

# IMPLÉMENTATION D'UN MODÈLE DU SCORING



**Victoire MOHEBI**

Janvier 2023



**OPENCLASSROOMS**



# AGENDA

- **Problématique & mission**
- **Présentation des données**
- **Nettoyage de données & EDA**
- **Modélisation**
- **Optimisation du modèle**
- **Interprétabilité du modèle**
- **Dashboard interactive**
- **Conclusion & piste d'amélioration**

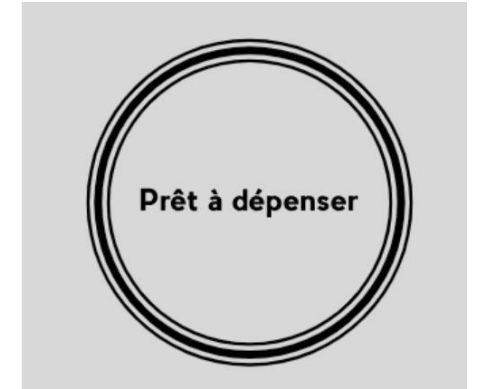


# PROBLÉMATIQUE & MISSION



# PROBLÉMATIQUE

- Société financière « Prêt à dépenser » propose des crédits à la consommation
- Nécessité d'un outil de *scoring* pour savoir si le client rembourse ses dettes
- Impératif de transparence vis-à-vis des décisions d'octroi de crédit (explicabilité)



# MISSION



- Développer un modèle qui donne la prédiction de défaut d'un client (classification binaire)
- Un modèle de *scoring* pour calculer la probabilité qu'un client rembourse son crédit ou fasse défaut
- Construire un *dashboard* interactif pour le chargé de client pour l'interprétabilité du modèle

# PRÉSENTATION DES DONNÉES



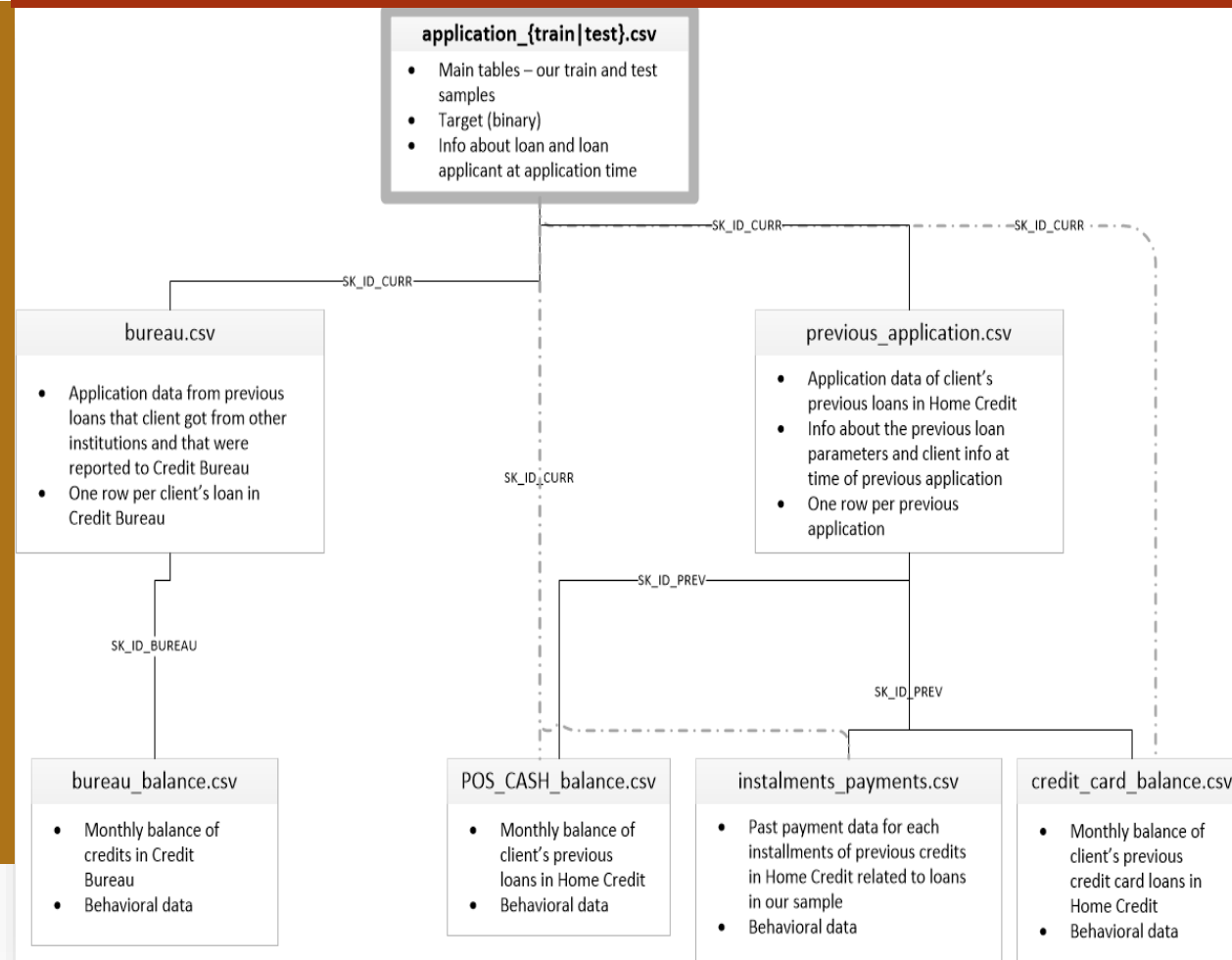
- **Source de donnée :** <https://www.kaggle.com/c/home-credit-default-risk/data>
- **Le jeu de données est composé de 8 fichiers CSV**

Nom de fichier	Description du fichier
application_train.csv Application_test.csv	Les principales données de formation avec des informations sur chaque demande de prêt chez Prêt à dépenser.
Bureau.csv	Données concernant les crédits antérieurs du client auprès d'autres institutions financières
Bureau_balance.csv	Données mensuelles détaillées sur les crédits précédents dans le fichier bureau
credit_card_balance.csv	Données mensuelles sur les cartes de crédit précédentes que les clients ont eues avec Prêt à dépenser.
Installment_payment.csv	Historique de paiement pour les prêts précédents chez Prêt à dépenser.
Previous application.csv	Demandes précédentes de prêts chez Prêt à dépenser des clients qui ont des prêts dans le fichier application_train
POS_CASH_balance.csv	Données mensuelles sur les clients précédents.



