



# Formation Data Scientist OpenClassrooms

Projet 7: Implémentez un modèle de scoring

[Projet commencé avant le 14/12/2022](#)

Etudiant : Monine Chan

Evaluateur : Kezhan Shi

Dimanche 26 Février 2023



# PLAN

1. Rappel la problématique et du jeu de données (5min)
2. Explication de l'approche de modélisation (10min)
3. Présentation du dashboard (5 min)
4. Conclusion



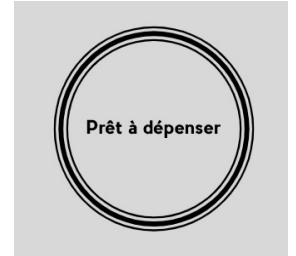
# PLAN

1. Rappel la problématique et du jeu de données (5min)
2. Explication de l'approche de modélisation (10min)
3. Présentation du dashboard (5 min)
4. Conclusion

# 1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

## PROBLÉMATIQUE

- La société financière « Prêt à dépenser » souhaite **mettre en œuvre un outil de « Scoring crédit »** pour **calculer la probabilité** qu'un client rembourse le crédit et classer la demande en crédit accordé ou refusé en se fondant sur multiples sources de données.
- Les clients souhaitent plus de **transparence** sur les raisons de la décision d'octroi du crédit.
- Le but de ce projet est de :
  - Construire un **modèle qui prédira la probabilité** de faillite d'un client à rembourser le crédit (**classification binaire**).
  - Construire un **dashboard interactif** qui permet de faire la prédiction et d'aider à interpréter la décision d'octroi du prêt.



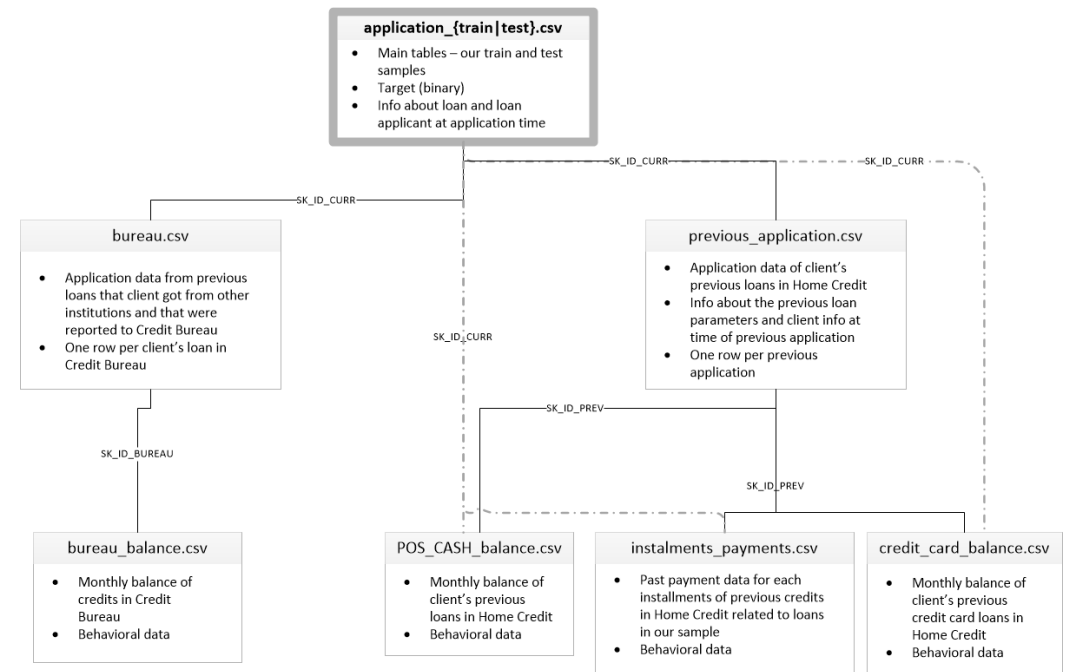
# 1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

## JEUX DE DONNEES



- Il y a 8 fichiers .csv listant les données des clients (transactions bancaires, salaire, demande de prêt etc.)
- La classe à prédire est TARGET :
  - TARGET = 0 ⇔ prêt accordé
  - TARGET = 1 ⇔ prêt refusé
- Le jeu de données **application\_train** est utilisé pour l'**entraînement** et la **validation** des modèles : il contient TARGET.
- Le jeu de données **application\_test** ne contient pas TARGET : utilisé pour la compétition Kaggle.
- On utilise le kernel Kaggle fourni pour merger les 8 dataframes dont résulte un dataframe mergé de 307 511 lignes.

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



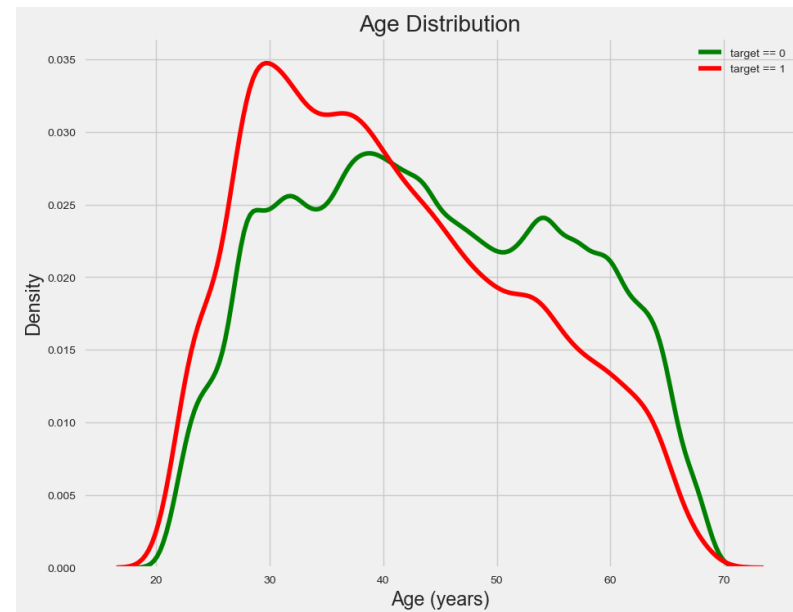
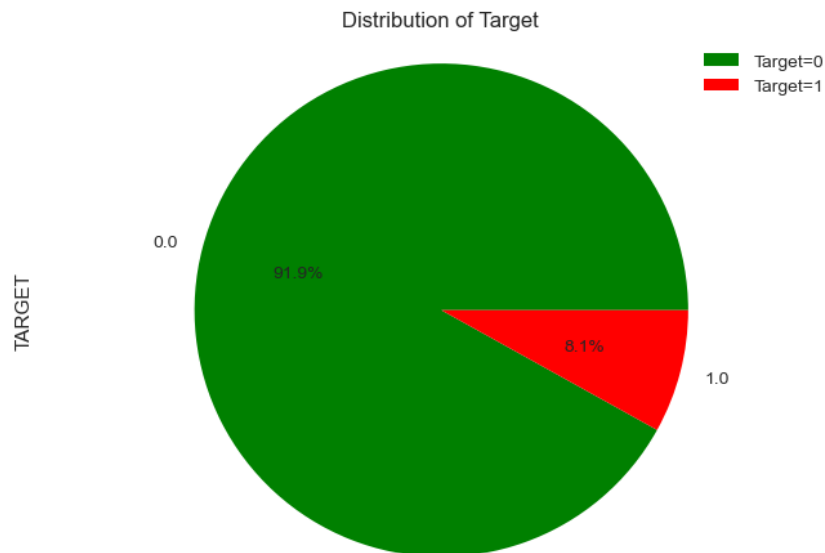
# 1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

