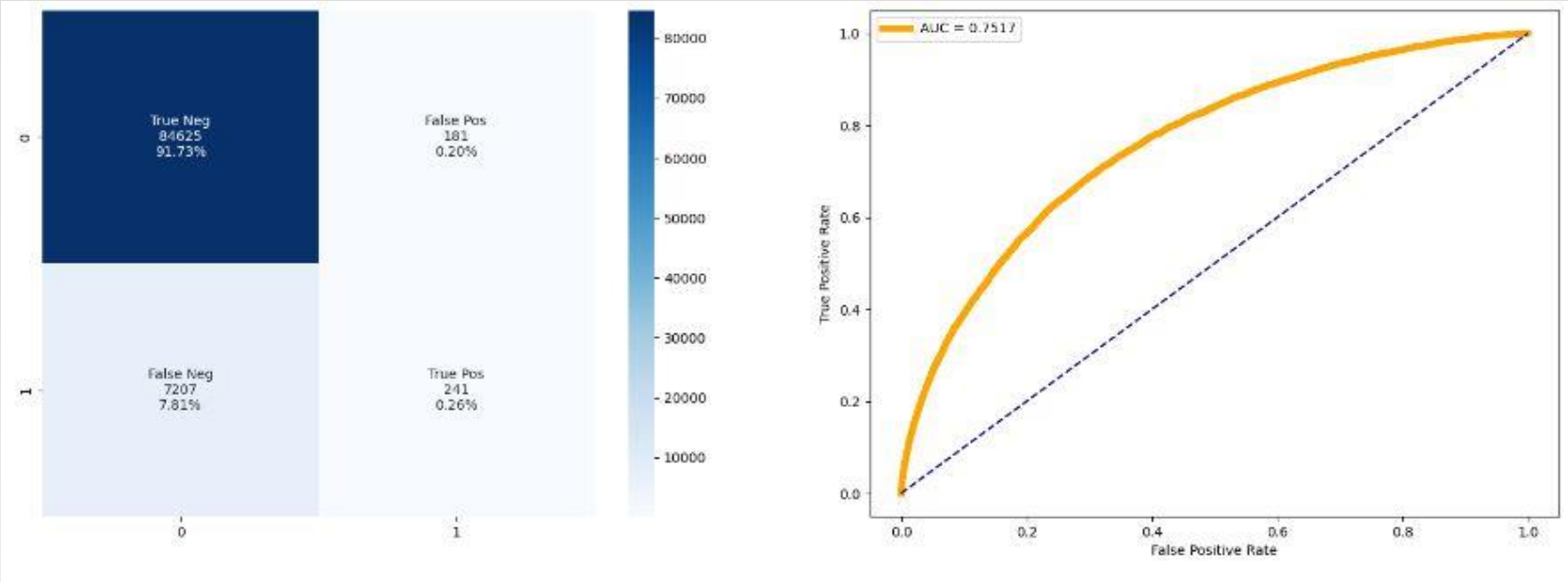
En toute logique bancaire, le Recall sera plus le paramètre à plus pondérer que la précision car on préférera limiter un risque de perte financière plutôt qu'un risque de perte de client potentiel. Dans le cas de notre étude et sans objectifs métier connus, nous avons sélectionné un coefficient Beta de manière empirique afin de corriger au mieux l'erreur 2

```
100%|████████████████████████████████████████████████████████████████████████████████| 10/10 [10:52<00:00, 65.27s/trial, best loss: 0.970206898815728]
beta= 2
False Pos: 1.00%
False Neg: 7.40%
Ratio_False:0.14
[[83883  923]
 [ 6823  625]]
{'AUC': 0.7435678636649774, 'Precision': 0.4037467700258398, 'Recall': 0.08391514500537058, 'F1': 0.13895064473099156}
100%|████████████████████████████████████████████████████████████████████████████████| 10/10 [13:57<00:00, 83.79s/trial, best loss: 0.9770960238954955]
beta= 3
False Pos: 0.28%
False Neg: 7.77%
Ratio_False:0.036
[[84550  256]
 [ 7167  281]]
{'AUC': 0.7585227611674417, 'Precision': 0.5232774674115456, 'Recall': 0.03772824919441461, 'F1': 0.07038196618659986}
100%|████████████████████████████████████████████████████████████████████████████████| 10/10 [12:21<00:00, 74.11s/trial, best loss: 0.980079570126031]
beta= 4
False Pos: 0.75%
False Neg: 7.52%
Ratio_False:0.1
[[84114  692]
 [ 6940  508]]
{'AUC': 0.7434012041458976, 'Precision': 0.42333333333333334, 'Recall': 0.0682062298603652, 'F1': 0.11748381128584645}
100%|████████████████████████████████████████████████████████████████████████████████| 10/10 [11:59<00:00, 71.91s/trial, best loss: 0.9831813091291843]
beta= 5
False Pos: 0.67%
False Neg: 7.51%
Ratio_False:0.09
[[84185  621]
 [ 6927  521]]
{'AUC': 0.7499570733157244, 'Precision': 0.45621716287215414, 'Recall': 0.0699516648764769, 'F1': 0.12130384167636785}
100%|████████████████████████████████████████████████████████████████████████████████| 10/10 [12:24<00:00, 74.49s/trial, best loss: 0.9854066577664308]
beta= 6
False Pos: 0.30%
False Neg: 7.76%
Ratio_False:0.039
[[84527  279]
 [ 7163  285]]
{'AUC': 0.7577070876721149, 'Precision': 0.5053191489361702, 'Recall': 0.03826530612244898, 'F1': 0.07114328507239141}
```

