IMPLEMENTER UN MODELE DE SCORING

KOFFI KONAN

PLAN DE PRÉSENTATION

I Problématique et présentation du jeu de données

- Problématique
- Jeu de données

II Approche de modélisation et interprétation

- Preprocessing et Features engineering
- Techniques de gestions de classes déséquilibrées
- Métrique d'évaluation adéquate
- Choix de techniques de gestion d'équilibre et de modèles
- Fonction coût et optimisation métier du modèle
- Interprétabilité du modèle choisi

III Présentation du Dashboard

- Mise en place d'une api via FastAPI et déploiement sur le cloud Heroku
- Mise en place d'un Dashboard via Streamlit et déploiement sur le cloud Heroku

Problématique

I PROBLÉMATIQUE ET PRÉSENTATION DES JEUX DE DONNÉES

Contexte

"Prêt à dépenser"
crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

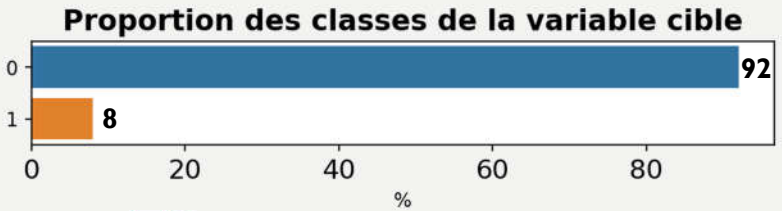
Mission

- ❑ Mise en œuvre d'un **outil de "scoring crédit"** pour calculer la **probabilité de défaut de paiement**
- ❑ Développement d'un **modèle de classification**
- ❑ Modèle de ML doté d'une **interprétabilité** afin de garder la **transparence sur les prises de décision d'octroi de crédit**
- ❑ Développement d'un **Dashboard interactif** pour le service de la **relation client** et pour les **clients** eux-mêmes

Constat sur les données à notre disposition

- **Variables explicatives X (features)**
- **Variable cible y (possibilité de prédiction)**