## VARIABLES : DÉSÉQUILIBRE ET ÂGE



- On voit tout d'abord que le jeu de données est déséquilibré puisque 92 % de la classe 0 est représenté contre seulement 8% pour la classe 1.

- On a également accès aux variables qui caractérisent les client comme leur âge.



# 1. RAPPEL DE LA PROBLÉMATIQUE ET DU JEU DE DONNÉES

## TRAITEMENT DES DONNÉES



**Step 1 (Kernel Kaggle):**  
Merge des 8 dataframes  
Encodage et feature engineering

**Step 2:**  
Traitement des NaN et  
suppression des valeurs 'inf'

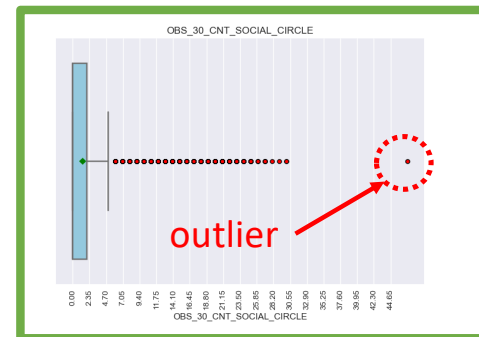
**Step 3:**  
Identification des outliers  
via boxplot et suppression

**Step 4:**  
Préparation des jeux  
d'entraînement/validation  
et de test  
(TARGET présente)

E  
X  
E  
M  
P  
L  
E

Création de la variable :  
PAYMENT RATE =  
AMT\_ANNUITY / AMT\_CREDIT  
montant du crédit remboursé  
par an / montant du crédit  
→ **Taux de remboursement**

NaN remplacé par la médiane.  
Valeurs 'inf' supprimées



Suppression de TARGET pour  
les features (X)  
Création du jeu de données y  
(TARGET uniquement)

N  
O  
T  
B  
O  
O  
K  
S

P7\_Step1\_Traitement\_Kernel\_Ka  
ggle\_Nettoyage.ipynb

P7\_Step2\_Traitement\_NaN\_En  
codage.ipynb

P7\_Step3a\_Traitement\_des\_Outli  
ers\_df\_train\_test.ipynb  
P7\_Step3b\_Traitement\_des\_Outli  
ers\_df\_valid.ipynb

P7\_Step4\_Split\_X\_y\_train\_test  
\_valid.ipynb



# PLAN

1. Rappel la problématique et du jeu de données (5min)
2. Explication de l'approche de modélisation (10min)
3. Présentation du dashboard (5 min)
4. Conclusion



## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### CHOIX DES MÉTRIQUES

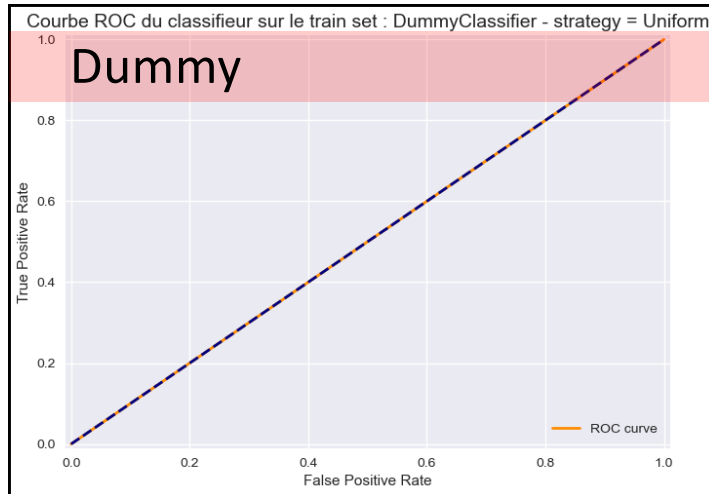


		Matrice de confusion	
		Classe 0 : prêt accordé – Classe 1 : prêt non accordé	
		Valeurs prédites	
		Prédit 0	Prédit 1
Vraies Valeurs	Vrai 0	Vrai Négatif Prêt accordé et client solvable	Faux positif Prêt non accordé mais client solvable « Dommage, on aurait pu accorder le prêt »
	Vrai 1	Faux négatif Prêt accordé mais client non solvable « On n'aurait jamais dû accorder le prêt ! »	Vrai positif Prêt non accordé et client non solvable

- On choisit une métrique qui permet de se rendre compte du compromis entre faux positifs et faux négatifs : on choisit donc le score ROC\_AUC qu'on pourra comparer au score de la compétition Kaggle.
- On définira également un coût métier pour optimiser le modèle.

