

# Projet 7

## Implémentez un modèle de scoring

Zeynep Erdem  
27-10-2023





# Sommaire

- Rappel de la Problématique et Environnement
- Présentation et nettoyage des données
- Analyses Exploratoires des Données
- Présentation de démarche de modélisation et choix des métriques
- Présentation de la synthèse des résultats et Visualisation du tracking via MLFlow UI
- Interprétabilité globale et locale du modèle
- Présentation du pipeline de déploiement
- Présentation de l'analyse de data drift
- Conclusion et Recommandations
- Présentation et démo du dashboard

# Rappel de la Problématique

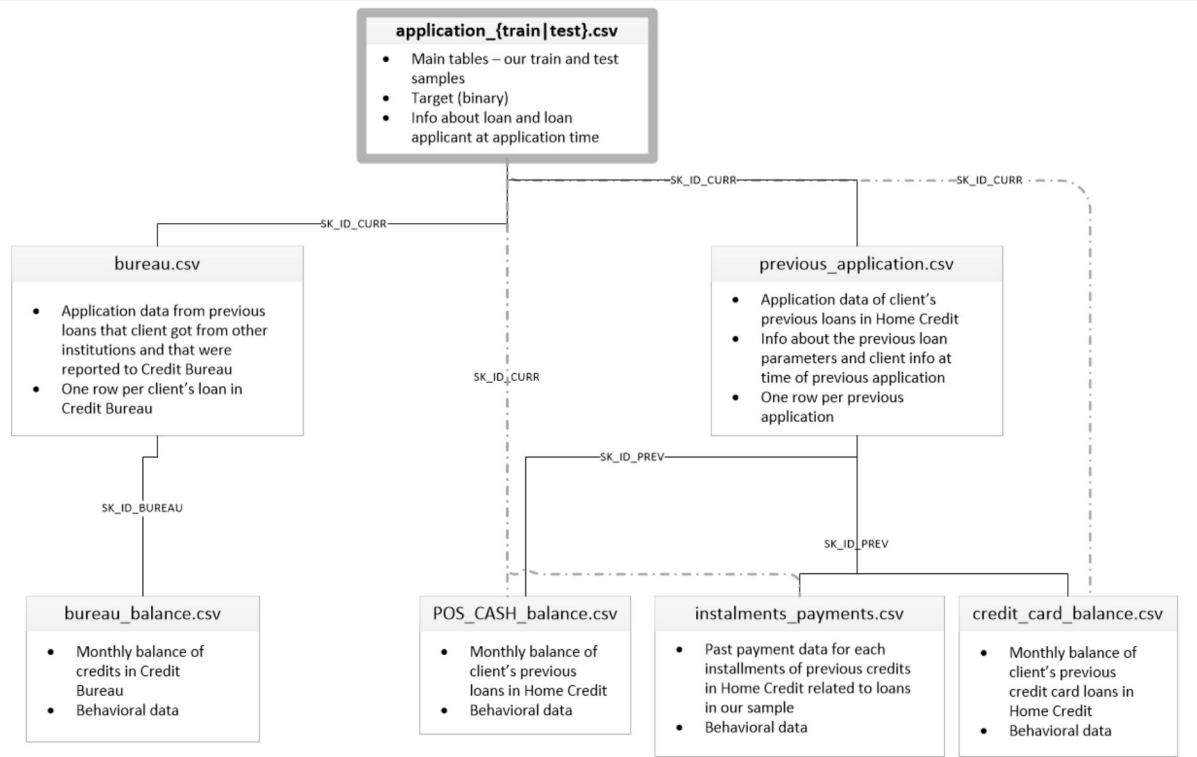


- ❖ "Prêt à dépenser", souhaite développer un outil de scoring crédit.
- ❖ Répondre à la demande de transparence de la part des clients.
- ❖ Construire un modèle de scoring pour prédire la probabilité de faillite d'un client.
- ❖ Construire un dashboard interactif permettant d'interpréter les prédictions faites par le modèle
- ❖ Mettre en production le modèle prédiction à l'aide d'une API, ainsi que le dashboard.

## Environnement

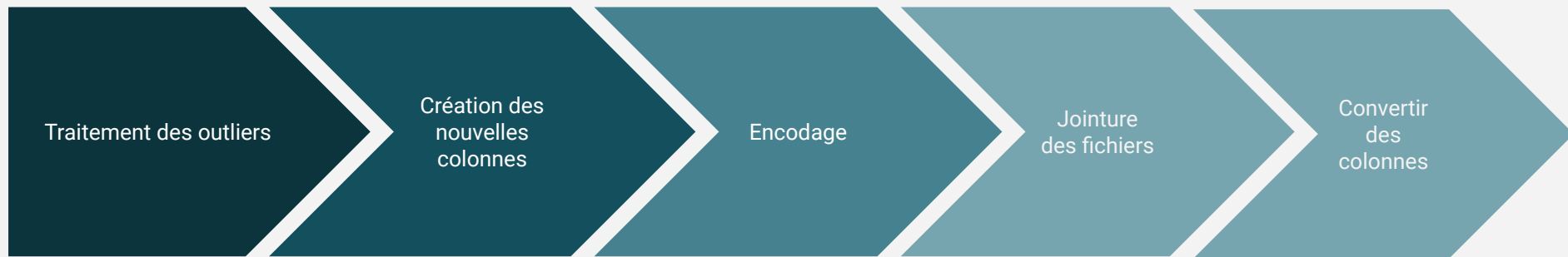
- Python: 3.9.17
- Pandas: 2.0.3
- Numpy: 1.23.5
- Seaborn: 0.12.2
- Matplotlib: 3.7.2
- Missingno: 0.4.2
- Sklearn: 1.2.2
- Mlflow: 2.6.0
- Shap: 0.42.1
- Plotly: 5.15.0

# Présentation du Jeu de Données



- Les données proviennent de la compétition Kaggle "Home Credit Default Risk"
- Il y a 10 fichiers avec 346 colonnes.
- Ils sont liés par des clés.
- Il y a 25 % de valeur manquantes.