- 7 fichiers sont reliées par des « Primary key »
- 218 variables comportementales et financières sur l'emploi, le cadre de vie, l'historique de crédit, pour chaque client
- +300000 demandes de crédit sont enregistrées

## Modèle de base de données

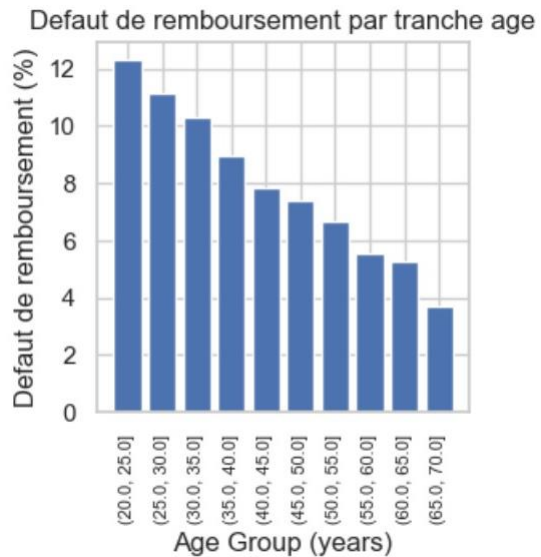
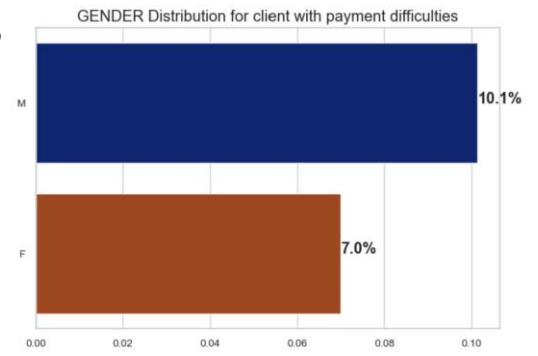
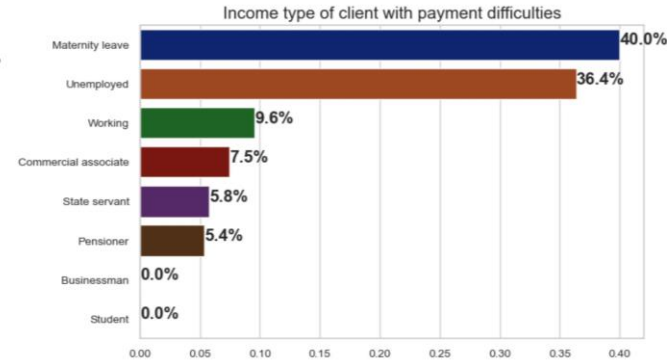
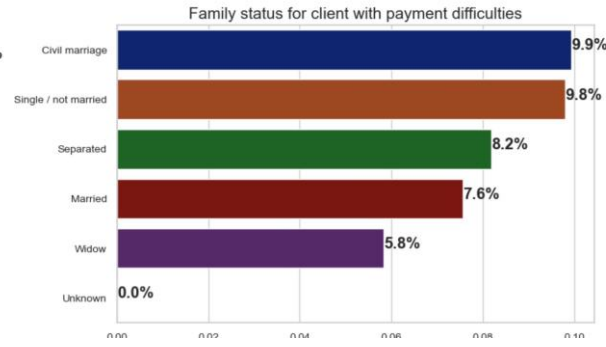
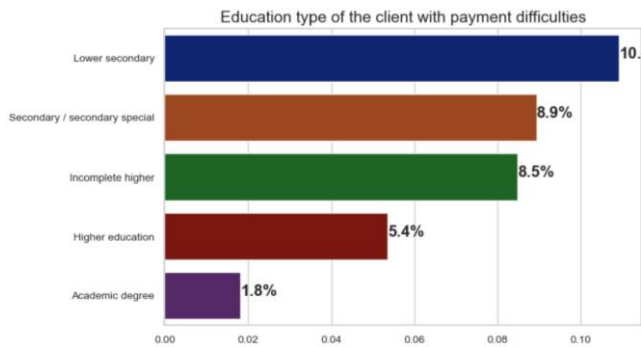




# NETTOYAGE DE DONNÉES & ANALYSE EXPLORATOIRE



# ANALYSE EXPLORATOIRE



- La catégorie "Lower Secondary" a le plus fort taux de non remboursement du prêt (~11%).
- En termes de pourcentage de non-remboursement du prêt, le mariage civil a le pourcentage le plus élevé (~10 %), les veufs étant la plus faible.
- Les client en congé de maternité ont un taux 40% de prêts non remboursés, suivis des chômeurs (~36%).
- Les hommes sont plus susceptible de ne pas rembourser leurs prêts (10 %), par rapport aux femmes (7 %).
- Les clients jeunes sont plus susceptibles de ne pas rembourser le prêt !
- Le taux d'impayés est supérieur à 10 % pour les trois tranches d'âge les plus jeunes et inférieur à 5 % pour la tranche d'âge la plus élevée.

# PRÉ-TRAITEMENT DES DONNÉES

## ➤ **Feature engineering :**

Les pré-traitements effectués ( jointure ente les dataframes, agrégation des variables numériques, encodage des variables catégorielles, rajout des features métier sont inspirés de ce [notebook](#)

## ➤ **Traitement des valeurs aberrantes :**

Pour certaines variables catégorielles, des valeurs aberrantes apparaissent comme « XNA » (comme “CODE\_GENDER”)  
Pour le variable numériques, DAYS\_EMPLOYED, les valeur aberrante > 1000 ans ont été remplacées pap NaN.