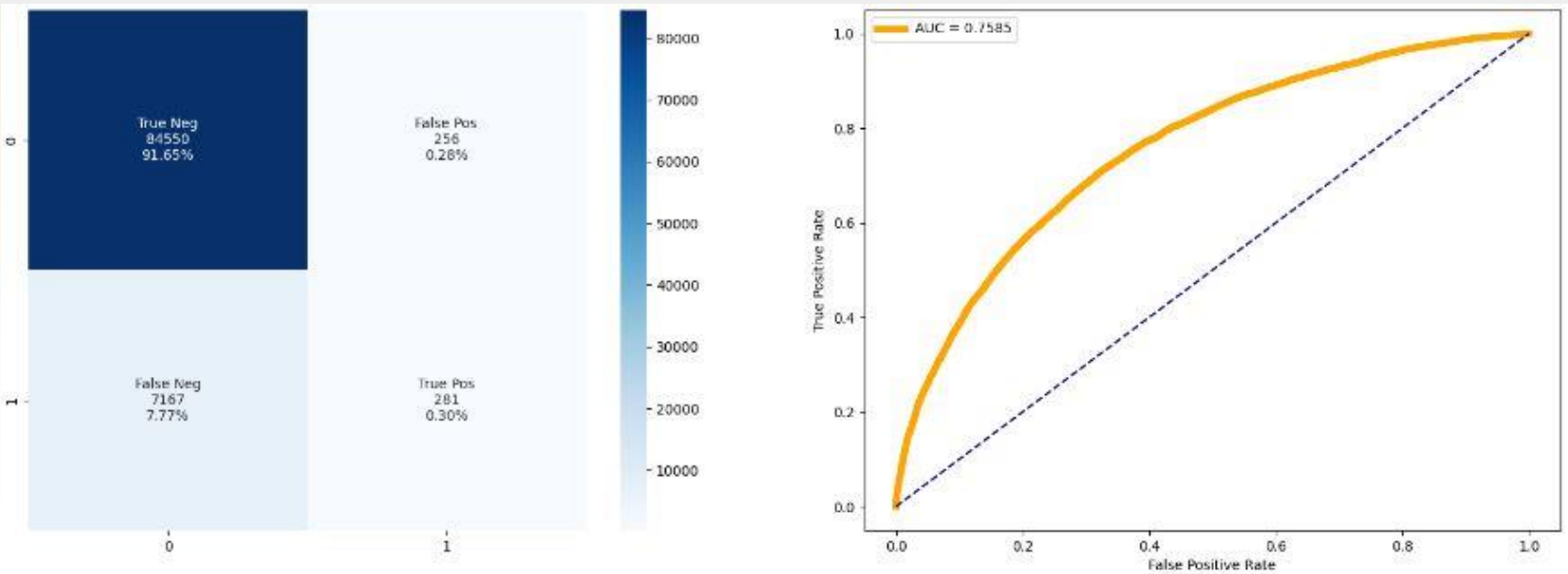
$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

III.8 Courbes Roc du modèle classification Cloud et modèle classification objectifs métiers

Courbe Roc et Matrice de confusion du Modèle de Classification Cloud avec des belles performance [AUC=0.7517](#)

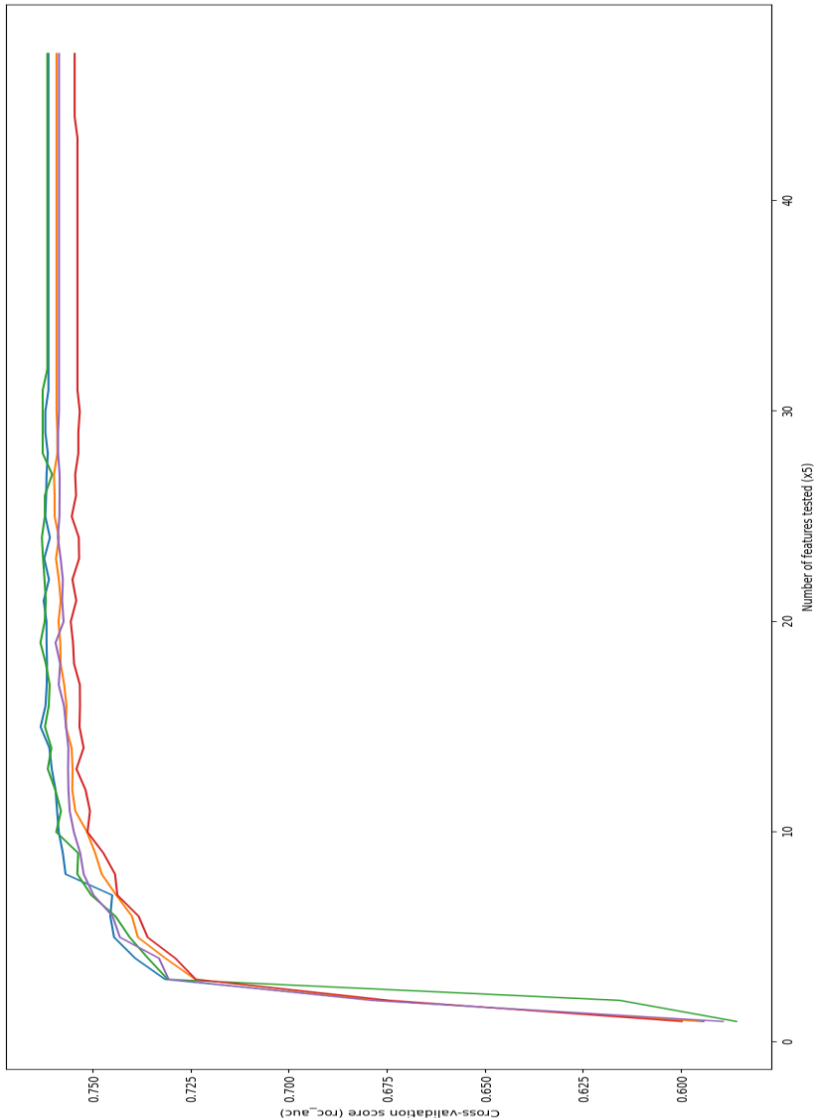


Courbe Roc et Matrice de confusion du modèle de Classification avec objectif métier . Il présente un bon score Roc Auc mais avec un mauvais Recall. Les objectifs métiers pourraient apporter une meilleure optimisation. [AUC=0.7517](#)