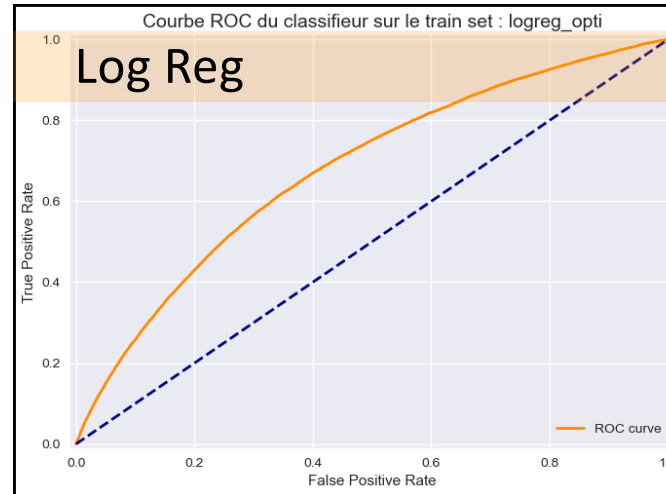
## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### EVALUATION DE DIFFERENTS MODELES



ROC\_AUC Dummy = **0.5**

-

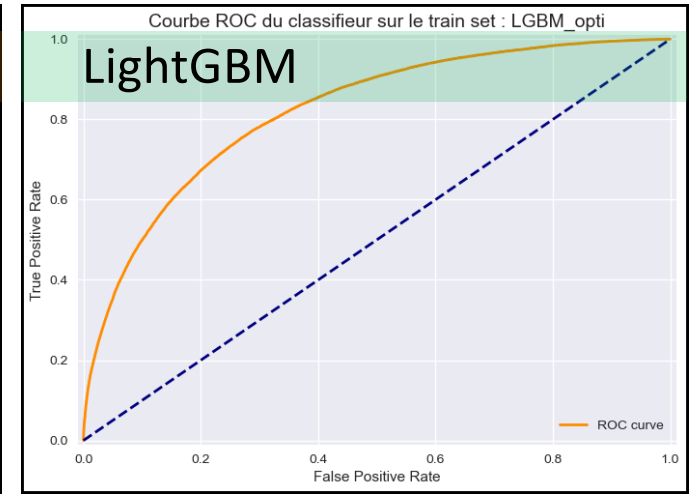


ROC\_AUC Log Reg = **0.6719**

C : 1

penalty = l2

solver : newton-cg



ROC\_AUC LightGBM = **0.7814**

learning\_rate: 0.1, max\_depth: 5,

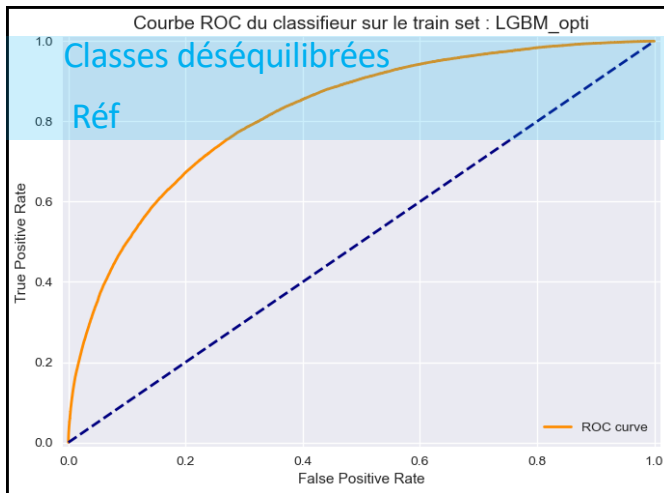
min\_data\_in\_leaf: 50, num\_iterations:

200, num\_leaves: 20

- Les modèles Log Reg et Light GBM ont été chacun optimisés avec Grid SearchCV.
- Score ROC\_AUC Kaggle de **0.7715/0.7754** (privé/public)
- Score ROC\_AUC sur le jeu de données non vu durant l'entraînement : **0.766**
- On choisira donc ce modèle LightGBM pour la suite de la modélisation.

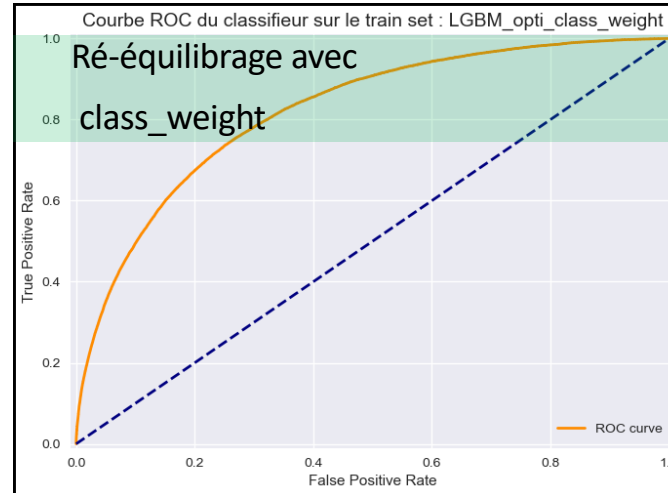
## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### PRISE EN COMPTE DU DÉSÉQUILIBRE DES CLASSES AVEC LIGHTGBM



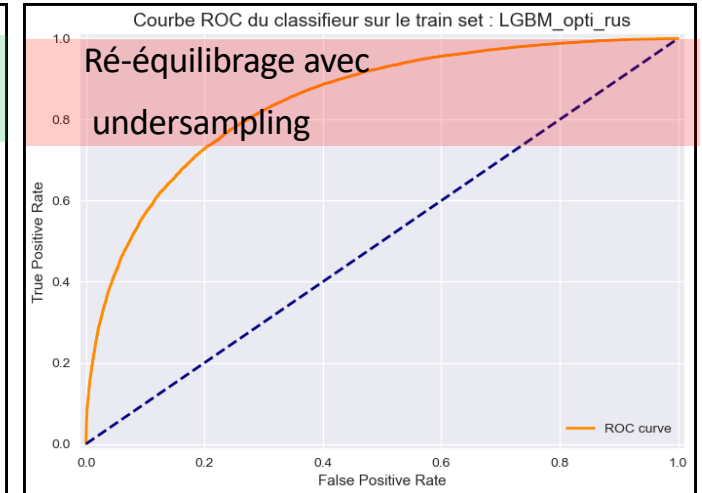
**ROC\_AUC Réf. = 0.7814**

learning\_rate: 0.1, max\_depth: 5,  
min\_data\_in\_leaf: 50, num\_iterations:  
200, num\_leaves: 20



**ROC\_AUC class\_weight = 0.7819**

Réf. +  
class\_weight={0: 0.35, 1: 0.65}



**ROC\_AUC undersampling = 0.7787**

Réf.

- 2 méthodes : utiliser le paramètre class\_weight de LightGBM (optimisation via GridSearchCV) ou faire un undersampling.
- La technique qui donne de meilleurs résultats est class\_weight : on conservera ce modèle.

## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### FONCTION COÛT MÉTIER



$$\text{coût métier} = FN + 10FP$$

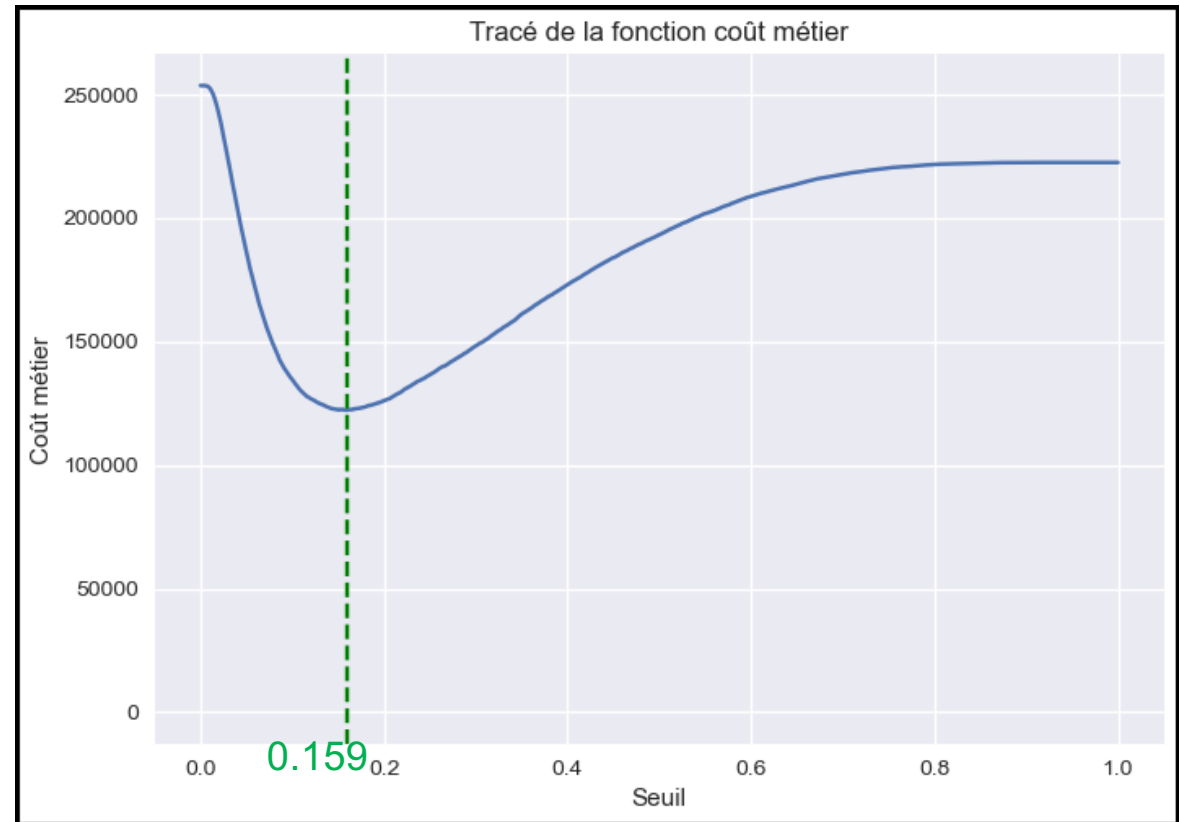
- FN : Nombre de Faux Négatifs, FP : Nombre de Faux Positifs
- On considère que la présence d'un faux négatif (personne non solvable qui s'est vu attribuée un prêt) est 10 fois plus coûteuse que la présence d'un faux positif (personne solvable qui s'est vu refuser un prêt).
- Le but va être de trouver le seuil de la probabilité que le client fasse défaut tel que ce seuil minimise le coût (le client fait défaut quand il appartient à la classe 1).

## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### DÉTERMINATION DU SEUIL POUR LE COUT METIER

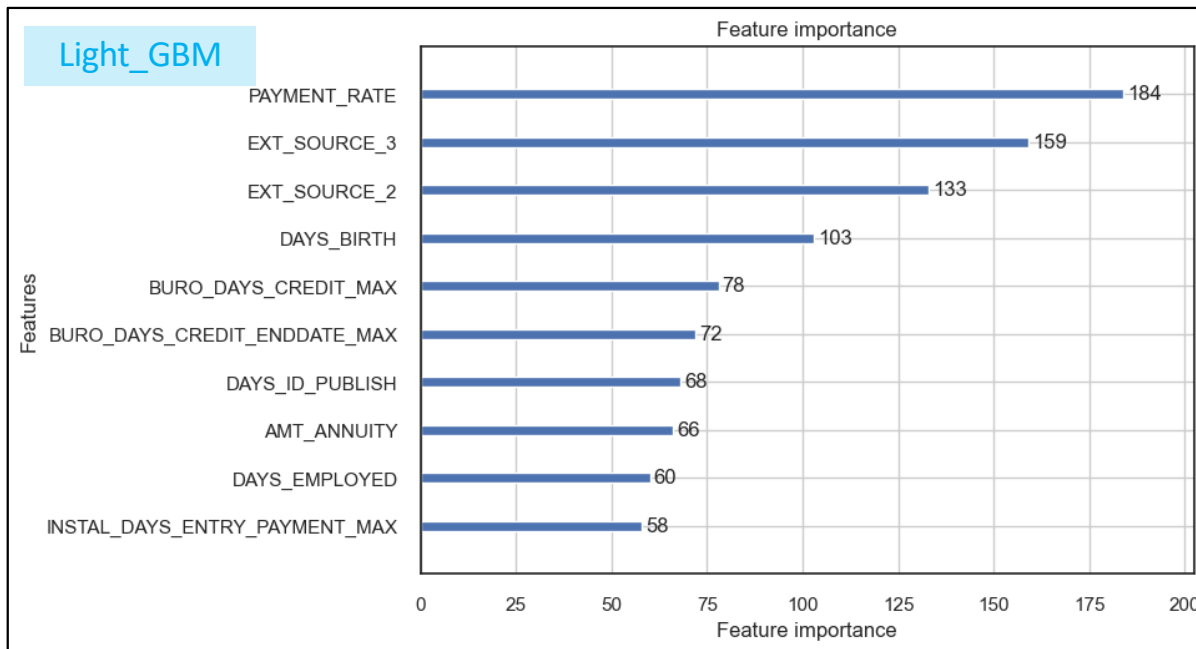


- On trace la courbe de coût métier pour différentes valeurs de seuils.
- Le minimum du coût métier est réalisé pour un seuil de 0.159.
- Si la probabilité de faire défaut est  $> 0.159$ , le prêt ne sera pas accordé.