## ➤ **Traitement des valeurs manquantes :**

Les variables ayant plus de 40% de valeurs manquantes ont été supprimées. Les autres valeur manquantes sont remplacées par le “median”. **Le dataframe final a 356,251 lignes et 171 colonnes**

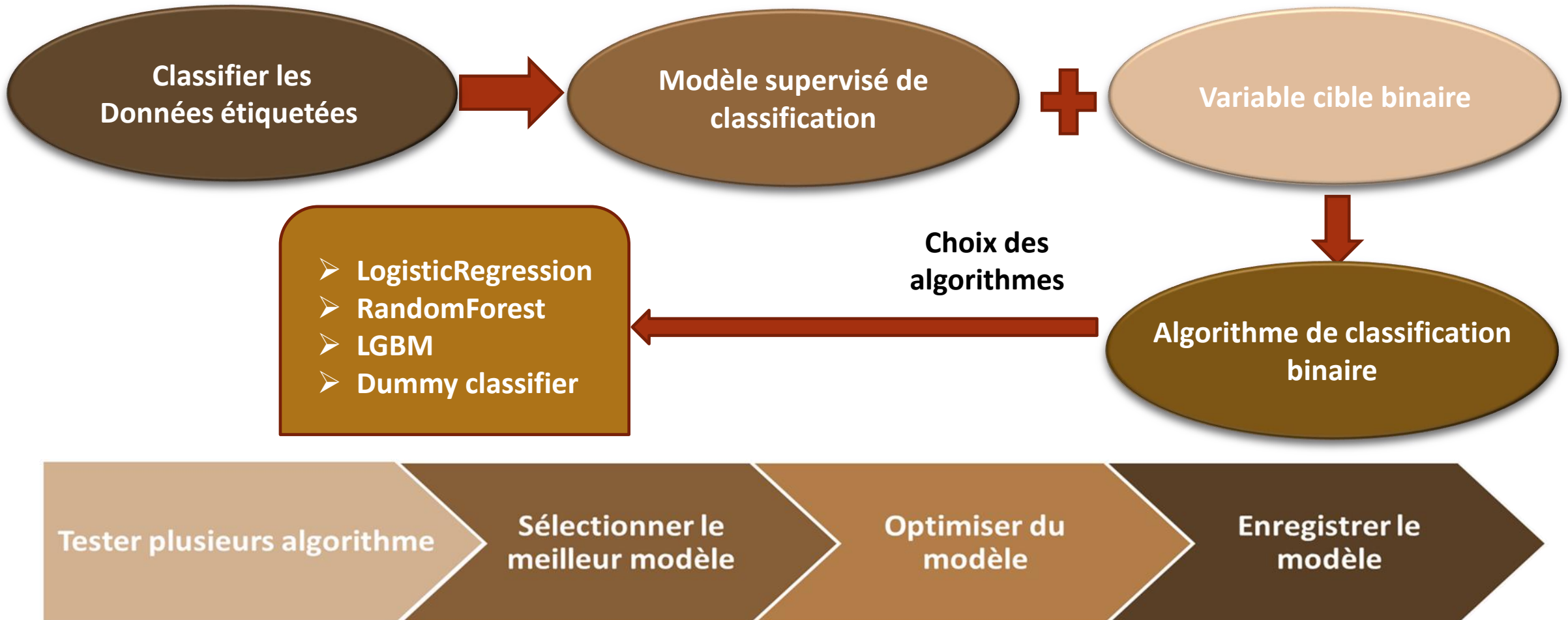
➤ **Feature scaling :** on a eu recours à la Standardisation qui est le processus de transformer une variable en une autre qui répondra à la loi normale.

➤ **Split Train/Test set :** 80% des données pour l’entraînement et la validation du modèle. 20% restantes pour le tester. Dans cette opération les mêmes proportions des différentes classes ont été gardé ( avec l’argument *stratify*)

# MODÉLISATION



# PROCESSUS DU MODÉLISATION

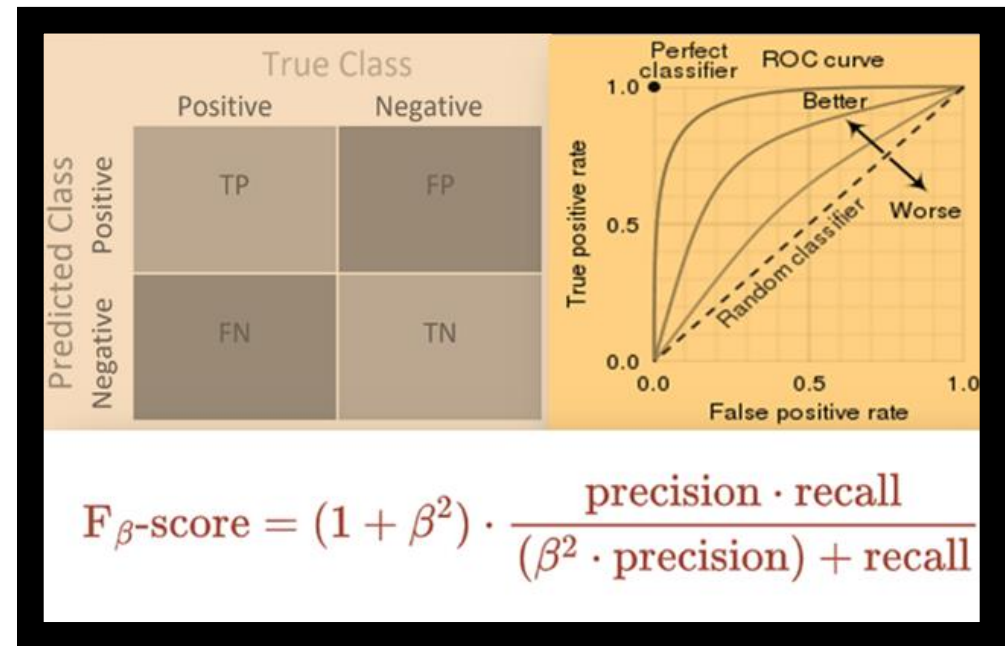


# MÉTRIQUE SPÉCIFIQUE

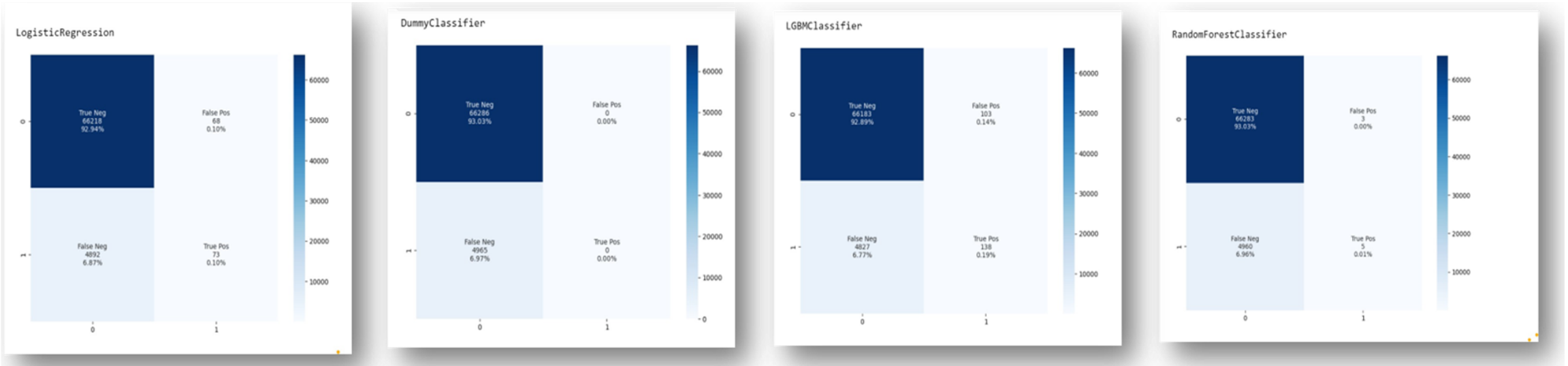
- L'utilisation de mesures plus simples comme le "*Precision*" ou "*Accuracy*" peut être trompeuse.
- Les clients non solvables que l'algorithme ne détectera pas (Faux Négatif, target =0) coûteront plus cher à la société financière que le coût d'un client solvable prédit comme non solvable.
- Minimiser le taux de Faux Négatif