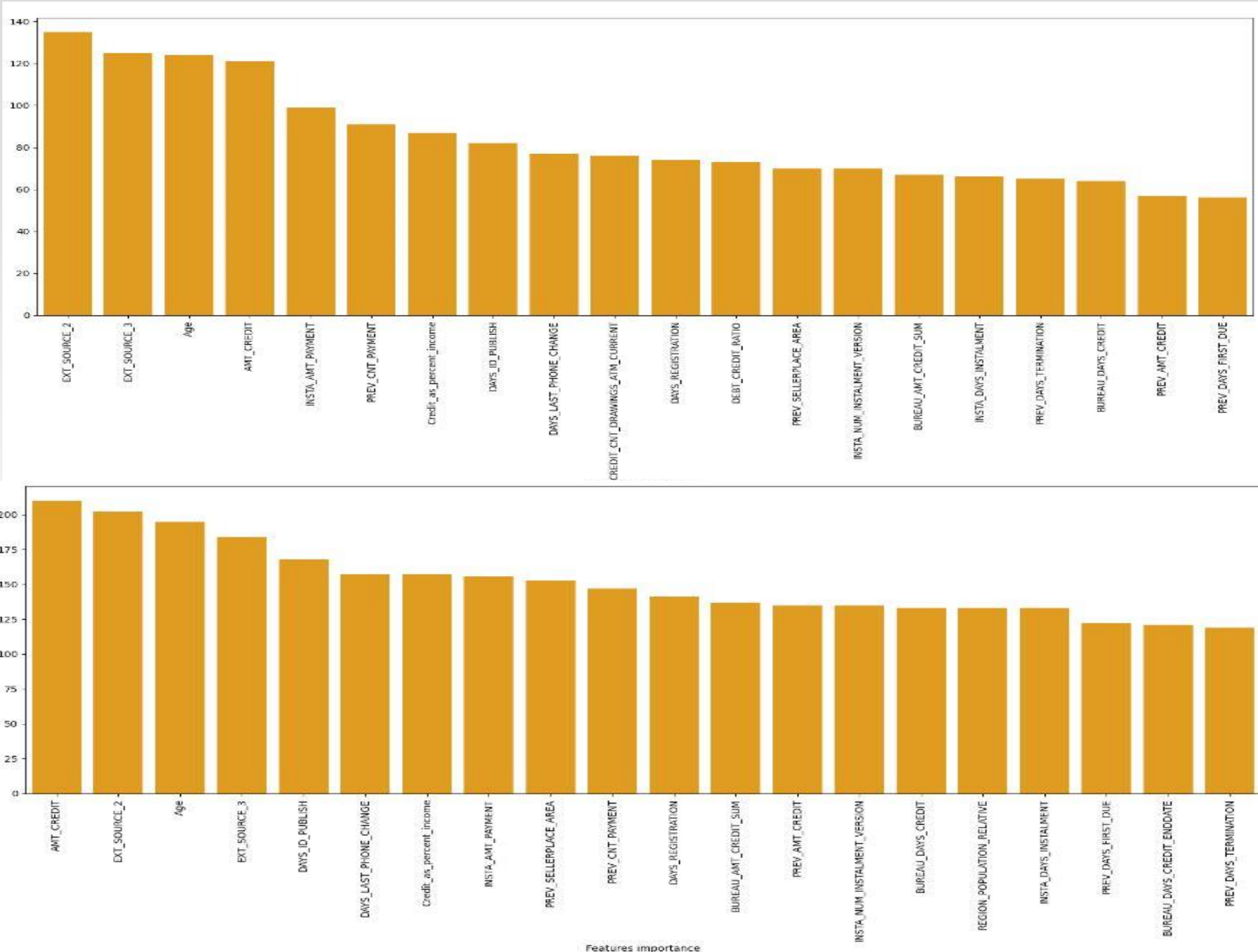


III.9_Selection des features/Features importants

Selection des Features :C'est le processus qui permet de sélectionner automatiquement ou manuellement les features qui contribuent à une variable cible. Les algorithmes à arbre de décision sont adaptés et la rfc est l'une des techniques couramment utilise en ML.



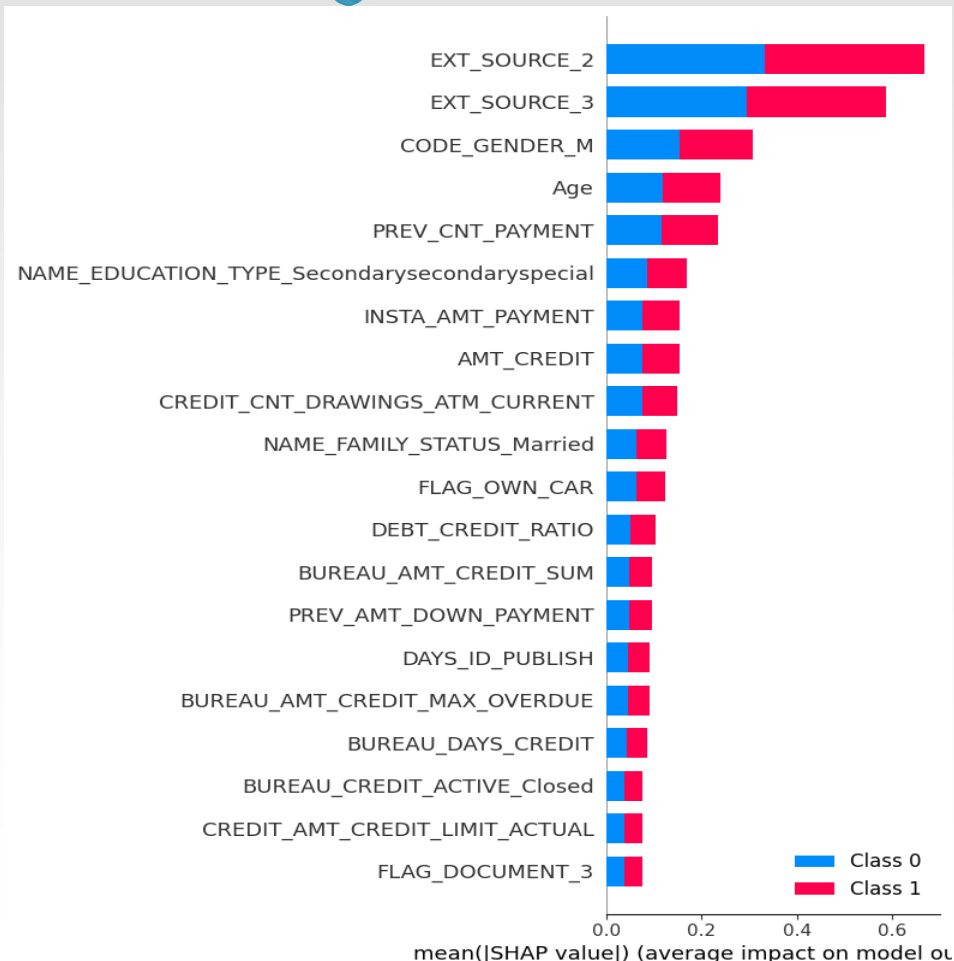
Importance des features :Il fait référence à des techniques qui attribuent un score aux caractéristiques d'entrée en fonction de leur utilité pour prédire les variables cibles.