↳ ***Problème d'apprentissage supervisé***



↳ ***Problème de classes déséquilibrées***

I PROBLÉMATIQUE ET PRÉSENTATION DES JEUX DE DONNÉES

Problématique

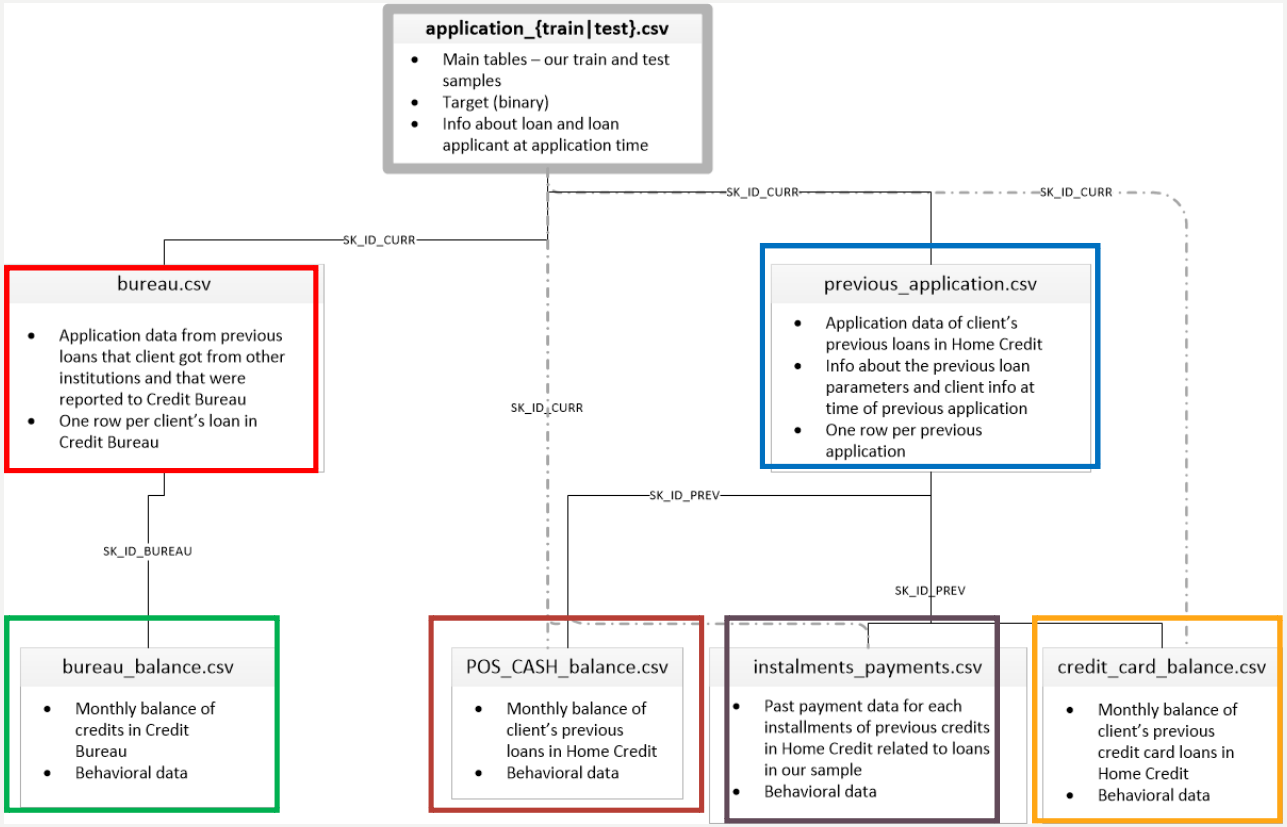
Problématique

- Comment gérer les classes déséquilibrées pour espérer généraliser au mieux notre modèle de classification ?
- Quelles métriques pour les problèmes de classification et quelle est la plus adaptée pour répondre à des impératifs métiers ?
- Par quels moyens pourrait-on réaliser un Dashboard interactif avec surtout le fait que nous travaillons dans un environnement Python ?

Problématique

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Preprocessing et feature engineering



Création de nouvelles features (*)

- *PAYMENT_RATE*
- *PAYMENT_PERC*
- *INCOME_PER_PERSON*
- ...

Traitement des valeurs manquantes

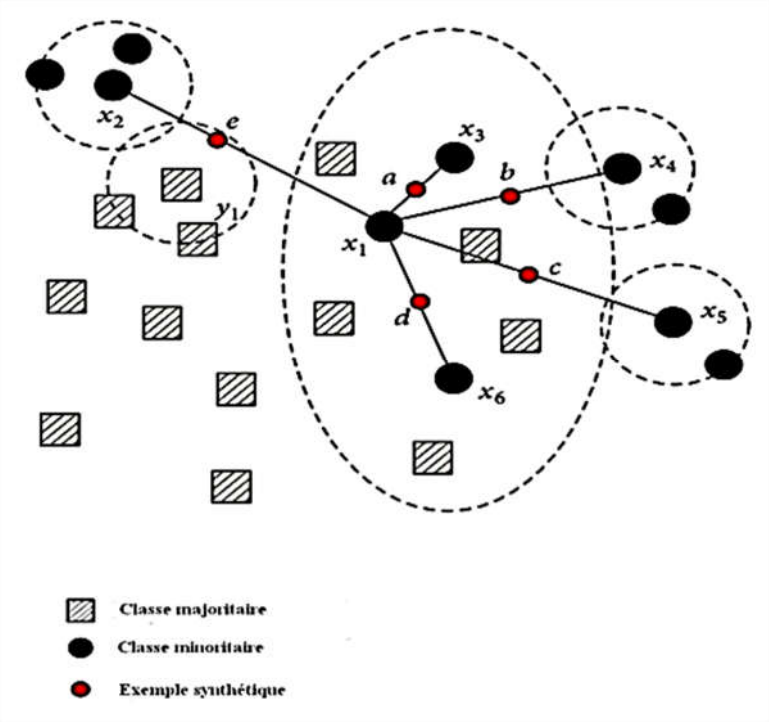
- **Médiane** pour les variables numériques

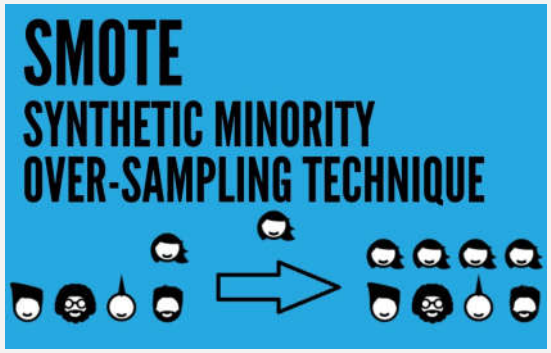
(*) https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Techniques de gestion de classes déséquilibrées

Rééchantillonnage adapté aux données: la méthode ***SMOTE (*)***





Apprentissage ML

Data augmentation

(*) https://kobia.fr/imbalanced-data-smote/

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Techniques de gestion de classes déséquilibrées