## Choix des métriques

- Matrice de confusion
- Métrique technique : AUC\_ROC
- Métrique du métier : F5-Score



# COMPARAISON DES MODÈLES



	Model	AUC	Accuracy	Precision	Recall	F1	F5	Time
3	LGBMClassifier	0.780642	0.930429	0.516949	0.024572	0.046914	0.025506	5.63856
1	LogisticRegression	0.763982	0.930485	0.545455	0.014502	0.028252	0.015066	4.55621
2	RandomForestClassifier	0.720825	0.930457	1.0	0.002014	0.00402	0.002094	237.468794
0	DummyClassifier	0.5	0.930317	0.0	0.0	0.0	0.0	0.191926

**LGBMClassifier a simultanément le meilleurs :**

- **AUC\_ROC (score technique)**
- **F5 (score métier)**
- **Temps d'entraînement**



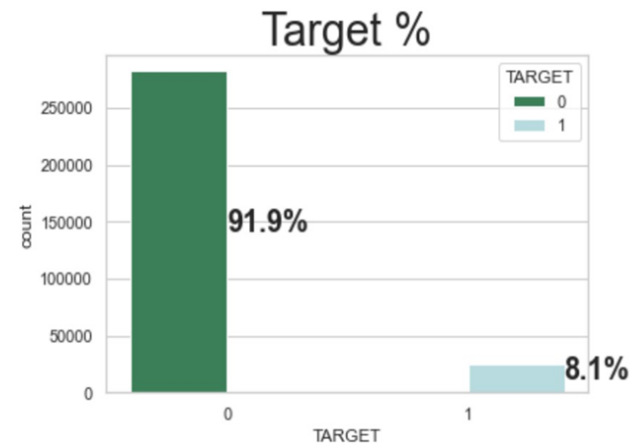
# DONNÉES DÉSÉQUILIBRÉES, QUOI FAIRE?

**Target = 1 : Client ne peut pas payer le crédit**

**Target = 0 : Client peut payer le crédit**

Les deux modalités de la variable cible ne sont pas représentées de façon égale dans l'échantillon.  
La classe 0 est fortement majoritaire.

Tenter de ré-équilibrer l'échantillon pour aider les algorithmes à mieux détecter les individus de la classe minoritaire

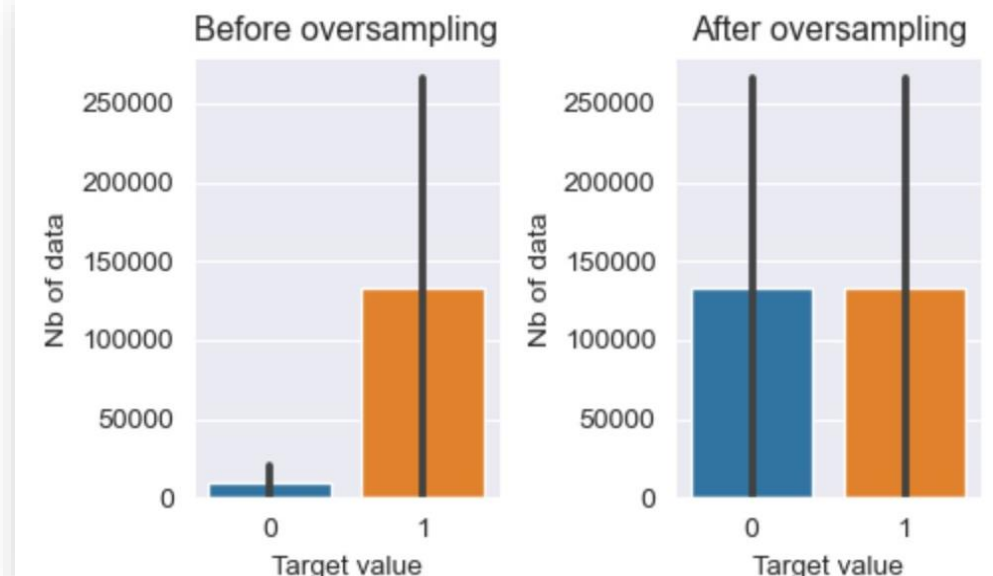


- 92 % des prêts ont été remboursés
- 8 % des individus ont été non-solvables.

# TECHNIQUE DE RÉÉCHANTILLONNAGE

- Une technique largement adoptée pour traiter des ensembles de données très déséquilibrés est appelée rééchantillonnage.
- Elle consiste à retirer des échantillons de la classe majoritaire (sous-échantillonnage) et/ou à ajouter d'autres exemples de la classe minoritaire (sur-échantillonnage).