IV. INTERPRETABILITE DU MODELE ML AVEC SHAP

IV.1 INTERPRETABILITE GLOBALE DU MODELE ML AVEC SHAP

Notebook3:.



$$\text{Output modèle} = \sum_{i=0}^N E_i$$

N=Nombre de features qui entraînent le modèle

Pour chaque variable, une distribution est tracée sur la façon dont il contribue aux résultats..

L'entité EXT_SOURCE_2 est la variable qui a le plus d'impact sur la sortie du modèle.

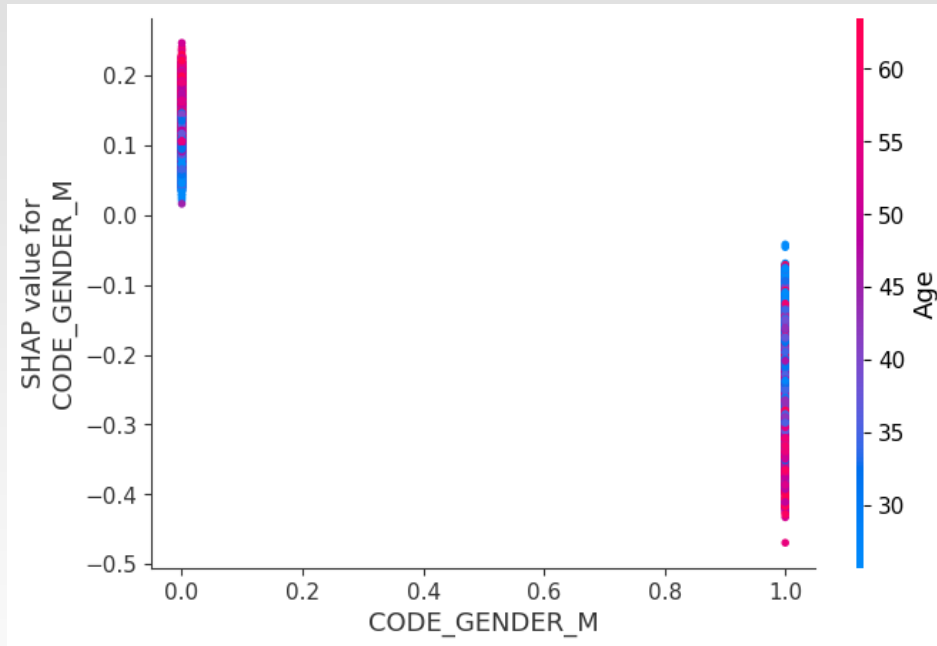
Plus il y a du rouge rouges et plus la probabilité d'obtenir un prêt est faible. on voit que beaucoup de clients sont dans ce cas.

Cas: Analyse locale de la dépendance de la feature (Genre)à la feature (Age)

Class 0
Class 1

Masculin

```
shap.dependence_plot("CODE_GENDER_M", shap_values[1], valid_x)
```

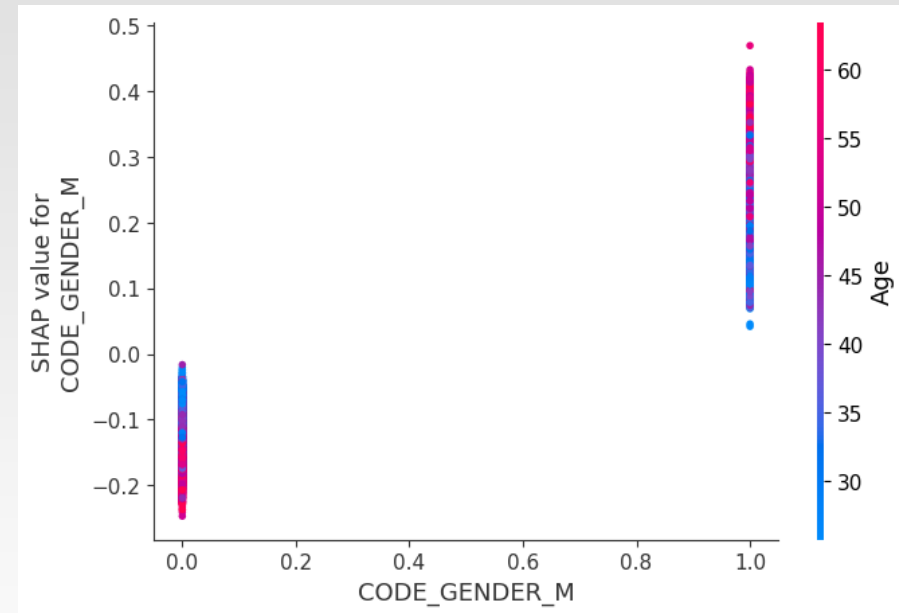


Interprétation du statut du prêt :

25 à 35ans : Prêt potentiellement non sécurisé
35 à 55ans: Prêt potentiellement sécurisé
55 à 62ans: Prêt potentiellement non sécurisé

Féminin

```
1 shap.dependence_plot("CODE_GENDER_M", shap_values[0], valid_x)
```



Interprétation du statut du prêt:

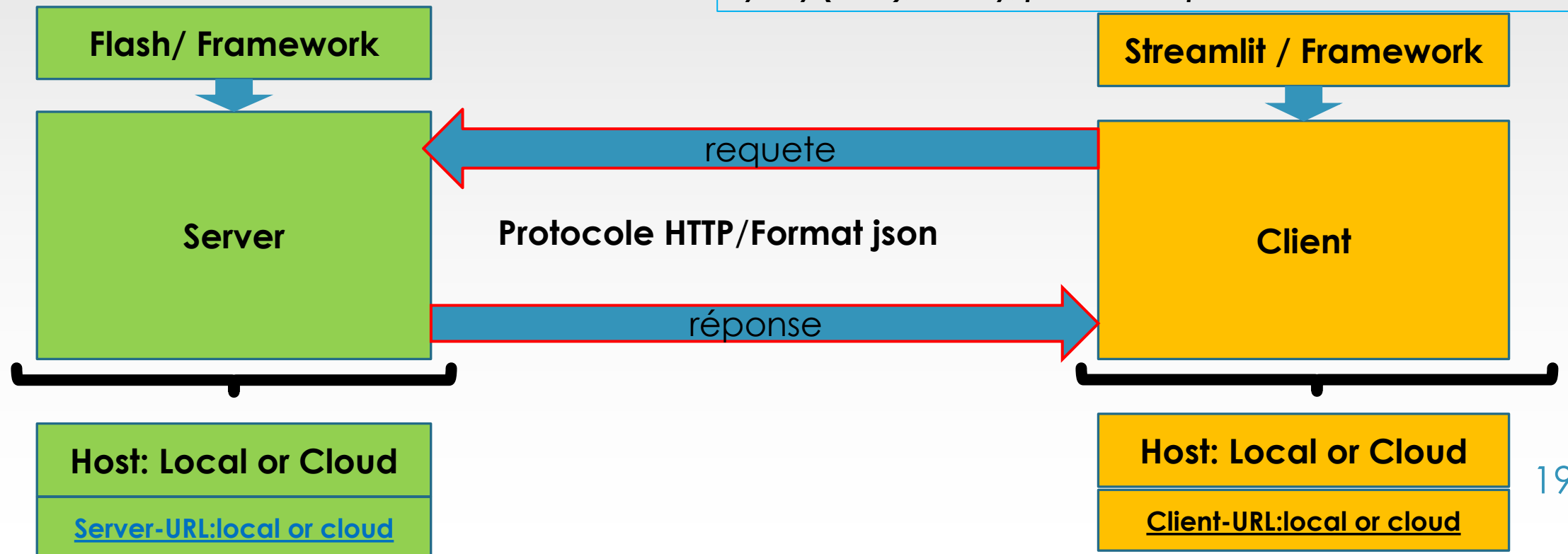
25 à 36 ans :Prêt potentiellement non sécurisé
36 à 48ans:Prêt potentiellement sécurisé
48 à 62ans:Prêt potentiellement non sécurisé

V. DASHBOARD

v.1 Architecture Client-Serveur/Protocole HTTP

Flask est un micro framework open-source de développement web en Python. Il est classé comme micro Framework car il est très léger. Flask a pour objectif de garder un noyau simple mais extensible.

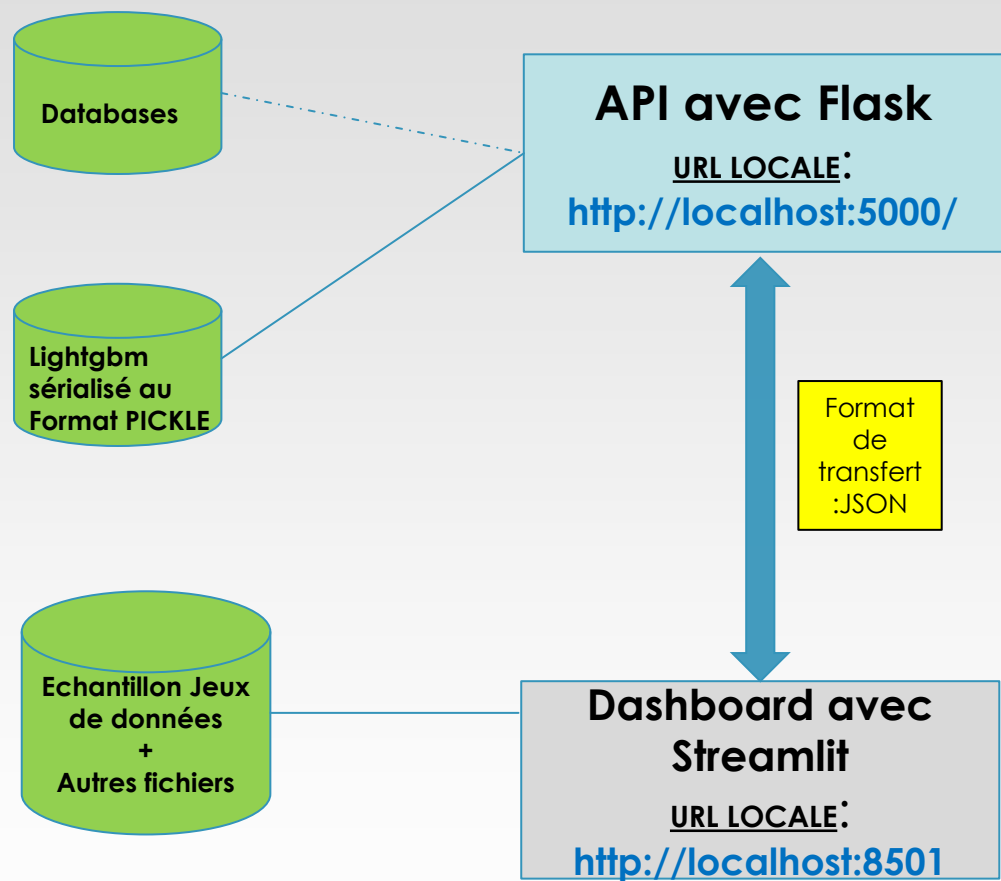
Streamlit est un framework d'application open source en langage Python. Cela nous aide à créer rapidement des applications Web pour la Data Science et le Machine Learning. Il est compatible avec les principales bibliothèques Python telles que scikit-learn, Keras, PyTorch, SymPy (latex), NumPy, pandas, Matplotlib, etc.