Ajustement des poids de classes: la méthode des **class weights**

Class_weight = **'balanced'**

$$w_j = n_{samples}/(n_{classes} \times n_{samplesj})$$

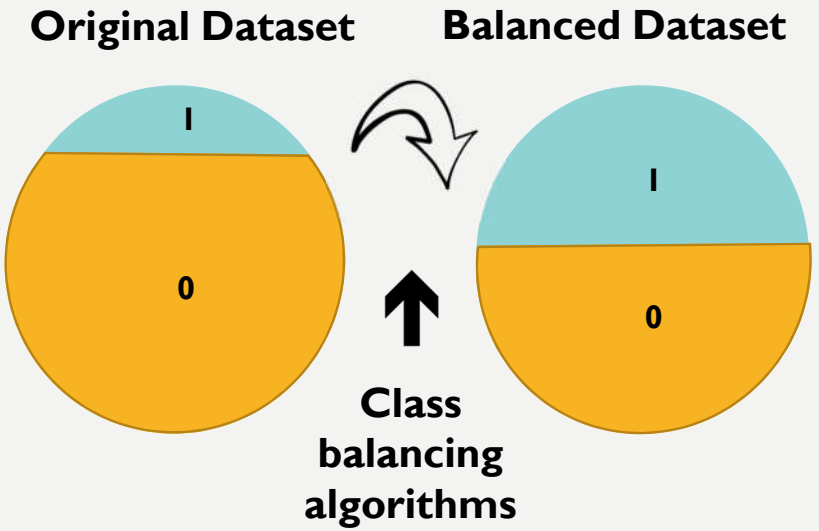
où :

- w_j *est le poids de la classe j*
- $n_{samples}$ *est le nombre total d'observations dans le dataset*
- $n_{classes}$ *est le nombre de classes présentes dans le dataset*
- $n_{samplesj}$ *est le nombre total d'instances de la classe j*

Objectif

Une mauvaise classification d'une observation de la classe minoritaire est plus lourdement pénalisée que la mauvaise classification d'une observation de la classe majoritaire

➡ **Ajustement de la fonction coût du modèle**

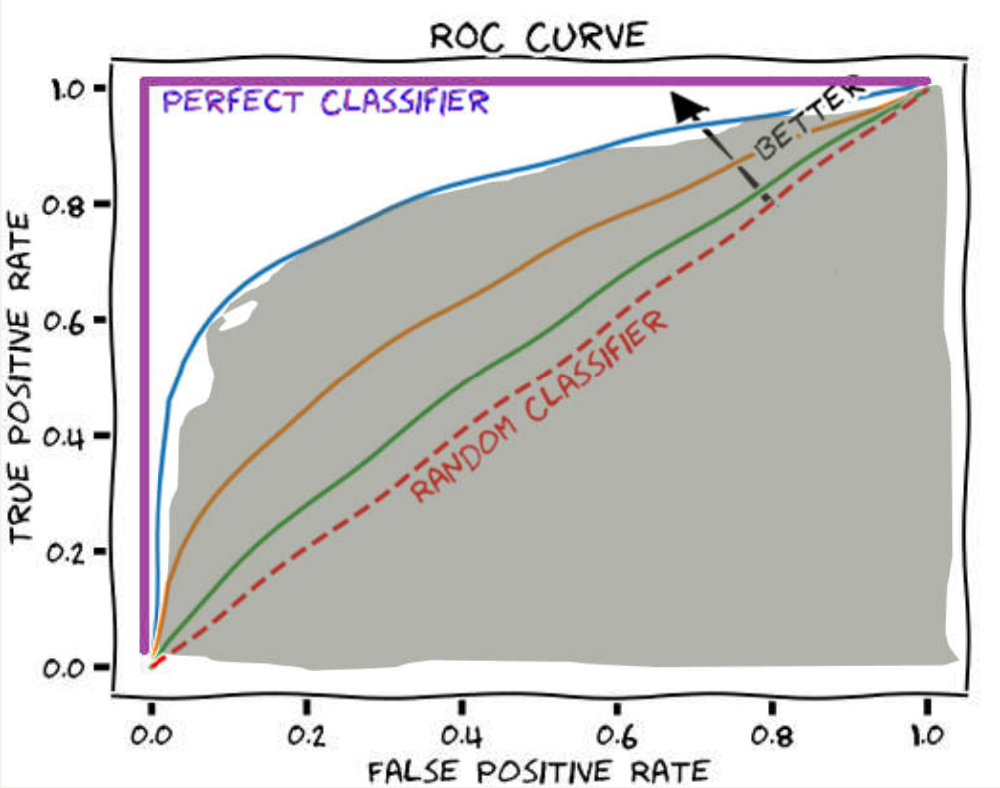


II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Métrique d'évaluation adéquat : l'aire sous la courbe ROC (*)