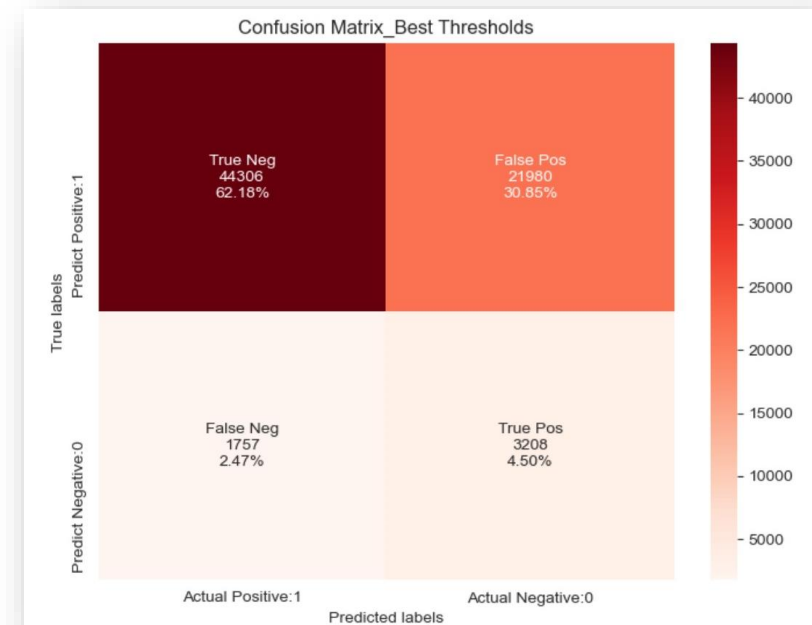
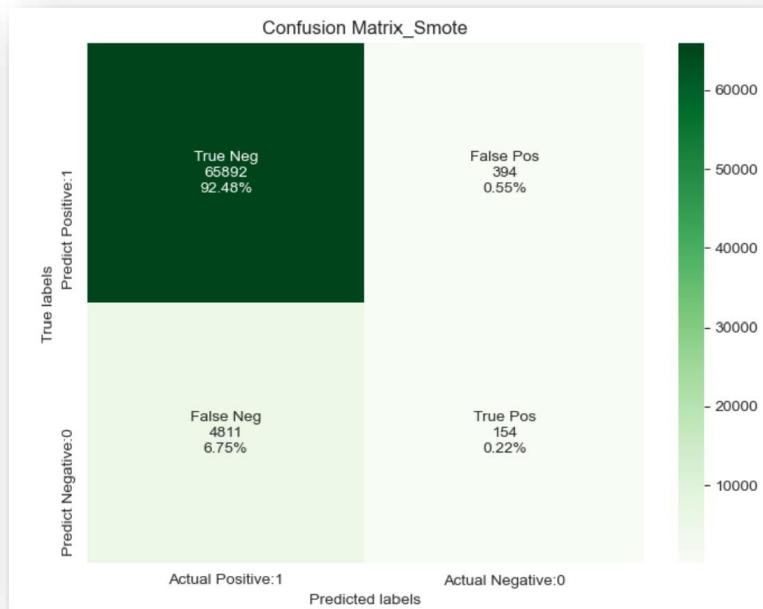
- On utilise *SMOTE* (Synthetic Minority Oversampling Technique) pour suréchantillonner la classe minoritaire (prêts non remboursés, cible = 1)
- On utilise l'argument *class\_weight = balanced* des modèles *LGBMClassifier*



# OPTIMISATION DU MODÈLE SMOTE

Optimisation des hyperparamètres avec *GridSearchCV* pour obtenir le meilleur *AUC\_ROC* (score technique) ou *F5* (score métier)

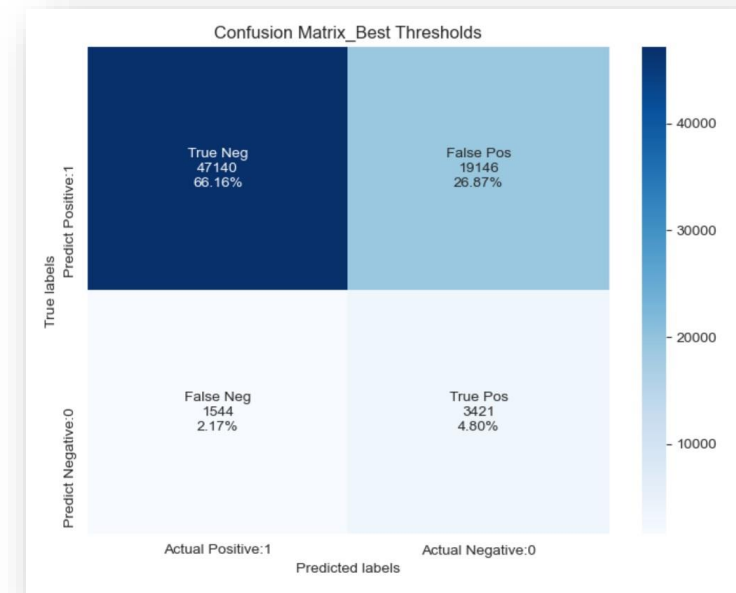
Trouver la meilleur seuil de solvabilité avec la *méthode Statistique J de Youden*:  $J = TP - FN$   
Choisir la plus grande valeur de J.



# OPTIMISATION DU MODÈLE CLASS\_WEIGHT

Optimisation des hyperparamètres avec *GridSearchCV* pour obtenir le meilleur *AUC\_ROC* (score technique) ou *F5* (score métier)

Trouver la meilleur seuil de solvabilité avec la méthode *Statistique J de Youden*:  $J = TP - FN$   
Choisir la plus grande valeur de J.



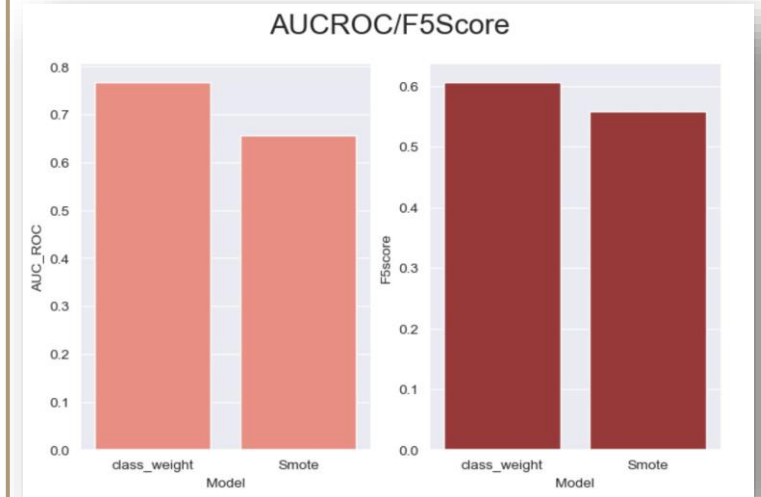
# SÉLECTIONNER LE MEILLEUR MODÈLE

- On constate que le modèle avec *class\_weight* a le meilleur performance : score métier & score technique plus élevés