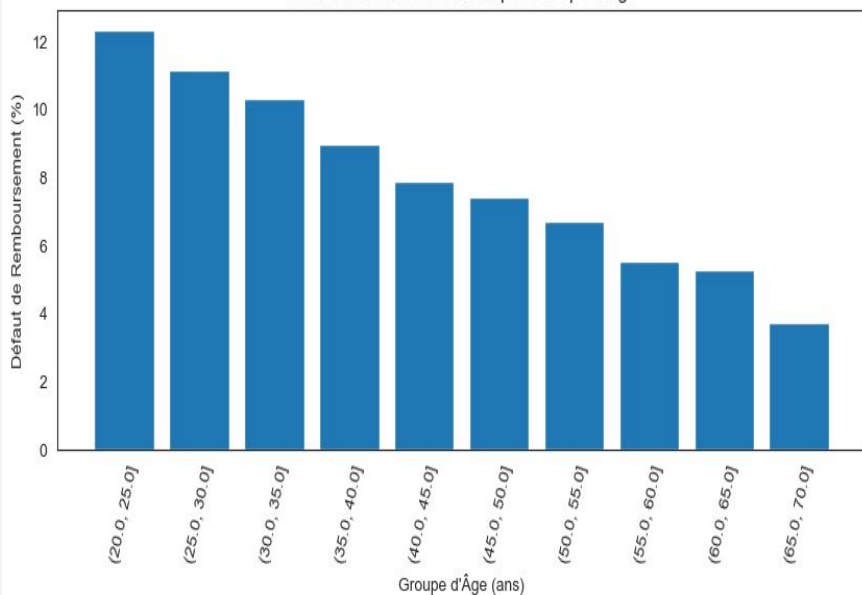
# Les étapes du nettoyage



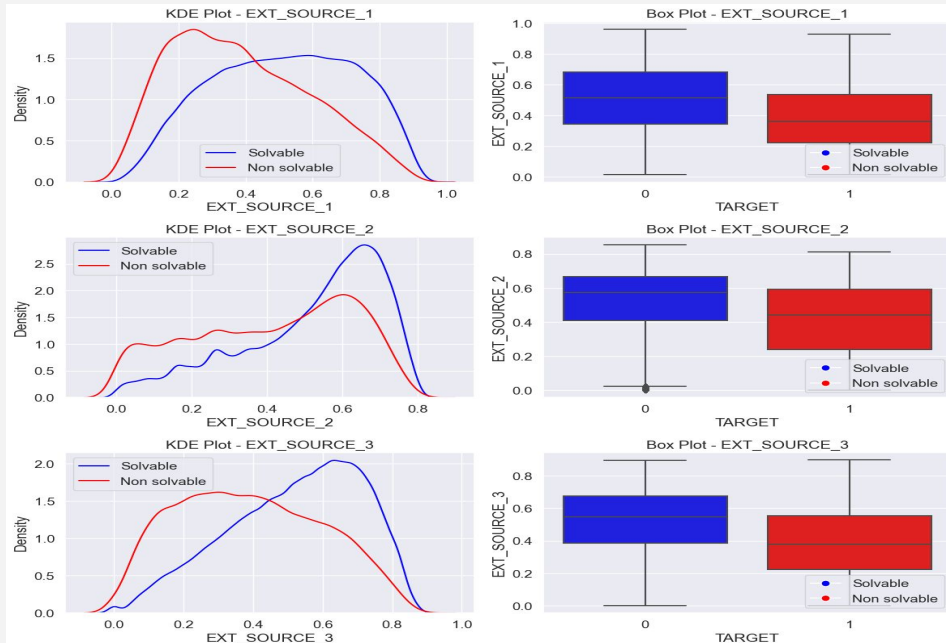
- Elimination de XNA de la colonne CODE\_GENDER
- Remplacement des valeurs 365243, par np.nan dans la colonne DAYS\_EMPLOYED (1000 ans de travail)
- Calculs arithmétiques  $\text{PAYMENT\_RATE} = \text{AMT\_ANNUITY} / \text{AMT\_CREDIT}$
- Aggregations ['max', 'mean', 'sum']
- Label encoding pour binary colonnes
- One hot encoding pour les autres colonnes catégorielles
- 10 fichiers avec 346 colonnes
- 1 fichier avec 797 colonnes
- Changement des colonnes DAYS\_BIRTH et DAYS\_EMPLOYED aux valeurs positives et en ans