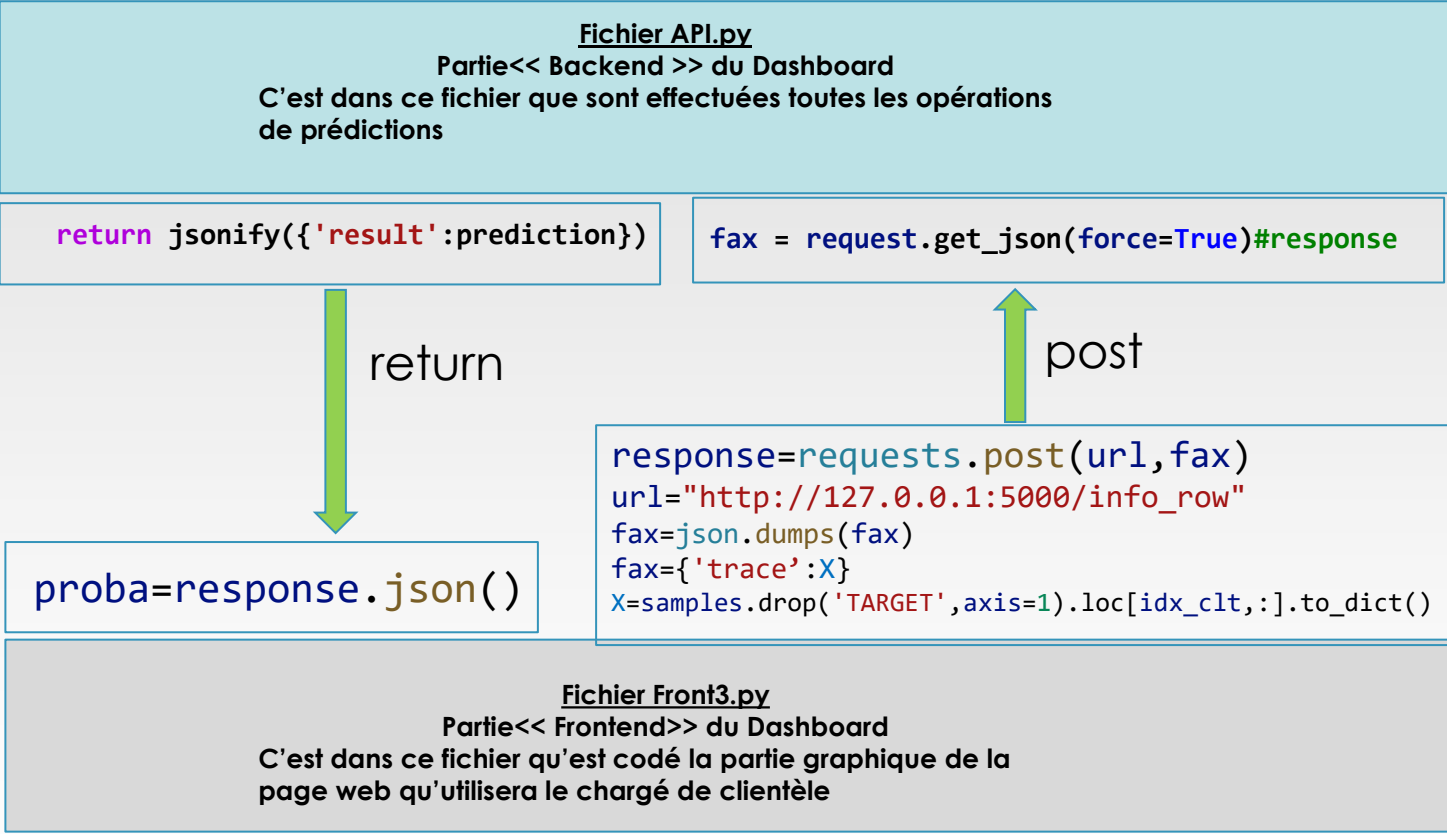
v.2_Déploiement sur un serveur local: ➡



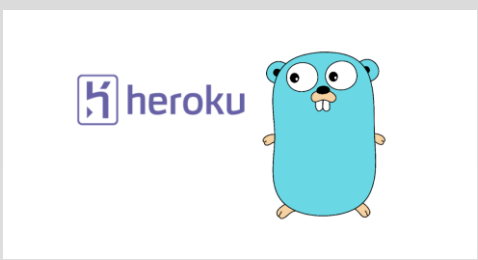