Classification Report_Class-weight					
	precision	recall	f1-score	support	
0.0	0.97	0.71	0.82	66286	
1.0	0.15	0.69	0.25	4965	
accuracy			0.71	71251	
macro avg	0.56	0.70	0.53	71251	
weighted avg	0.91	0.71	0.78	71251	

- On constate que le modèle arrive à détecter 69% des classes 1

Classification Report after SMOTE					
	precision	recall	f1-score	support	
0.0	0.96	0.67	0.79	66286	
1.0	0.13	0.65	0.21	4965	
accuracy			0.67	71251	
macro avg	0.54	0.66	0.50	71251	
weighted avg	0.90	0.67	0.75	71251	



# INTERPRÉTABILITÉ DU MODÈLE



# FEATURE IMPORTANCE

## GLOBAL VS LOCAL



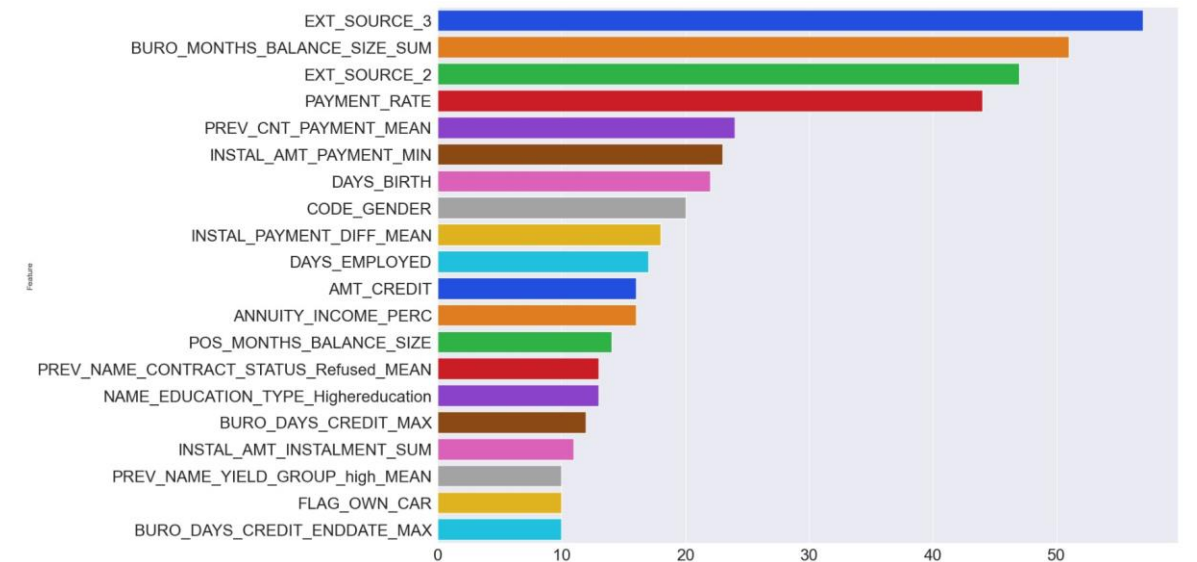
- L'interprétabilité : comprendre les étapes et les décisions prises par le modèle
- La possibilité de comprendre davantage les aspects suivants :
  - Quelles variables sont importants pour le modèle ?
  - Pourquoi le modèle est-il arrivé à cette conclusion ?
- Local feature importance se concentrent sur la contribution des variable pour une prédiction spécifique
- Global feature importance prennent en compte toutes les prédictions



# INTERPRÉTABILITÉ GLOBALE

## Feature importance

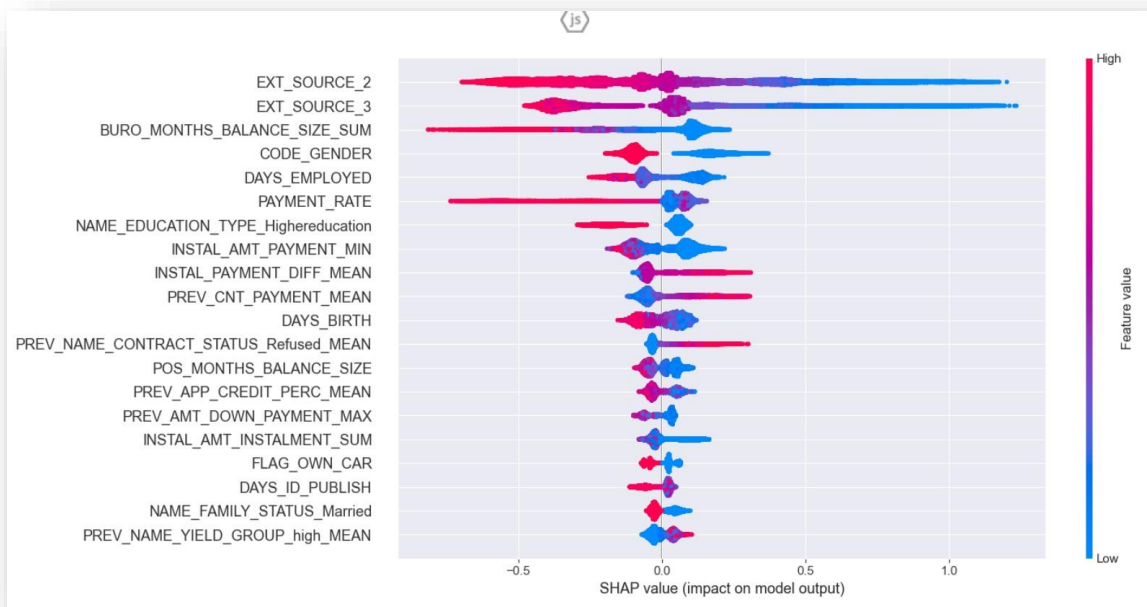
- 20 plus importantes variables que le classifieur LGBM a utilisé pour prédire la probabilité de remboursement du prêt
- On constate que les variables external sources et les mensualité payé par le client sont plus important pour le modèle