# Analyses Exploratoires des Données

Défaut de Remboursement par Groupe d'Âge

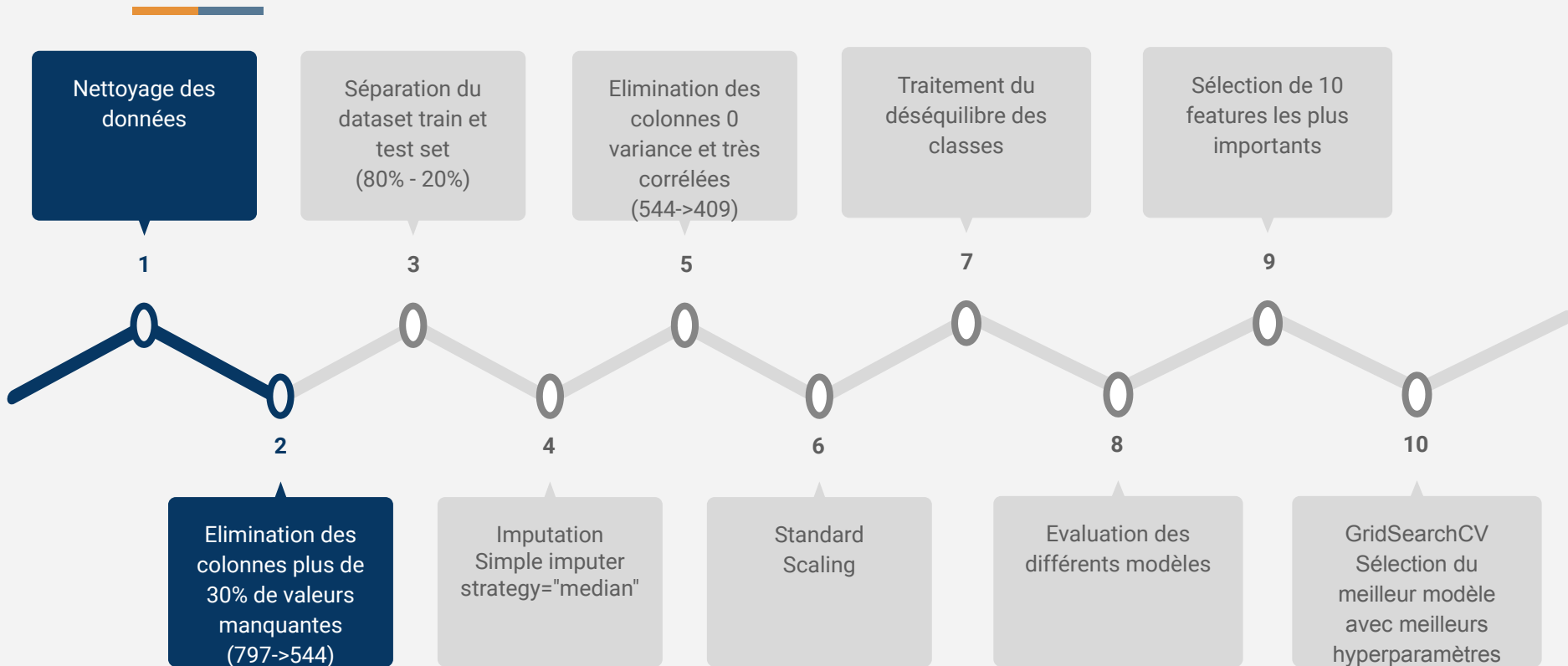


Le taux de défaut de remboursement est plus élevé chez les clients de moins de 30 ans.



Les clients solvables ont un score EXT\_SOURCE plus élevé que les non solvables.

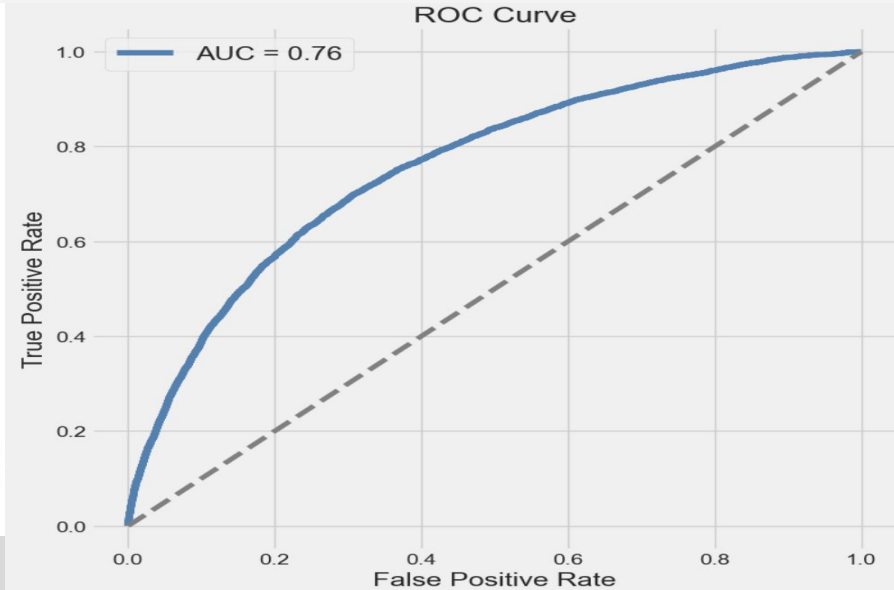
# La démarche de modélisation



# Les métriques d'évaluation et Fonction coût métier

		Predicted		
		Negative (0)	Positive (1)	
Actual	Negative (0)	True Negative <b>TN</b>	False Positive <b>FP</b> (Type I error)	<b>Specificity</b> $= \frac{TN}{TN + FP}$
	Positive (1)	False Negative <b>FN</b> (Type II error)	True Positive <b>TP</b>	<b>Recall, Sensitivity, True positive rate (TPR)</b> $= \frac{TP}{TP + FN}$
		<b>Accuracy</b> $= \frac{TP + TN}{TP + TN + FP + FN}$		<b>Precision, Positive predictive value (PPV)</b> $= \frac{TP}{TP + FP}$
				<b>F1-score</b> $= 2 \times \frac{Recall \times Precision}{Recall + Precision}$