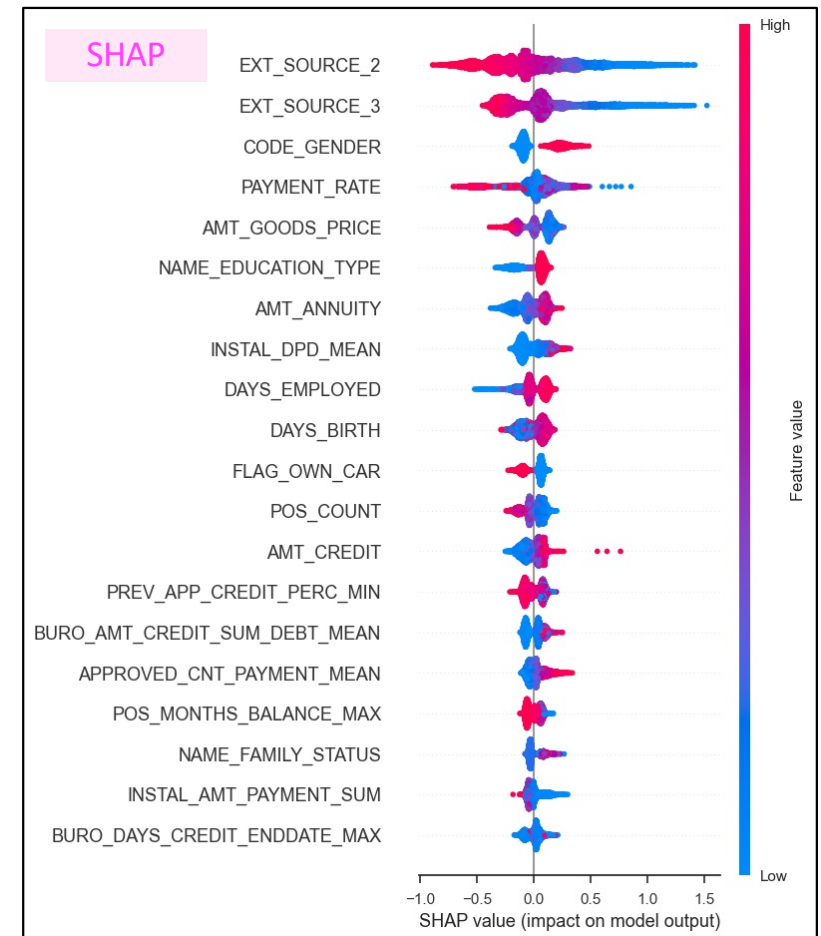
## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION

### INTERPRÉTABILITÉ GLOBALE DU MODÈLE

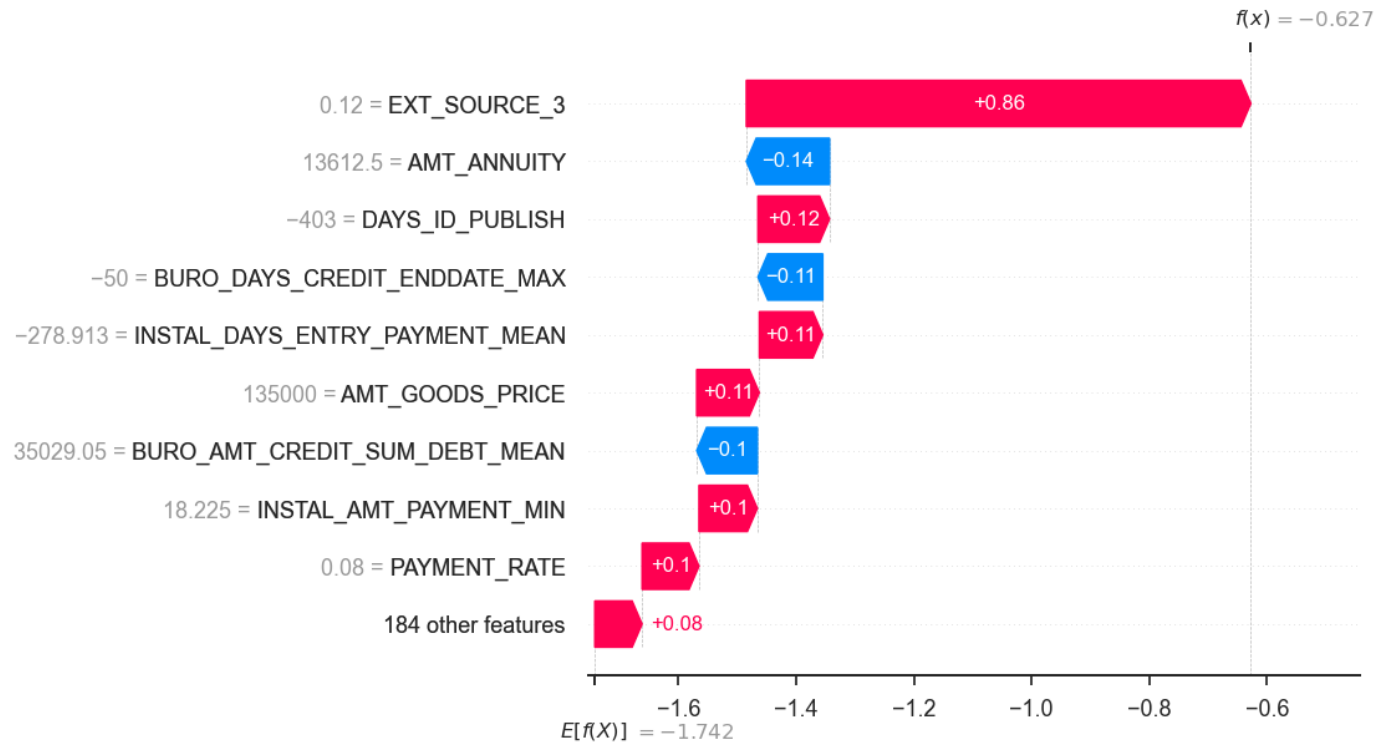


➤ Les paramètres les plus importants et communs aux 2 méthodes sont :

- **PAYMENT\_RATE**: taux de remboursement,
- **EXT\_SOURCE\_2, EXT\_SOURCE\_3**: scores issus d'autres institutions.



## 2. EXPLICATION DE L'APPROCHE DE MODÉLISATION INTERPRÉTABILITÉ LOCALE DU MODÈLE



- Pour ce client (7395), le prêt a été refusé.
- On retrouve le paramètre EXT\_SOURCE\_3 comme étant le plus déterminant pour ce client dans le refus de sa demande de prêt.