Matrice de confusion

Matrice de confusion		Classes prédites	
		Négatif : 0	***Positif : 1***
Classes réelles	***Négatif : 0***	Vrais négatifs (VN)	Faux Positifs (FP)
	Positif : 1	Faux négatifs (FN)	Vrais Positifs (VP)

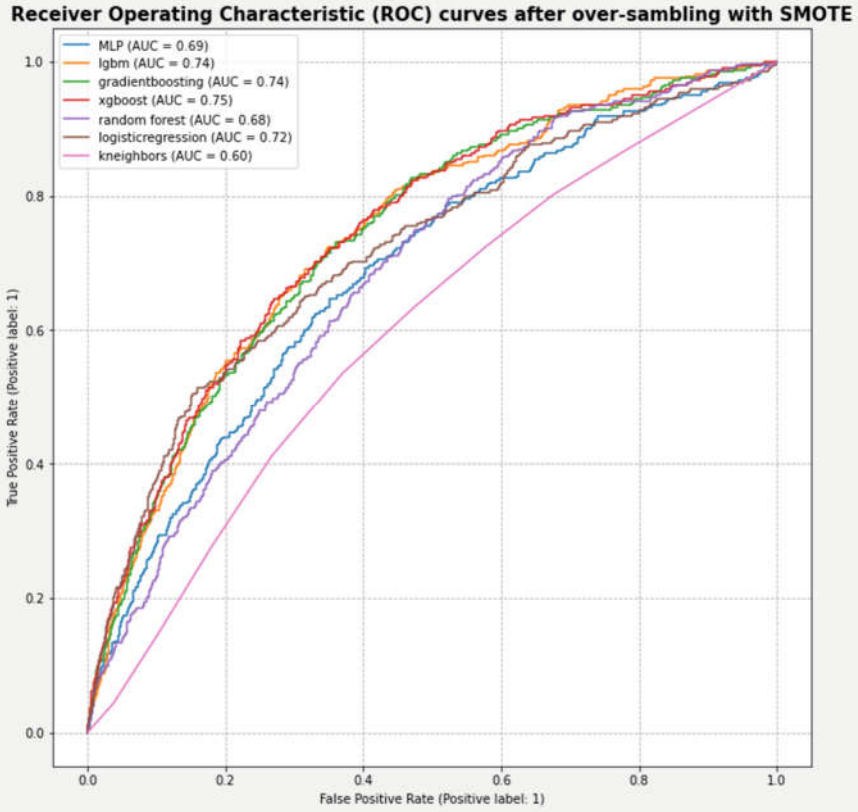


(*) https://progresser-en-maths.com/comment-evaluer-un-modele-de-classification-les-bonnes-metriques/