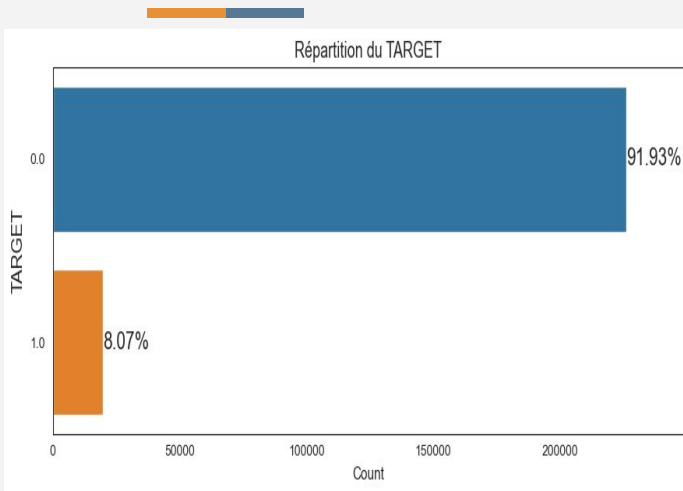
- FN - mauvais client, prédit comme bon client : donc crédit accordé et perte en capital / -10
- FP - bon client, prédit comme mauvais client : donc refus crédit et manque à gagner en marge/ -1
- TN - mauvais client, prédit comme mauvais client : donc refus crédit et pas perte/ 0
- TP - bon client, prédit comme bon client : donc crédit accordé et pas perte/ 0



```
total = (coeff_tn*tn + coeff_fp*fp + coeff_fn*fn + coeff_tp*tp)
max_gain = (tn + fp)*coeff_tn + (tp + fn)*coeff_tp
min_gain = (tn + fp)*coeff_fp + (tp + fn)*coeff_fn
gain = (total - min_gain) / (max_gain - min_gain)
```



# Traitement du déséquilibre des classes

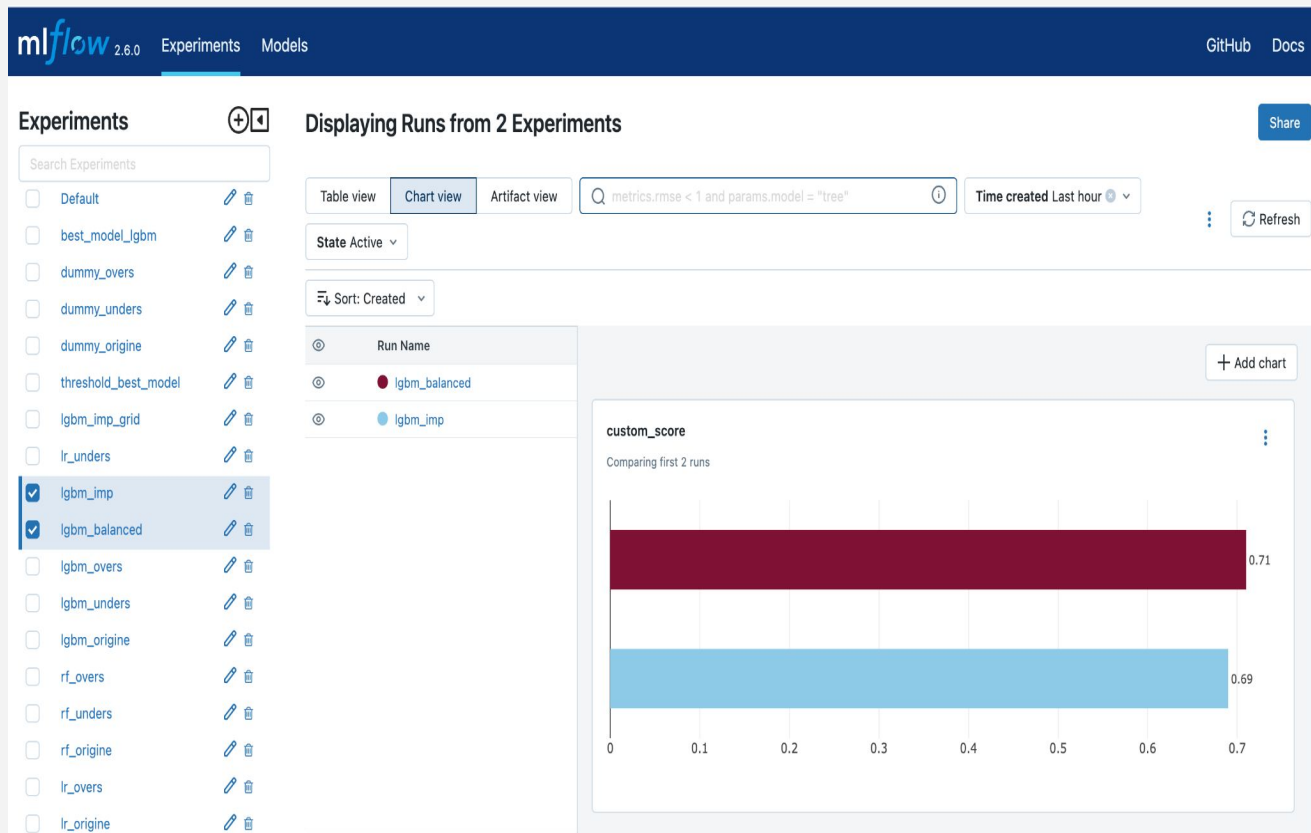


- On constate le changement considérable de l'accuracy.
- On obtient les meilleurs résultats avec la méthode class weight balanced.