# PLAN

1. Rappel la problématique et du jeu de données (5min)
2. Explication de l'approche de modélisation (10min)
- 3. Présentation du dashboard (5 min)**
4. Conclusion



### 3. PRÉSENTATION DU DASHBOARD FRAMEWORKS UTILISÉS



	Framework	Pour faire quoi?	Lien
	MLFlow	Générer le code pour transformer le modèle LightGBM pour créer une API	<a href="https://github.com/mochan97/OpenClassrooms-Project-7-API-ML-Flow">https://github.com/mochan97/OpenClassrooms-Project-7-API-ML-Flow</a>
	Microsoft Azure Machine Learning Studio	Déployer un modèle de machine learning dans le cloud	URL du Endpoint de l'API (utilisé par le dashboard Streamlit): <a href="https://ocr-p7-api-mlflow-proba-qljnp.francecentral.inference.ml.azure.com/score">https://ocr-p7-api-mlflow-proba-qljnp.francecentral.inference.ml.azure.com/score</a>
	Streamlit	Créer le dashboard	<a href="https://github.com/mochan97/OpenClassrooms-Project-7-Dashboard">https://github.com/mochan97/OpenClassrooms-Project-7-Dashboard</a>
	Microsoft Azure Web App	Déployer le dashboard dans le cloud	<a href="https://streamlit-dashboard-p7.azurewebsites.net/">https://streamlit-dashboard-p7.azurewebsites.net/</a>
	Github	Versionner le code Déploiement CI/CD pour le dashboard	<a href="https://github.com/mochan97/OpenClassrooms-Project-7">https://github.com/mochan97/OpenClassrooms-Project-7</a>

# 3. PRÉSENTATION DU DASHBOARD

## PIPELINE DE DEPLOIEMENT CONTINU



Update README.md · mochan97

github.com/mochan97/OpenClassrooms-Project-7-Dashboard/actions/runs/4260892118

En pause Mettre à jour

Search or jump to... Pull requests Issues Codespaces Marketplace Explore

mochan97 / OpenClassrooms-Project-7-Dashboard Public

Pin Unwatch 1 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

← Build and deploy Python app to Azure Web App - Streamlit-dashboard-P7

Update README.md #22 Re-run all jobs

Summary

Jobs

- build
- deploy

Run details

- Usage
- Workflow file

Triggered via push 5 hours ago

mochan97 pushed 4245141 main

Status Success

Total duration 9m 19s

Artifacts 1

main\_streamlit-dashboard-p7.yml

on: push

build 1m 0s

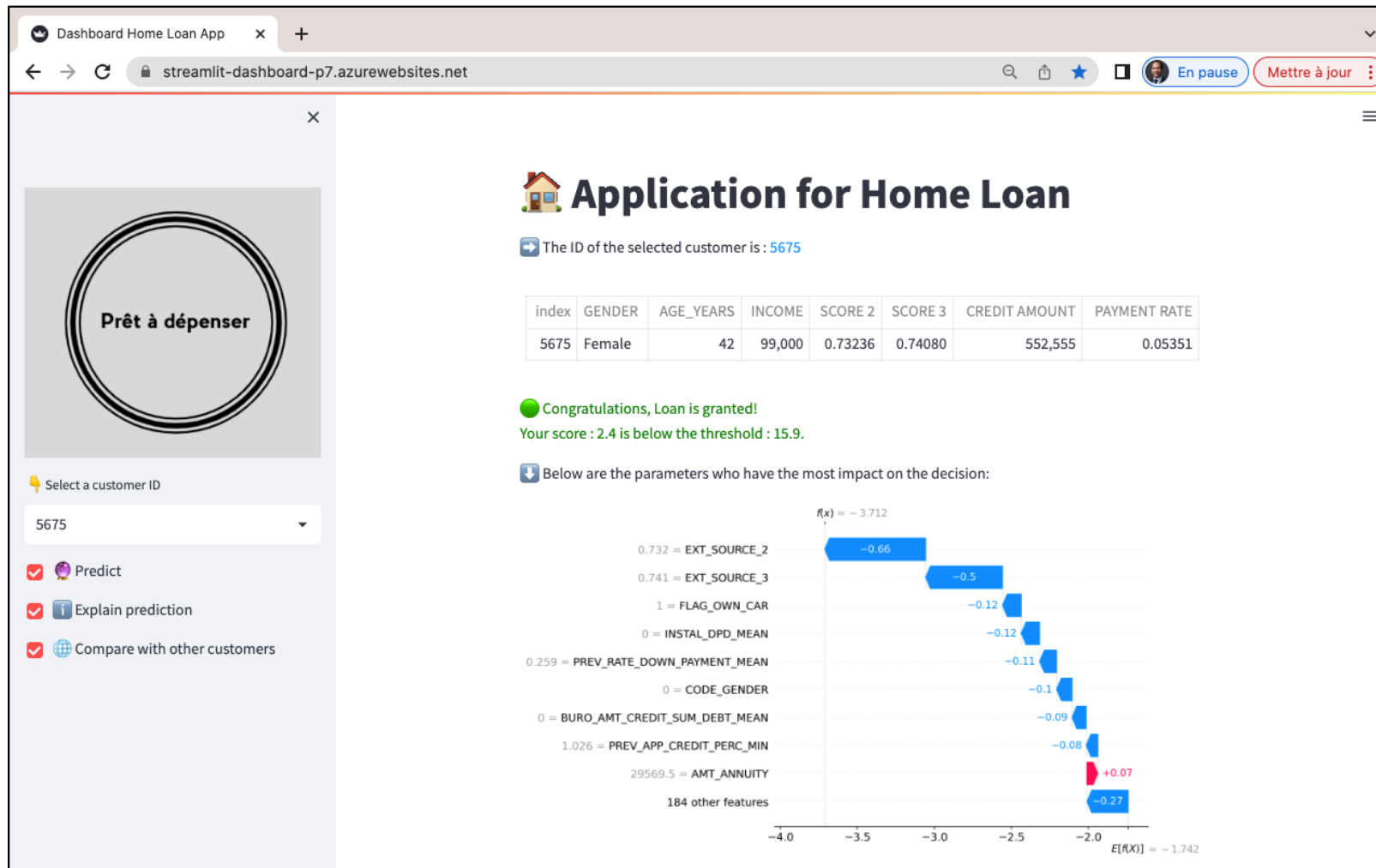
deploy 8m 1s

https://streamlit-dashboard-p7.azurew...