Etape1:Démarrer le serveur



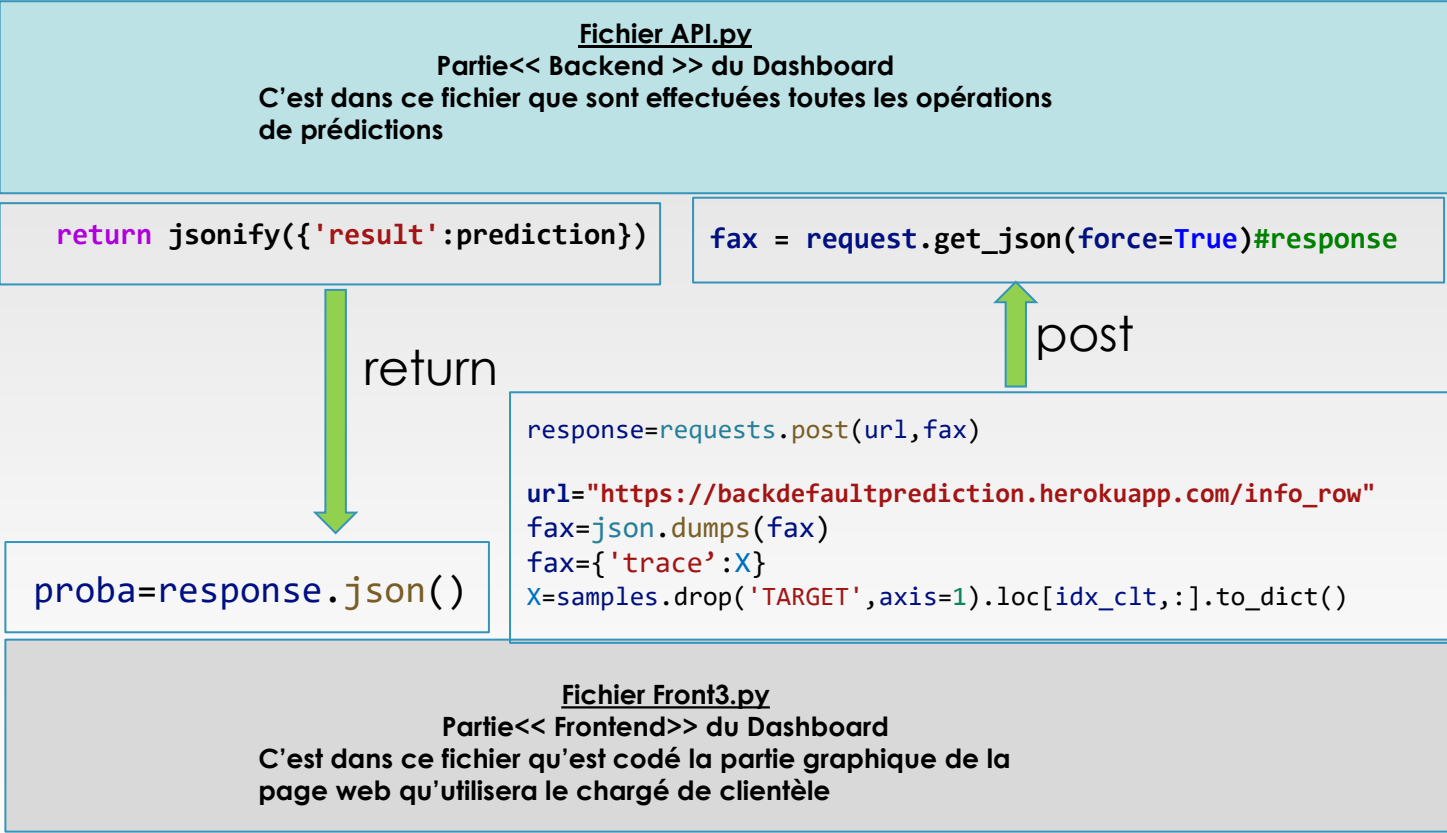
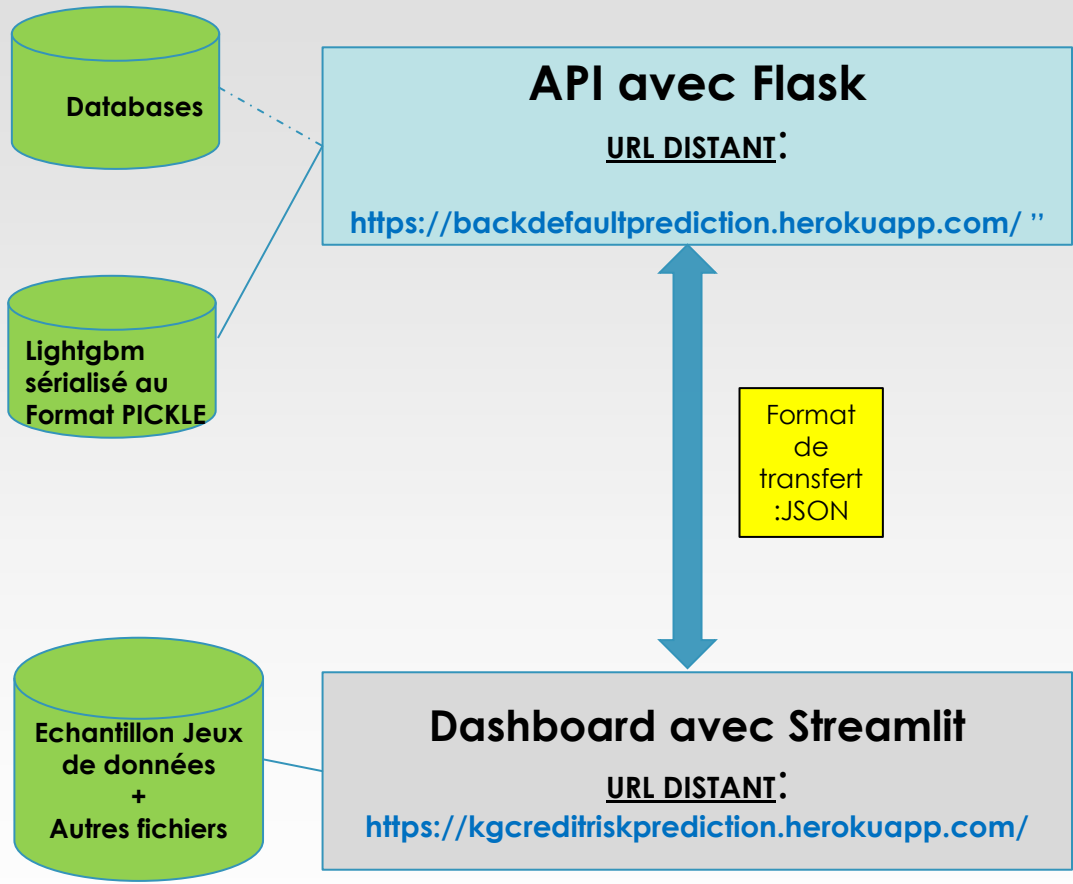
Etape2:Afficher le Dashboard