# INTERPRÉTABILITÉ GLOBALE

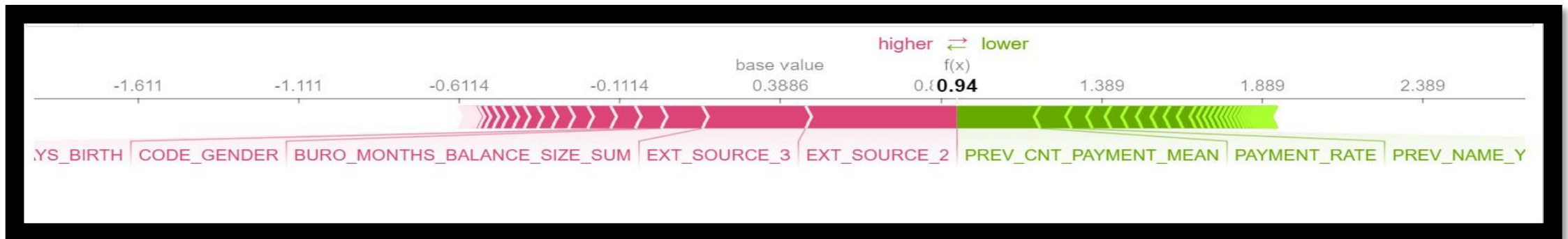
## Shaply Values

### SHAP Summary Plot



- Ensemble des SHAPley valeurs par observation.
- La couleur des points correspond à la valeur de la variable et le positionnement horizontal des points correspond à la SHAPley valeur.
- Pour chaque variable si les SHAPley valeurs sont négatives (situées sur la gauche) donc la variable est en défaveur de la prédiction CLASSE 1.
- Au contraire, pour les point rouge, les SHAPley valeurs sont positives (situées sur la droite) donc en faveur de la prédiction de la CLASSE 1.

# INTERPRÉTABILITÉ LOCALE







- Ce graphique montre quelles sont les principales variables affectant la prédiction d'une seule observation , et l'ampleur de la valeur SHAP pour chaque variable.
- En rouge, les variables qui ont un impact positif (contribuent à ce que la prédiction soit 1) et, en vert, celles ayant un impact négatif (contribuent à ce que la prédiction soit 0).
- Le crédit de cette personne en particulier a été refusé, car elle a été poussée plus haut par tous les facteurs indiqués en rouge.

# DASHBOARD INTERACTIF



# CONSTRUCTION DE DASHBOARD

Plateforme	Description
	<ul style="list-style-type: none"><li>- <b>GitHub</b> est un service web pour l'hébergement et la gestion de développement de logiciels</li><li>- <b>Git</b> qui est un logiciel libre pour le versionnage des projets</li></ul>
	<ul style="list-style-type: none"><li>- <b>FastAPI</b> est un framework de back-end moderne et rapide (haute performance) pour la création d'API avec Python 3.6+</li><li>- Les APIs contiennent tous les end points pour interagir avec d'autres logiciels.</li></ul>
	<p><b>Amazon Elastic Compute Cloud</b> ou <b>EC2</b> est un service proposé par Amazon permettant à des tiers de louer des serveurs sur lesquels exécuter leurs propres applications web.</p>
	<ul style="list-style-type: none"><li>- <b>Streamlit</b> est une plate-forme open-source de front_end pour créer des applications avec python.</li></ul>

## FastAPI :

URL Locale : <http://127.0.0.1:8000/docs>

URL Network: <http://35.180.66.152/>

## ➤ Dashboard interactif :

URL Local : <http://localhost:8501/>

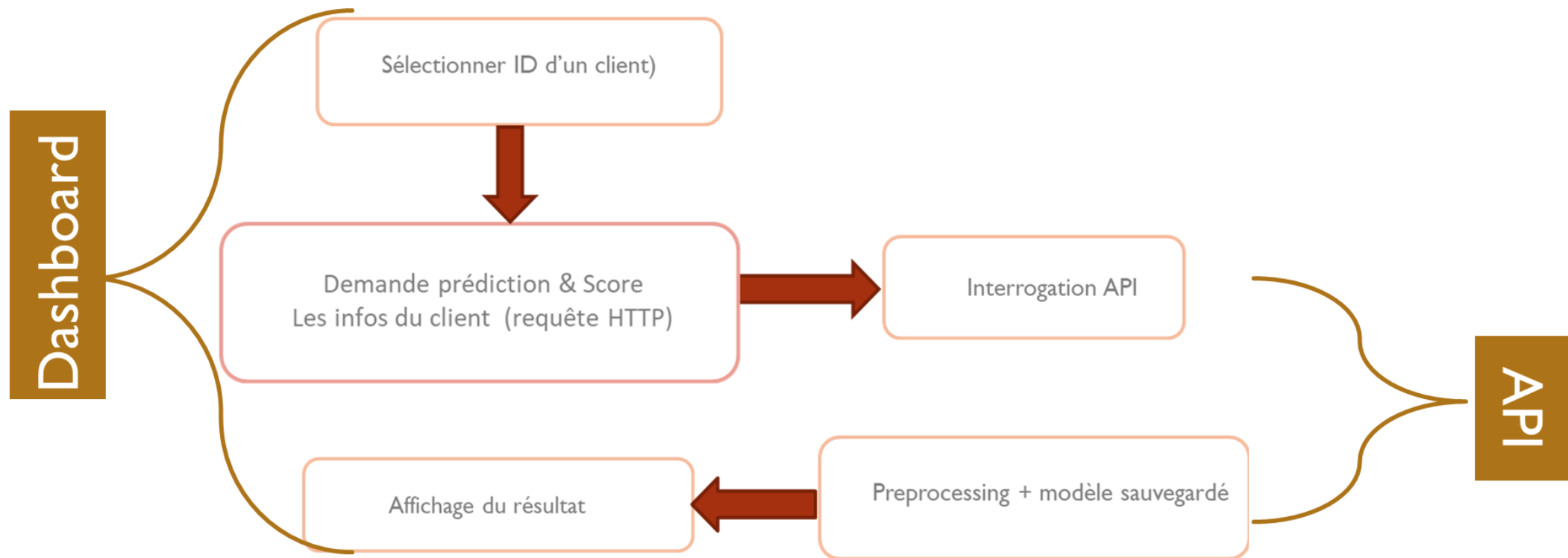
URL Network : <https://victoire76-projet07-dashboard-dashboard-2sfss9.streamlit.app/>