### 3. PRÉSENTATION DU DASHBOARD

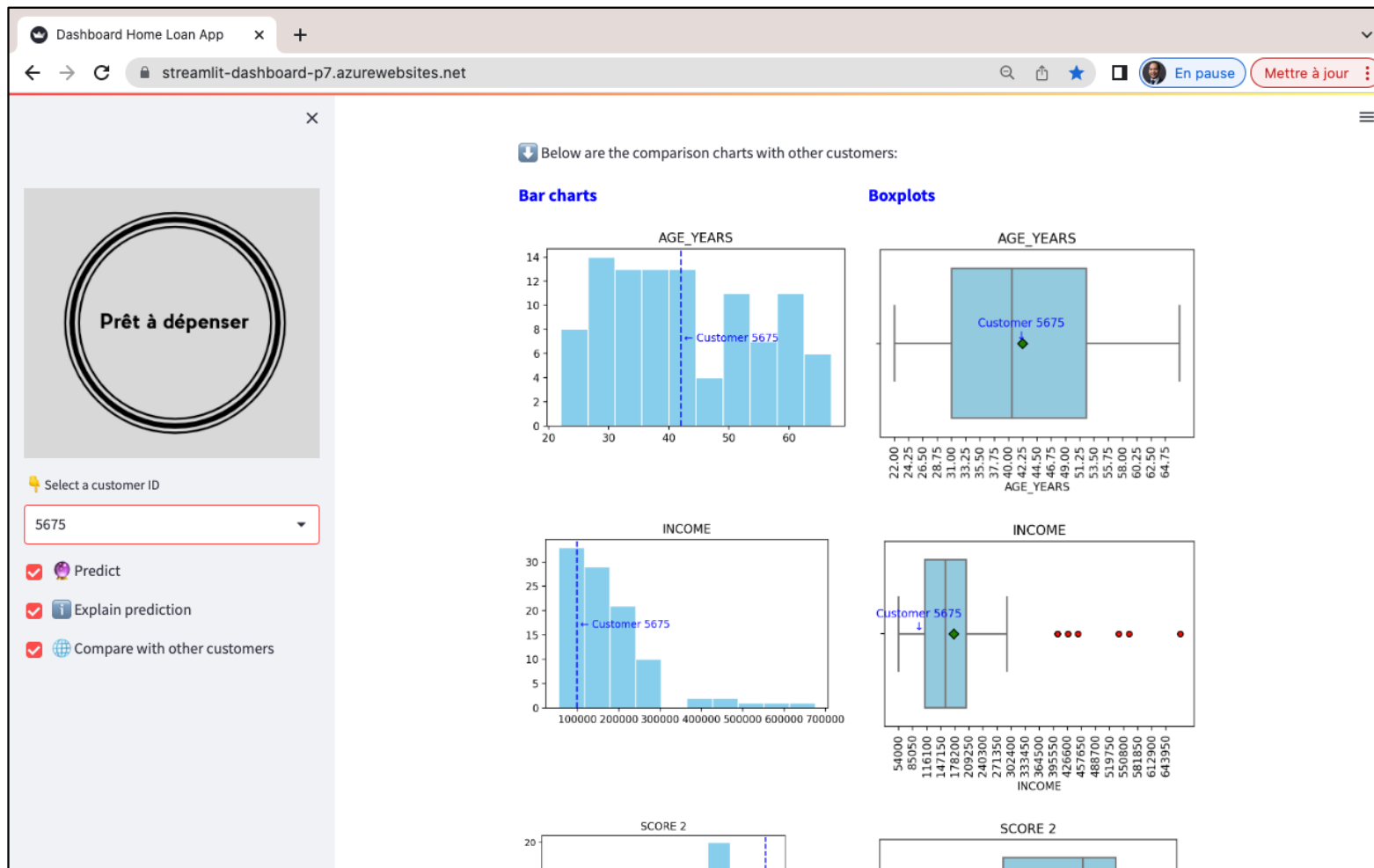
## VUE UTILISATEUR : PRÉDICTION ET FEATURES LOCALES



**NB:** Les valeurs SHAP ne sont pas calculées dans le dashboard mais chargées (le calcul a été fait dans le Notebook Jupyter). Pas de solution trouvée avec Azure pour déployer dans le cloud autrement.