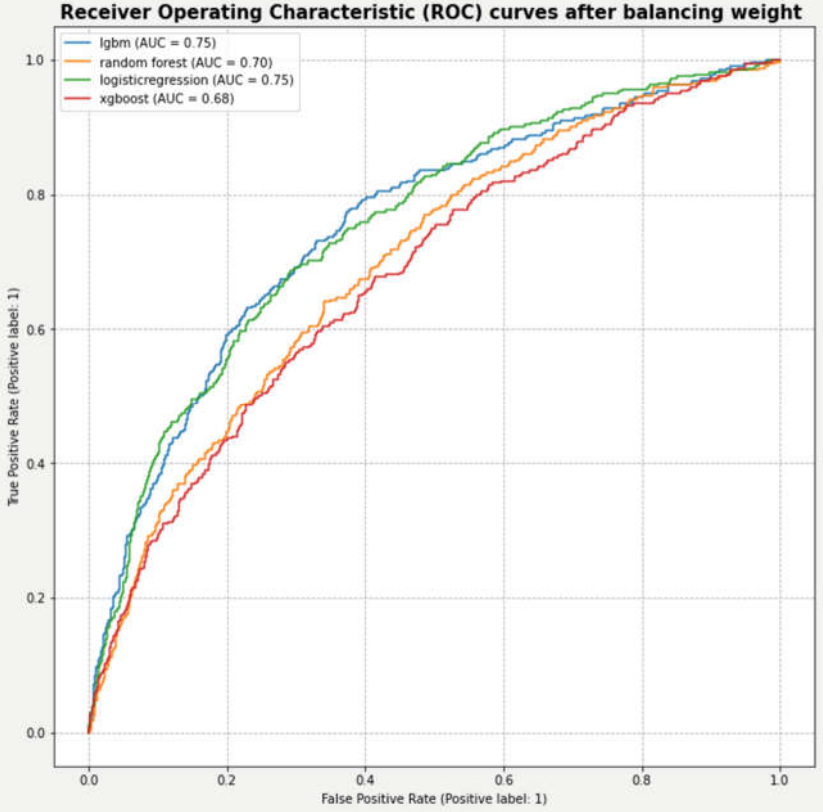
Approche de modélisation

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Après SMOTE



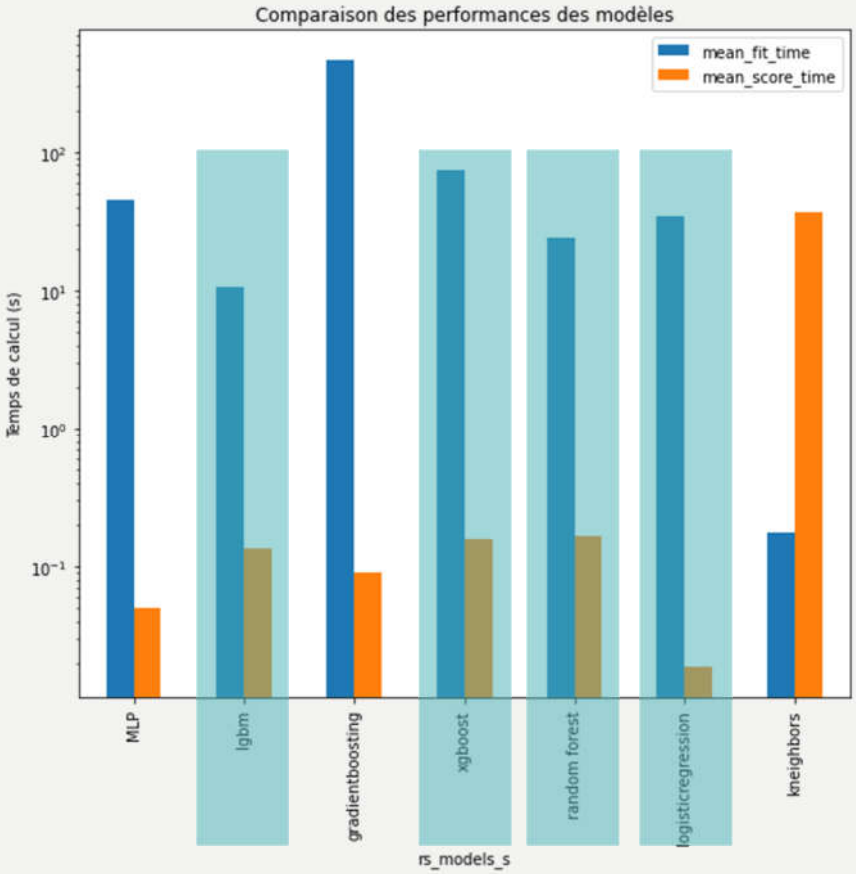
Après ajustement des poids
RandomForest , Lightgbm Logisticregression, et Xgboost