- RandomUnderSampler - sous-échantillonnage
  - supprime aléatoirement des échantillons de la classe majoritaire
- SMOTE - Suréchantillonnage
  - augmente la taille de la classe minoritaire
- Model(class\_weight="balanced")
  - attribue des poids plus importants aux classes minoritaires pendant l'entraînement

model	accuracy	precision	recall	f1_score	roc_auc_score	custom_score	execution_time
lgbm_origine	0.92	0.55	0.03	0.06	0.78	0.55	14.535
lgbm_unders	0.70	0.17	0.70	0.28	0.77	0.70	5.846
lgbm_overs	0.92	0.52	0.03	0.05	0.77	0.54	28.776
lgbm_balanced	0.72	0.18	0.69	0.29	0.78	0.71	14.838

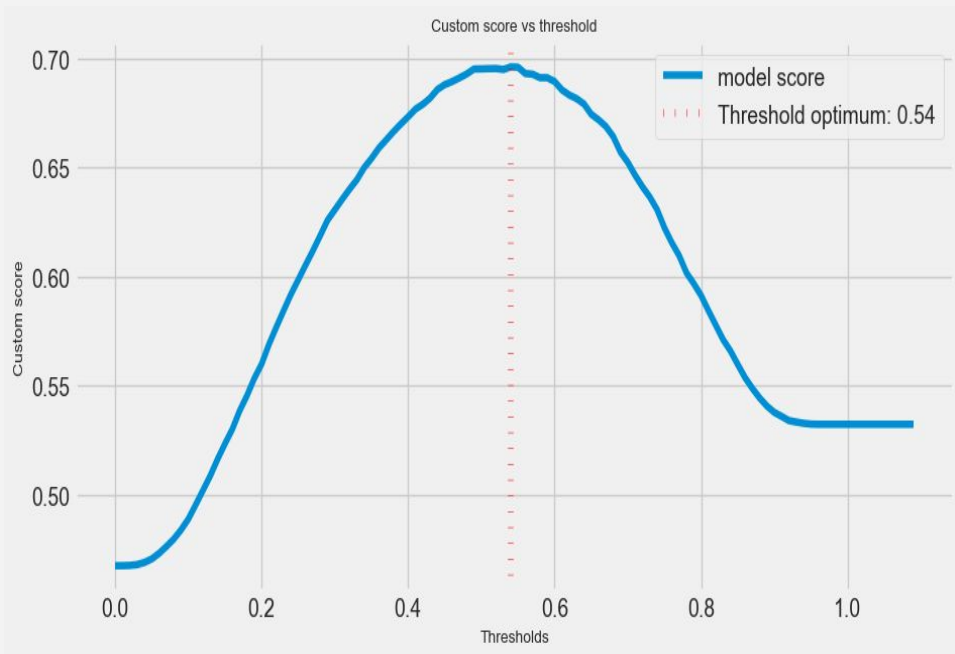
# Sélection de 10 features les plus importants



	Avant	Après
custom score	0.71	0.69
roc auc	0.78	0.76
time (s)	14.7	1.87

- Sélection des 10 features les plus importantes
- Légère perdre de performance
- Gagne en termes de durée d'exécution et l'interprétabilité du modèle

# L'algorithme d'optimisation et le seuil optimum pour le score métier.



- On a choisi LightGBM et effectué un GridSearchCV afin d'optimiser les hyperparamètres
- learning\_rate: 0.05  
n\_estimators: 350
- Le seuil optimal: 0.54  
score métier de: 0.7

# Présentation de la synthèse des résultats

Run Name	Created	Duration	Metrics					
			accuracy	custom_score	f1_score	precision	recall	rocauc
threshold_best_model	✓ 11 minutes ago	6.6s	0.74	0.7	0.28	0.18	0.63	0.76
best_model_lgbm	✓ 30 minutes ago	6.9s	0.7	0.7	0.27	0.17	0.68	0.76
lgbm_balanced	✓ 42 minutes ago	34.2s	0.72	0.71	0.29	0.18	0.69	0.78
lgbm_overs	✓ 44 minutes ago	1.0min	0.92	0.54	0.05	0.52	0.03	0.77
lgbm_unders	✓ 45 minutes ago	15.8s	0.7	0.7	0.28	0.17	0.7	0.77
lgbm_origine	✓ 49 minutes ago	34.3s	0.92	0.55	0.06	0.55	0.03	0.78
rf_balanced	✓ 57 minutes ago	4.3min	0.92	0.53	0	0.59	0	0.73
rf_overs	✓ 1 hour ago	10.7min	0.92	0.54	0.05	0.35	0.03	0.72
rf_unders	✓ 1 hour ago	46.2s	0.69	0.69	0.26	0.16	0.68	0.75
rf_origine	✓ 1 hour ago	5.7min	0.92	0.53	0	0.78	0	0.72
lr_balanced	✓ 1 hour ago	37.6s	0.7	0.7	0.27	0.17	0.69	0.76
lr_overs	✓ 1 hour ago	1.2min	0.71	0.68	0.26	0.17	0.65	0.75
lr_unders	✓ 1 hour ago	12.5s	0.69	0.69	0.27	0.17	0.69	0.76