v.2_Déploiement sur un serveur d'application distant: ➡



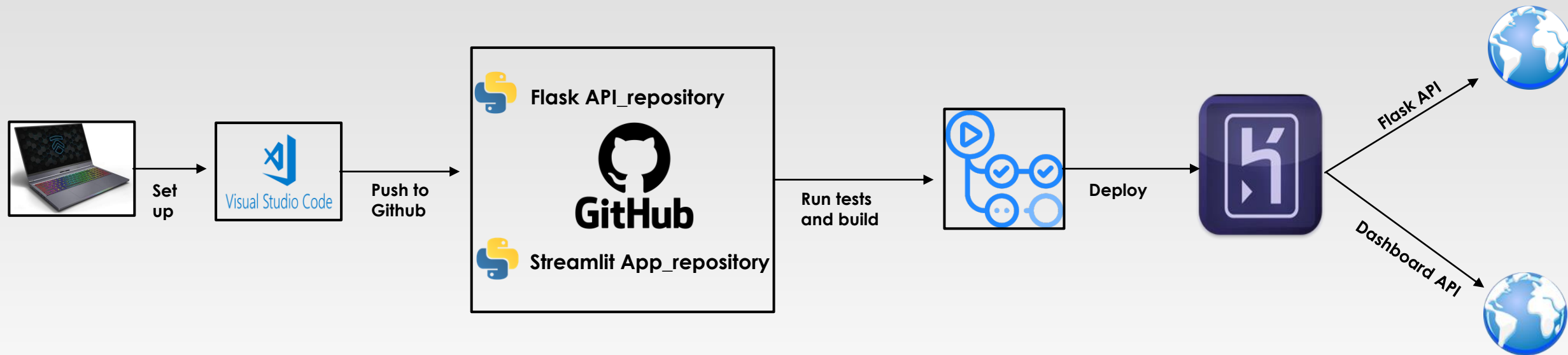
Etape1:Démarrer le serveur



Etape2:Afficher le Dashboard

v.4_Worflow de déploiement de l'Appli Scoring Client sur un serveur distant:

Déployer une application dans le cloud revient à construire un pipeline de déploiement continu d'un server d'application local vers un serveur d'application distant.
Pour notre projet ,nous avons appliqué le Worflow de déploiement suivant:



➤ Lien application Cloud:

<https://kgcreditriskprediction.herokuapp.com/>

➤ Lien site GitHub

https://github.com/didiergamassa/P7_Ops_Frontend_Deploy

https://github.com/didiergamassa/P7_Ops_Backend_Deploy

Lignes de commandes dans le dépôt du serveur d'application local :Ma machine

```
Git init
Git add
Git commit<<Name commit>>
Git remote<<Link Github repo
Git push_u origin master
```

Fichier.gitignore.txt pour
ignorer certains fichiers
csv

VI. BILAN DU PROJET

1. Rappel du projet:

Objectif principal : Construire une Web Application de Scoring Client avec deux Framework Python. Il s'agit de Flask en backend et Streamlit en frontend.

2. Bilan technique

Les choix techniques: Bon mais peuvent être optimisés

3. Bilan méthodologique

Les choix méthodologiques : Bon

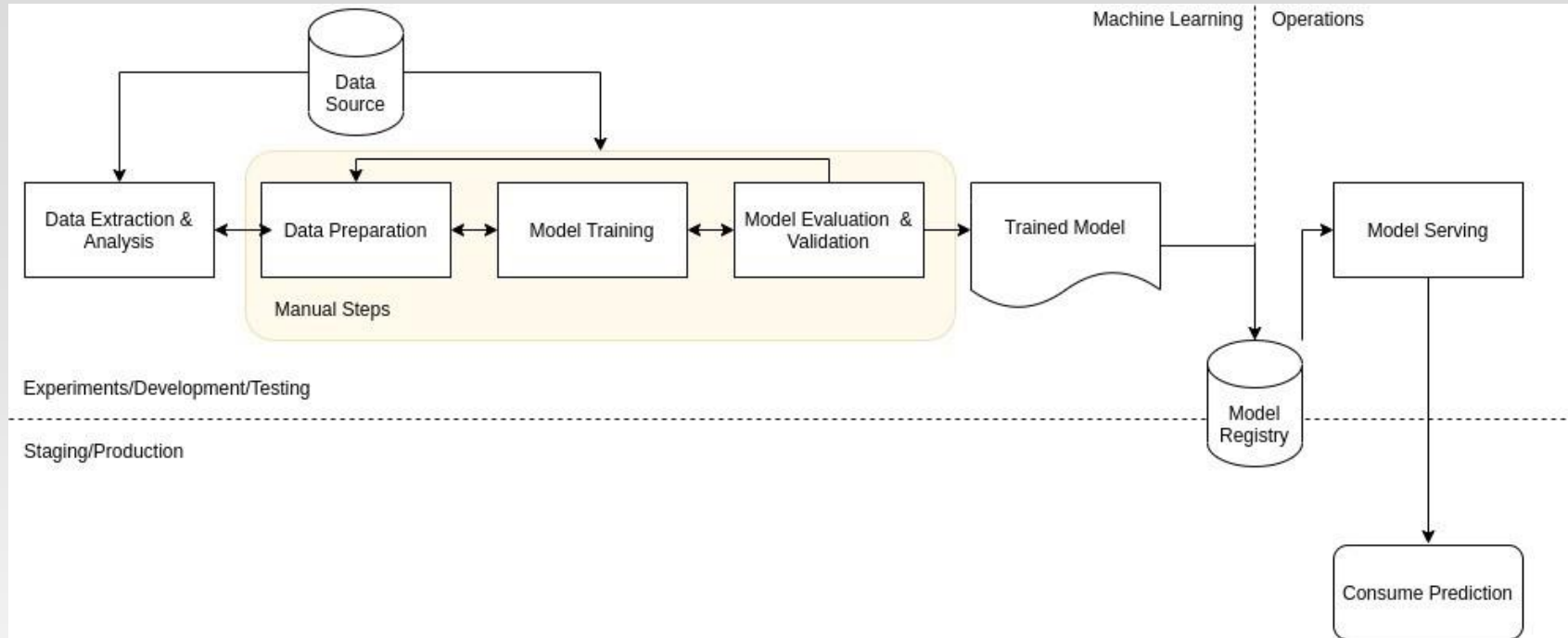
4. Améliorations:

4_1: Préciser les objectifs métiers afin de mieux cadrer le projet

4_: L'instabilité du Modèle de classification Objectifs métiers au réentraînement

4_2: Les choix techniques peuvent être optimisés en Deep Learning

5:Leçons apprises:




MLOPS: C'est un ensemble de pratiques qui vise à déployer et maintenir des modèles de machine Learning en production de manière fiable et efficace.

MLOps s'applique à l'ensemble du cycle de vie d'un modèle machine learning

Cycle de vie d'un modèle de machine Learning

Note Méthodologique et Notebook Python disponibles sur le lien ci-dessous:

 https://github.com/didiergamassa/P7_Ops_Frontend_Deploy/tree/main/.gitignore

FIN

Merci