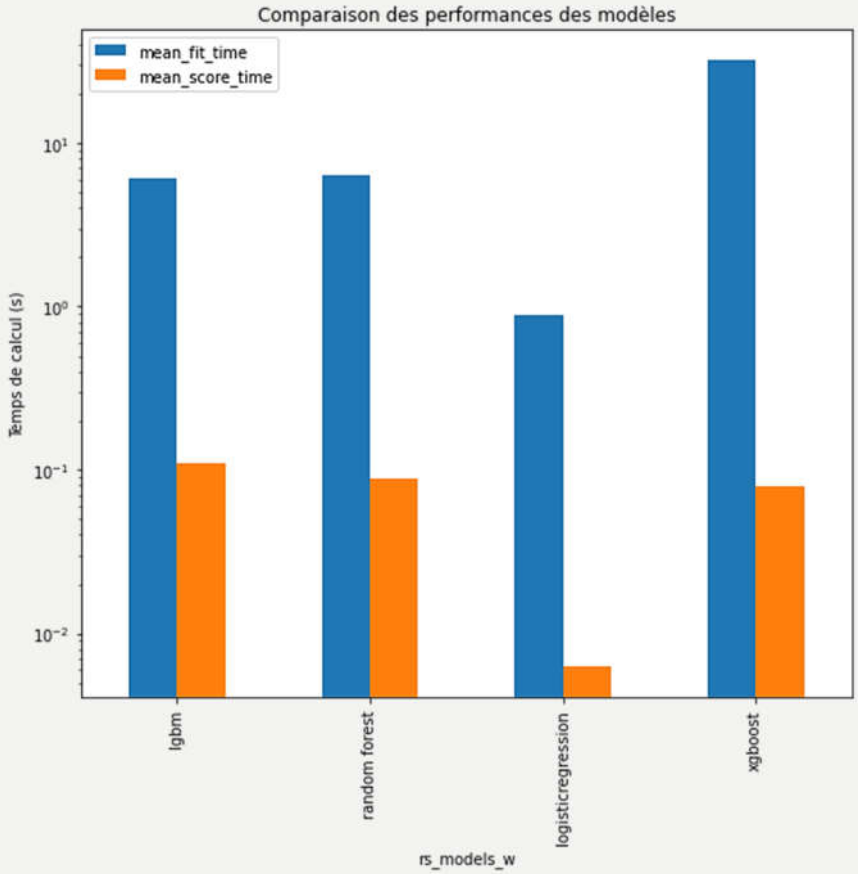
Approche de modélisation

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Après SMOTE



Après ajustement des poids
RandomForest , Lightgbm, Logisticregression, et Xgboost



Approche de modélisation

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Choix de techniques de gestion d'équilibre et de modèles

Aire sous la courbe ROC		MODELES DE CLASSIFICATION						
		knn	mlp	gboost	rf	logreg	xgboost	lgbm
Techniques	SMOTE	0.60	0.69	0.74	0.68	0.72	0.75	0.74
	Ajustement de poids	X	X	X	0.70	0.75	0.68	0.75

Modèles	Abréviation
KNeighbors	knn
MLP	mlp
GradientBoosting	gboost
RandomForest	rf
LogisticRegression	logreg
XGBoost	xgboost
LightGBM	lgbm

SMOTE

- ❑ Efficace contre le surapprentissage
- ❑ Plus coûteux en temps pour synthétiser les nouvelles instances
- ❑ Coûteux lors de la prédiction et l'apprentissage de données
- ❑ Déconseillé quand existence de variables catégorielles encodées dans notre dataset (*)

vs

Ajustement des poids des classes

- ❑ Mise en place très facile de l'ajustement via **class_weight** ou **scale_pos_weight**
- ❑ Plus rapide que SMOTE dans l'apprentissage de données et de prédiction
- ❑ peu d'algorithme dispose de ce paramètre

(*) https://kobia.fr/imbalanced-data-smote/

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Fonction coût et optimisation métier du modèle

La problématique **« métier »** est de prendre en compte qu'un faux négatif coûte en réalité environ 10 fois plus qu'un faux positif.

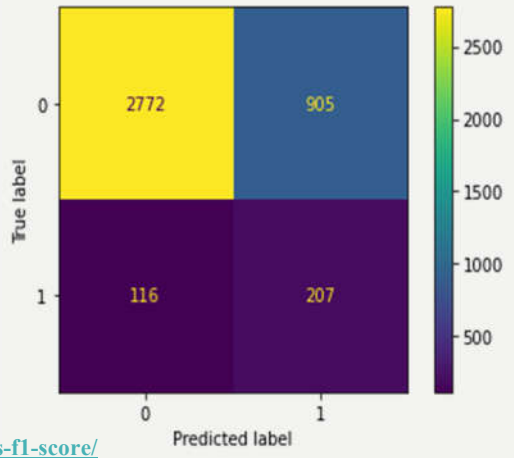
Métrique métier **Fβ-score (*)**

$$F_\beta\text{-score} = \frac{Vrais\ positifs}{Vrais\ positifs + \frac{1}{1+\beta^2}(\beta^2 Faux\ négatifs + Faux\ positifs)}$$

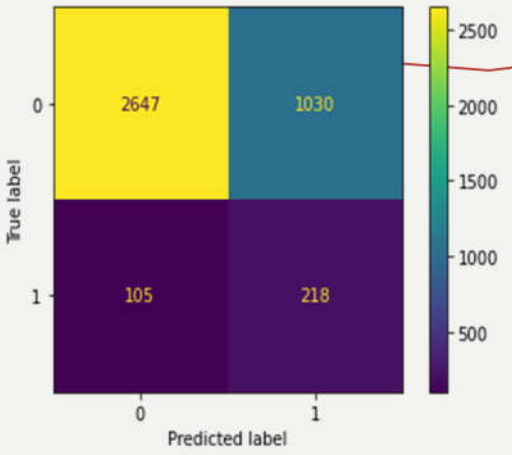
Matrices de confusion: exemple pour β = 2