Après  
sélection

Avant  
sélection

## Les modèles

Dummy Classifier

Régression Logistique

Random Forest

Light GBM

## Equilibrage

RandomUnderSampler

SMOTE

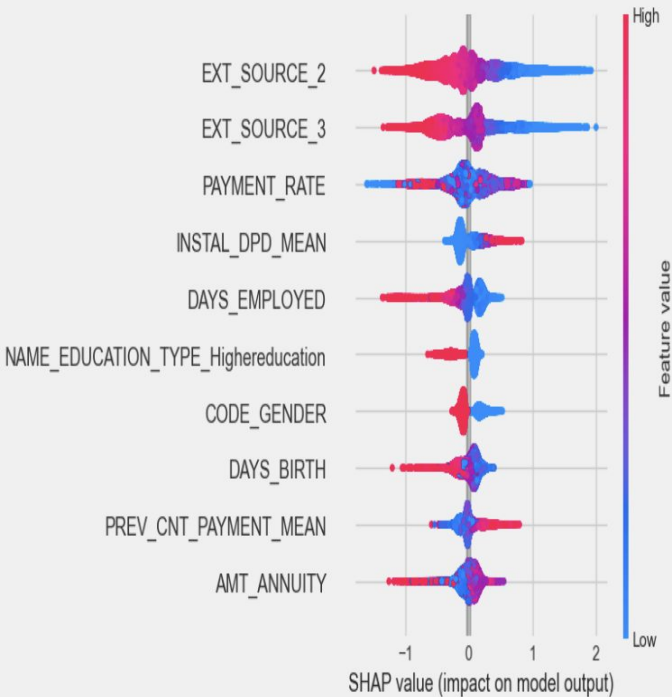
class\_weight="balanced"

## Feature sélection

GridSearchCV

Light GBM

# Interprétabilité globale du modèle

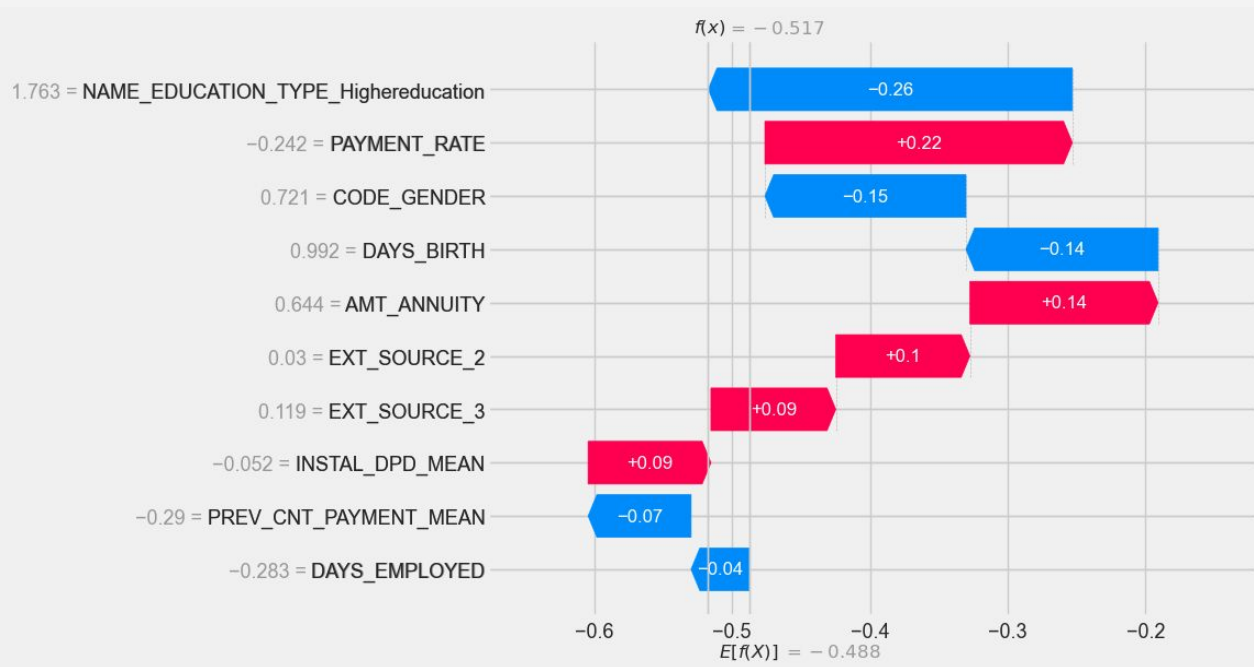


Feature	Explication
EXT_SOURCE_2 et 3	Score normalisé provenant d'une source de données externe
PAYMENT_RATE	Taux de paiement
INSTAL_DPD_MEAN	Nombre de jours de retard de paiement pour le crédit précédent (moyenne)
DAYS_EMPLOYED	Durée du travail (ans)
NAME_EDUCATION_TYPE_Highereducation	Niveau d'études le plus élevé (éducation supérieure)
CODE_GENDER	Genre female - 1 male - 0
PREV_CNT_PAYMENT_MEAN	Durée du crédit précédent (moyenne)
DAYS_BIRTH	Age (ans)
AMT_ANNUITY	Rente de prêt annuelle

- Faible valeur EXT\_SOURCE\_2 et 3 est associée à un risque accru de défaut
- Valeur élevée INSTAL\_DPD\_MEAN augment le risque de défaut