## ➤ GitHub repo pour *FastAPI* :

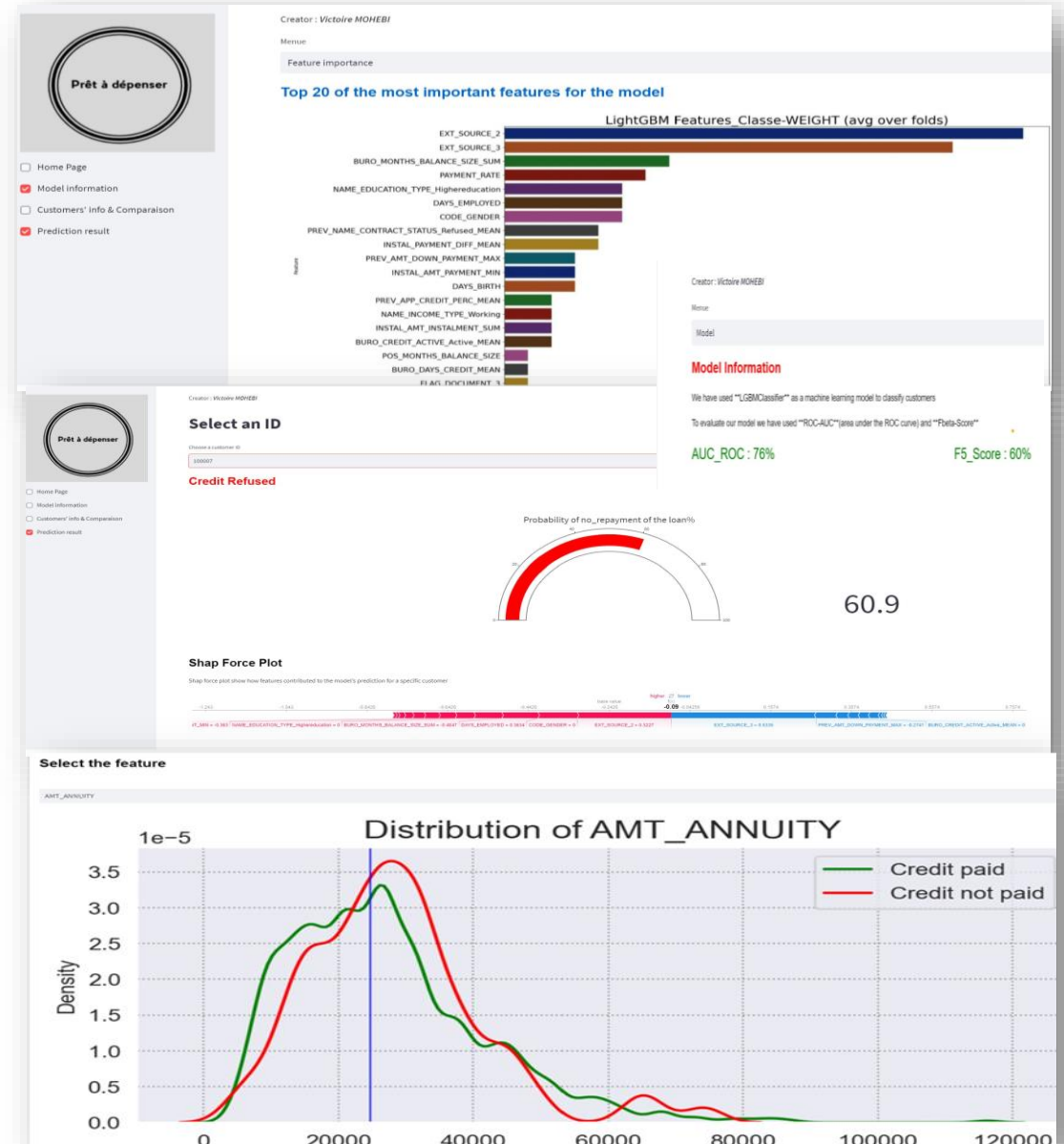
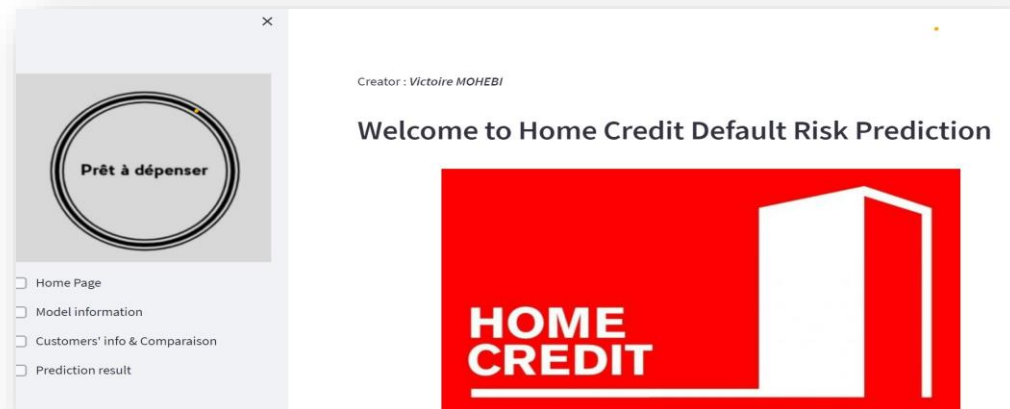
[https://github.com/Victoire76/OC\\_Project07\\_API\\_DASHBOARD](https://github.com/Victoire76/OC_Project07_API_DASHBOARD)

## ➤ Github repo pour *Share Streamlit* :

[https://github.com/Victoire76/Projet07\\_Dashboard](https://github.com/Victoire76/Projet07_Dashboard)



- Infos sur le modèle; ex. la métrique et l'importance des variables
- Possibilité de sélectionner un client selon son ID
- Infos principaux sur le client ; ex. Cadre de vie (l'âge, genre, emploi, éducation, le salaire), le montant du crédit
- Score (la probabilité) entre 0 et 1 que le client ne rembourse pas ses dette
- Explicabilité du modèle en comparer le client avec les autres sur certaine variables





# CONCLUSION & PISTE D'AMÉLIORATION



## Conclusion

- Meilleur modèle : LightGBMClassifier  
(AUC = 0.76 / F5score = 0.6)
- Métrique spécifique : la modélisation a été effectuée sur la base d'avoir le meilleur F5 score qui donne beaucoup d'importance au rappel (recall)
- Construction d'un Dashboard + API fonctionnels : mieux comprendre le résultat de la prédiction pour chaque client en prenant compte l'impact des variables sur la prédiction



## Piste d'amélioration

- Traiter des valeurs manquantes
- Echanger avec les experts du métier pour un meilleur *feature selection*
- Optimiser le temps de calcul
- Optimiser la performance du modèle en ajustant plus de paramètres
- Définir une fonction coût métier plus adaptée aux besoin de « Prêt à Dépenser »
- Construire un *Dashboard* plus complet :
  - Optimiser les performances pour un chargement de données plus rapide
  - Ajouter d'autres onglets pour une analyse complète



**MERCI DE VOTRE ATTENTION!**