- (Train set ,16000)
- (Test set, 4000)

Modèle évalué avec l' **AUC de ROC**



Modèle évalué avec **F2-score**



Minimisation des faux négatifs pour β > 1

(*) https://kobia.fr/classification-metrics-f1-score/

Approche de modélisation

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

Fonction coût et Optimisation métier du modèle

Matrices de confusion correspondantes à $\beta \cong \sqrt{10}$

$$(\ )\ F_3\ \ \text{score} = \frac{Vrais\ positifs}{Vrais\ positifs + \frac{1}{10}(9 \times Faux\ négatifs + 1 \times Faux\ positifs)}$$

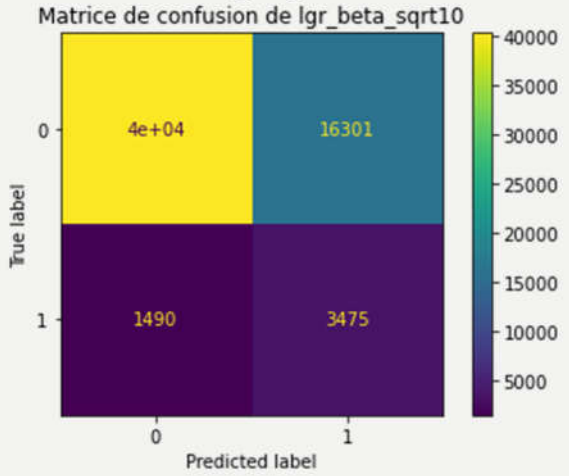
$$(\ )\ F_{\sqrt{10}}\ \ \text{score} = \frac{Vrais\ positifs}{Vrais\ positifs + \frac{1}{11}(10 \times Faux\ négatifs + 1 \times Faux\ positifs)}$$

Train set : 80%, Test set : 20% du jeu initial de données

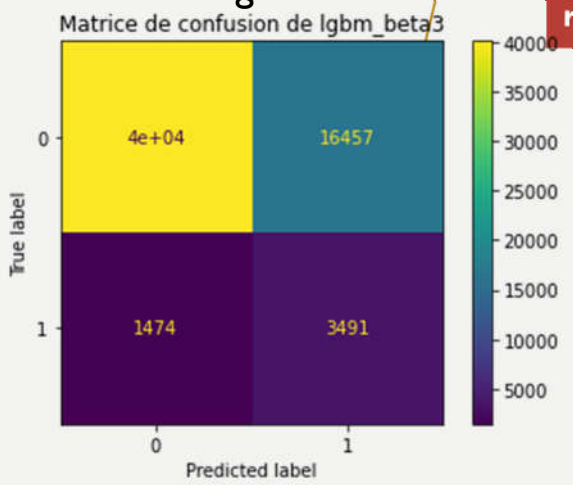
LightGBMClassifier

- class_weight='balanced'
- max_depth=3
- n_estimator=400
- num_leaves=127
- reg_alpha=0.5