### 3. PRÉSENTATION DU DASHBOARD

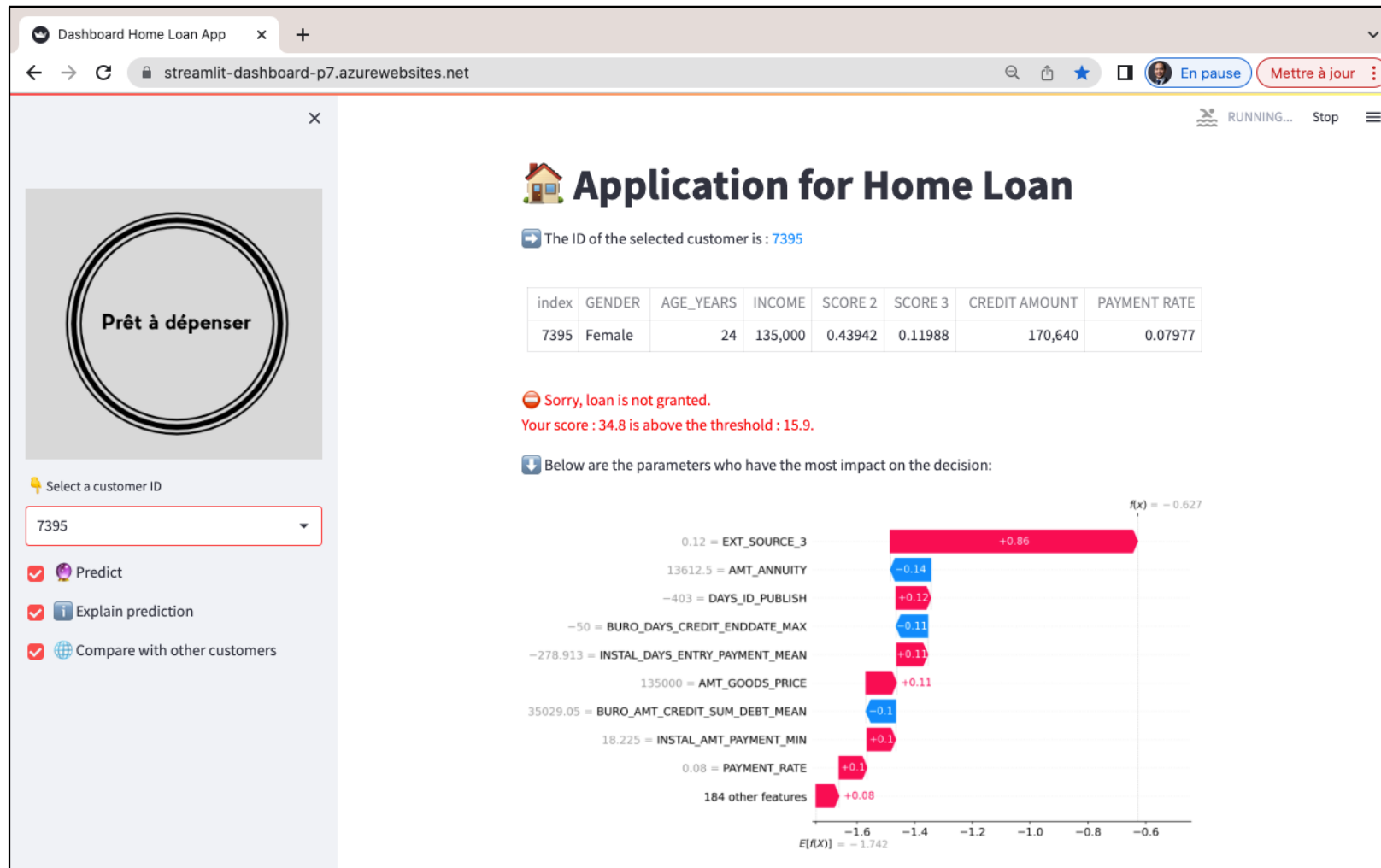
## VUE UTILISATEUR : PRÉDICTION ET COMPARAISON AUTRES CLIENTS





### 3. PRÉSENTATION DU DASHBOARD

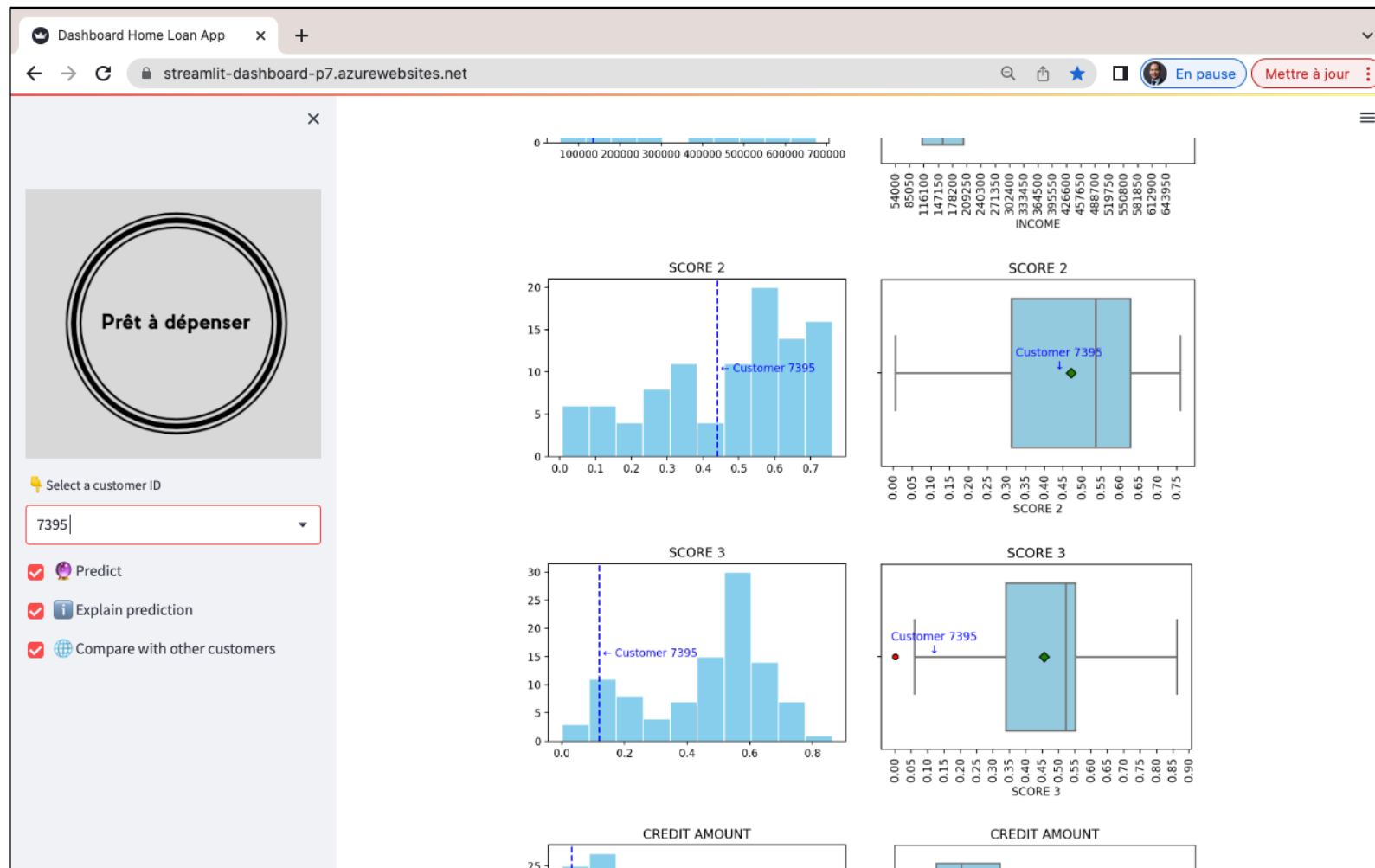
## VUE UTILISATEUR : PRÉDICTION ET FEATURES LOCALES



NB: Les valeurs SHAP ne sont pas calculées dans le dashboard mais chargées (le calcul a été fait dans le Notebook Jupyter). Pas de solution trouvée avec Azure pour déployer dans le cloud autrement.

### 3. PRÉSENTATION DU DASHBOARD

## VUE UTILISATEUR : PRÉDICTION ET COMPARAISON AUTRES CLIENTS





# PLAN

1. Rappel la problématique et du jeu de données (5min)
2. Explication de l'approche de modélisation (10min)
3. Présentation du dashboard (5 min)
4. Conclusion



## 4. CONCLUSION

- Un modèle de classification binaire a été mis en place avec une performance correcte.
- Le déséquilibre des classes a été pris en compte.
- Nous avons introduit une fonction coût métier pour minimiser les faux négatifs.

### Modèle : Light GBM

learning\_rate: 0.1, max\_depth: 5,  
min\_data\_in\_leaf: 50, num\_iterations: 200,  
num\_leaves: 20

class\_weight={0: 0.35, 1: 0.65}

ROC\_AUC = **0.7819**

ROC\_AUC Kaggle (privé/public) =  
**0.7715/0.7754**

- Ce modèle LightGBM a ensuite été déployé dans le cloud via une API et utilisable via l'interface du dashboard.

### Recommandations pour de futures améliorations:

- Affiner l'optimisation du modèle avec des balayages plus important et fins des hyperparamètres (limitation de ma machine).
- Tester d'autres modèles (RandomForrest Classifier, HistGradientBoostingClassifier)
- Recherche d'une fonction de coût plus complexe pour améliorer l'optimisation du coût.





**Merci de votre attention!**