← risque baisse    → risque augment

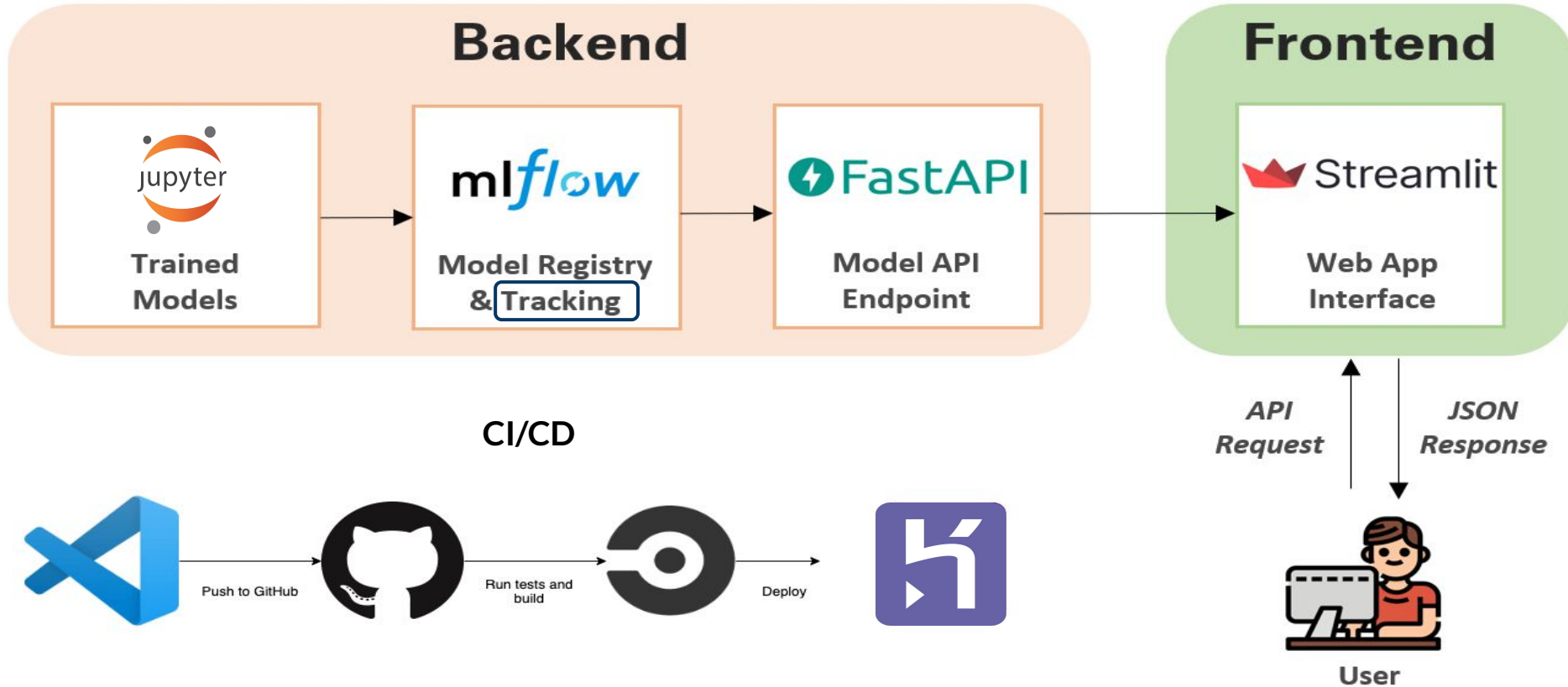
# Interprétabilité locale du modèle



- Client l'ID 343913  
NAME\_EDUCATION\_TYPE\_Highereducation et  
DAYS\_BIRTH ont réduit le  
risque de défaut
- Les features  
PAYMENT\_RATE et  
AMT\_ANNUITY ont  
augmenté le risque de défaut

← impact négatif    → impact positif

# Présentation du pipeline de déploiement

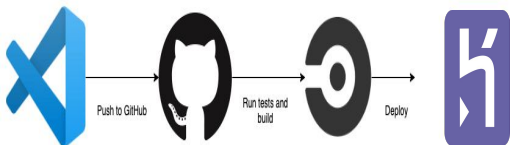


# Présentation du pipeline de déploiement

```
(p7) mbp-de-zeynep:tests zeynepdem$ pytest
===== test session starts =====
platform darwin -- Python 3.9.17, pytest-7.4.2, pluggy-1.0.0
rootdir: /Users/zeynepdem/Desktop/ds_prep/p7/tests
plugins: anyio-3.7.1
collected 3 items

test_api.py ...

===== 3 passed in 0.11s =====
(p7) mbp-de-zeynep:tests zeynepdem$
```



```
build
succeeded now in 1m 16s

> Set up job
> Run actions/checkout@v3
> Set up Python 3.9
> Install dependencies
< Test with pytest
  1 ▶ Run pytest
  7
  7 ===== test session starts =====
  8 platform linux -- Python 3.9.18, pytest-7.4.2, pluggy-1.3.0
  9 rootdir: /home/runner/work/p7_Home_Credit/p7_Home_Credit
 10 plugins: anyio-3.7.1
 11 collected 3 items
 12
 13 tests/test_api.py ...
 14
 15 ===== 3 passed in 1.74s =====

> Post Set up Python 3.9
> Post Run actions/checkout@v3
> Complete job
```

zeynepdemfr@gmail.com: Deployed 255dceeb  
Oct 6 at 10:09 PM · v7 · [Compare diff](#)

zeynepdemfr@gmail.com: Build succeeded  
Oct 6 at 10:07 PM · [View build log](#)

zeynepdemfr@gmail.com: Deployed e14ee38d  
Oct 6 at 6:42 PM · v6 · [Compare diff](#)

zeynepdemfr@gmail.com: Build succeeded  
Oct 6 at 6:40 PM · [View build log](#)



# Présentation de l'analyse de data drift

Dataset Drift

Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5

121  
Columns











9  
Drifted Columns

0.0744  
Share of Drifted Columns

Data Drift Summary

Drift is detected for 7.438% of columns (9 out of 121).