logreg



lgbm

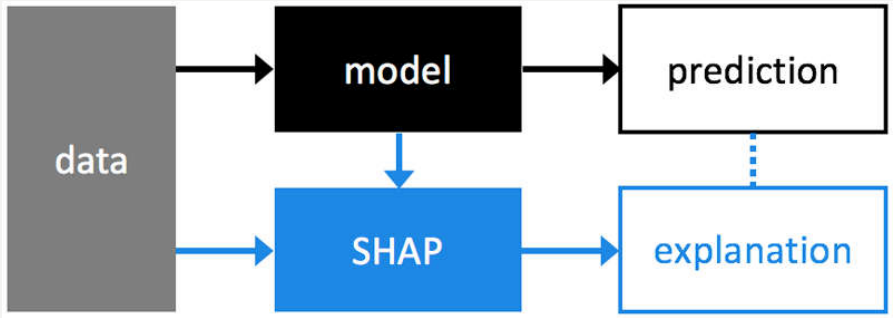


(*) https://kobia.fr/classification-metrics-f1-score/

II APPROCHE DE MODÉLISATION ET INTERPRÉTABILITÉ DU MODÈLE

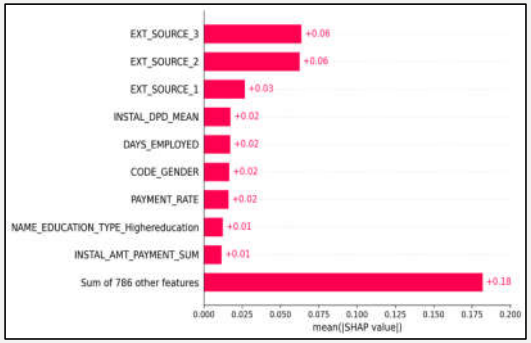
Interprétabilité du modèle choisi

SHAP



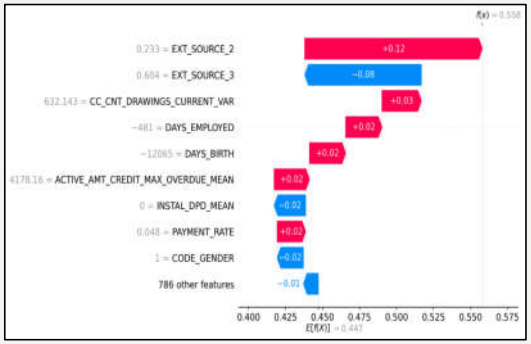
Interprétabilités globales

Interprétabilité par rapport au fonctionnement du modèle d'un point de vue général sur toutes les instances



Interprétabilités locales

Interprétabilité pour une instance donnée



https://medium.com/@ulalaparis/repousser-les-limites-dexplicabilit%C3%A9-un-guide-avanc%C3%A9-de-shap-a33813a4bbfc

https://shap.readthedocs.io/en/latest/index.html

Construction d'api

III PRÉSENTATION DU DASHBOARD

Mise en place d'une api FastAPI et de dashboard via Streamlit

FastAPI

Framework Back-end

- ❑ **Framework Web moderne** et rapide (haute performance)
- ❑ API avec Python **3.6+** basé sur des conseils de type Python standard
- ❑ Open-source

Utilité dans ce projet

- ➢ Score de défaut de paiement
- ➢ Décision sur la demande de crédit

Streamlit

Framework Front-end

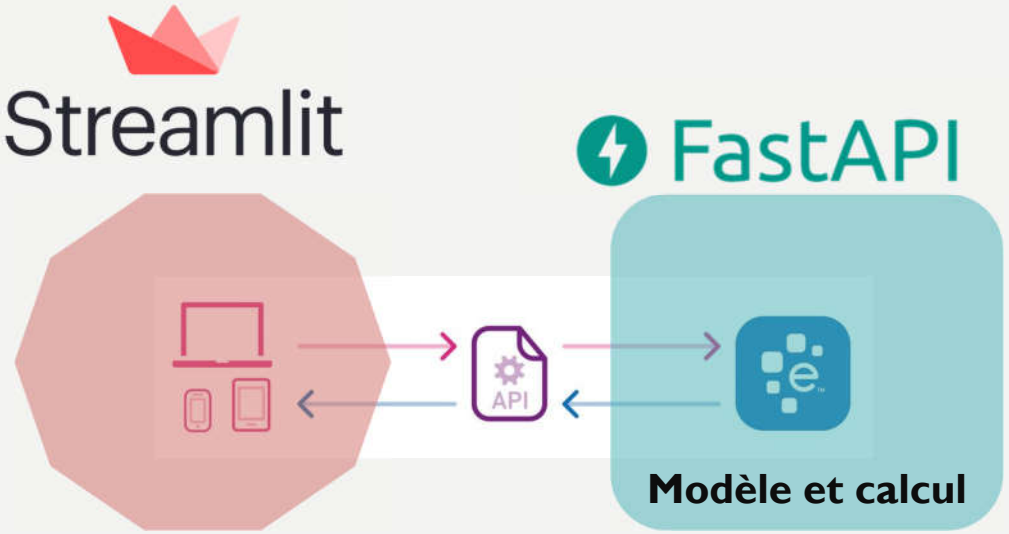
- ❑ Création d'app avec uniquement du code python
- ❑ Intégration facile de la visualisation de data dans l'app, grâce à de nombreux widgets prédéfinis
- ❑ Compatible avec la majorité des frameworks de dataviz (matplotlib, plotly, seaborn,..) et de Machine learning (pandas, pytorch,…)
- ❑ Prédiction et test de modèles de données avec des collaborateurs ou des clients
- ❑ Open-source

https://datascientest.com/streamlit-ou-loutil-pour-presenter-votre-travail-de-machine-learning

Construction d'api

III PRÉSENTATION DU DASHBOARD

Réalisation du Dashboard et déploiement sur Heroku



GITHUB

HEROKU

❖ Rendre disponible ses applications en ligne pour un large public

FastAPI : https://pret-a-depenser-heroku.herokuapp.com/docs

Streamlit : https://pret-a-depenser-board.herokuapp.com/

❖ Web Applications Monitoring

Merci pour votre attention