Q Search

Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
> AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.359052
> AMT_REQ_CREDIT_BUREAU_MON	num			Detected	Wasserstein distance (normed)	0.281765
> AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.210785
> AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.207334
> AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.161102

- La qualité et de la distribution des données au fil du temps
- Il nécessite une surveillance continue pour maintenir la précision du modèle.
- On a testé la librairie evidently
- On constate une légère data drift pour 9 colonnes

## Conclusion et Recommandations



- On a construit un modèle de scoring pour prédire la probabilité de faillite d'un client.
  - On a construit un dashboard interactif permettant d'interpréter les prédictions faites par le modèle.
  - On a mis en production le modèle prédiction à l'aide d'une API, ainsi que le dashboard.
- 
- Il serait préférable de collaborer avec les équipes métier pour créer une métrique plus adaptée à leurs besoins spécifiques.
  - Nous pouvons améliorer la sélection des features les plus importantes en explorant d'autres méthodes, avec les experts métier pour répondre aux exigences du domaine et augmenter l'interprétabilité du modèle.
  - On peut améliorer nos résultats en faisant des hyperparamètres tuning plus fin pour les modèles.

# Présentation et démo du dashboard




## Liens

**Le dossier Github :** [https://github.com/githubzey/p7\\_Home\\_Credit](https://github.com/githubzey/p7_Home_Credit)

**Api :** <https://apihomecredit-861d00eaed91.herokuapp.com/>

**Dashboard :** <https://dashboardhomecredit-1913c1e69feb.herokuapp.com/>



# Présentation et démo du dashboard


→ ↺ apihomecredit-861d00eae91.herokuapp.com/docs#/ 


## FastAPI 0.1.0 OAS 3.1


/openapi.json

### default

GET / Home  


POST /prediction\_manual Prediction Manual 

POST /predict Predict Credit 

POST /shaplocal/ Shap Values Local 

**Parameters** Cancel Reset

No parameters

**Request body** required application/json 

```
{  "data": "  [{"EXT_SOURCE_2\\":0.5202729458,\\\"EXT_SOURCE_3\\\":null,\\\"CODE_GENDER\\\":1,\\\"PAYMENT_RATE\\\":0.0482779168,\\\"DAYS_EMPLOYED\\\":null,\\\"INSTAL_DPD_MEAN\\\":0.5333333333,\\\"PREV_CNT_PAYMENT_MEAN\\\":12.0,\\\"NAME_EDUCATION_TYPE_Highereducation\\\":1,\\\"DAYS_BIRTH\\\":55.8,\\\"AMT_ANNUITY\\\":36459.0}]"}
```


**Server response**

**Code** **Details**

200

**Response body**

```
{  "proba": 0.37}
```

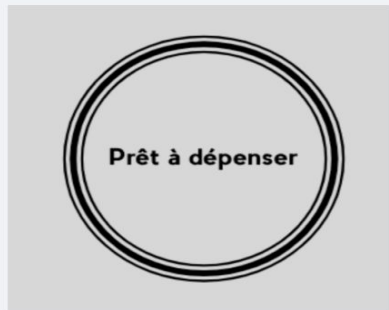
 Download

# Présentation et démo du dashboard

← → ↺ dashboardhomecredit-1913c1e69feb.herokuapp.com

🔖 ☆ ⚙️ 🗖️ 🌐

×



## Menu

Choisissez votre page.

- ☐ Home
- ☒ Information client
- ☐ Décision et explication
- ☐ Comparaison

## Décision Prêt à dépenser

Sélectionnez le numéro du client

343913

▼

### Les informations du client

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_
669	343,913	0	Cash loans	F	N	Y	0	112,500	·

### Informations de client :

ID client : 343913

Genre : F

Age : 55.8

Type d'éducation : Higher education

Statut familial : Married

Nombre d'enfant : 0

Type de contrat : Cash loans

Type de revenu : Pensioner

Revenu total : 112500.0

Prêt total : 755190.0

Durée travail(année) : None

Propriétaire maison/appartement : Y

Type de logement : House / apartment

# Présentation et démo du dashboard

← → ↺ dashboardhomecredit-1913c1e69feb.herokuapp.com

🔖 ☆ ⚙️ 🗖️ 🌐

×



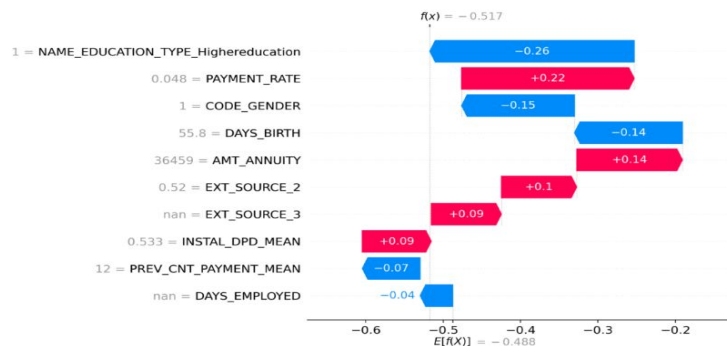
## Menu

Choisissez votre page.

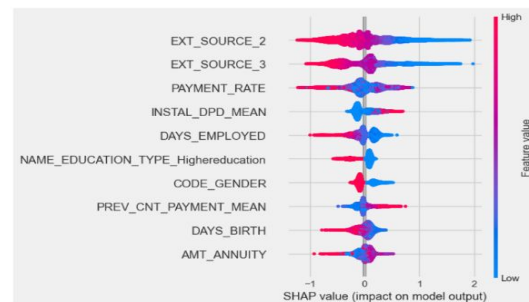
- ☐ Home
- ☐ Information client
- ☒ Décision et explication
- ☐ Comparaison

La probabilité de faillite du client : 37%

✓ Crédit accepté



Voir explication de graphique



Voir explication de graphique et des variables

# Présentation et démo du dashboard

