

# **Projet n°7**

Implémentation d'un modèle de scoring

**PRÊT À DÉPENSER**

**OPENCLASSROOMS**

# Sommaire

- Introduction
- Analyse et préparation des données
- Modélisation
  - Simulations
  - Modèle final (MlFlow)
- Features importance
  - Importance globale
  - Importance locale
- Optimisation
  - Score de performance
  - Optimisation du seuil de décision
- Data Drift
- Dashboard
- Conclusion

# INTRODUCTION

# Missions et objectifs

L'objectif de la société **Prêt à dépenser** est de mettre en place un outil de Scoring afin de déterminer la probabilité de faillite d'un client dans le cadre d'un remboursement de crédit.

Les missions ici sont :

- L'élaboration d'un modèle de classification binaire
- Analyse des features contribuant le plus au modèle
- Mise en production du modèle via une API

# **ANALYSE ET PRÉPARATION DES DONNÉES**

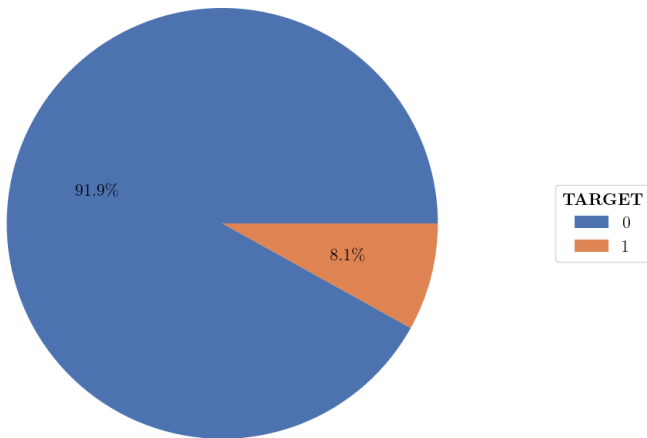
# Description des données

- Deux jeux :  $\begin{cases} \text{application\_train} \rightarrow (307511 \times 122) \\ \text{application\_test} \rightarrow (48744 \times 121) \end{cases}$
- Chaque client est identifié grâce à son `SK_ID_CURR`
- Aucun doublon détecté dans les jeux de données

Variables numériques	104
Variables catégorielles	16

# Répartition des clients par rapport à la cible

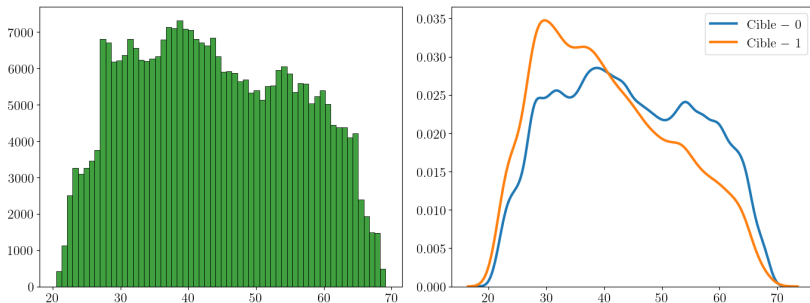
Proportions de la cible



# Distributions

DAYS BIRTH

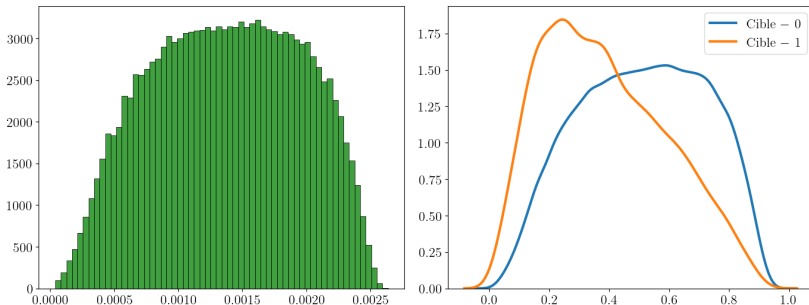
Distribution de DAYS BIRTH – (years)





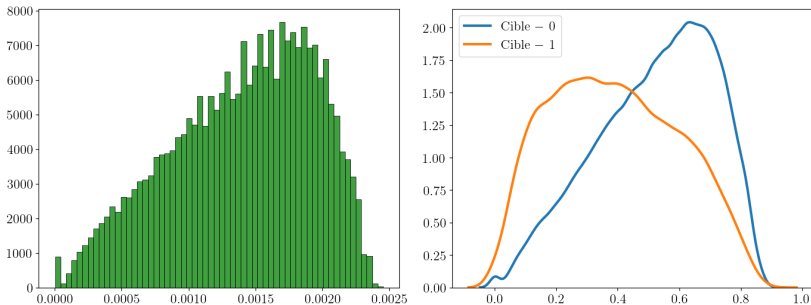
## EXT SOURCE 1

Distribution de EXT SOURCE 1



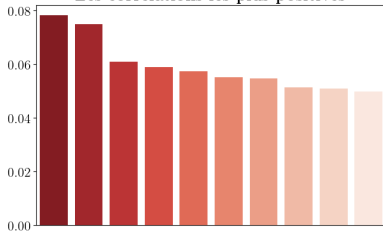
## EXT SOURCE 3

Distribution de EXT SOURCE 3



# Corrélations

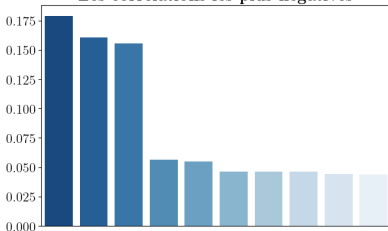
## Les corrélations les plus positives



### Variables

- DAYS.BIRTH
- DAYS.EMPLOYED
- REGION.RATING.CLIENT.W\_CITY
- REGION.RATING.CLIENT
- NAME.INCOME.TYPE.Working
- DAYS.LAST\_PHONE.CHANGE
- CODE.GENDER.M
- DAYS.ID.PUBLISH
- REG.CITY.NOT\_WORK.CITY
- NAME.EDUCATION.TYPE.Secondary / secondary special

## Les corrélations les plus négatives



### Variables

- EXT.SOURCE.3
- EXT.SOURCE.2
- EXT.SOURCE.1
- NAME.EDUCATION.TYPE.Higher education
- CODE.GENDER.F
- NAME.INCOME.TYPE.Pensioner
- DAYS.EMPLOYED.ANOMALY
- ORGANIZATION.TYPE.XNA
- FLOORSMAX.AVG
- FLOORSMAX.MEDI

# Encodage & Imputation

- Encodage des variables catégorielles avec
  - `LabelEncoder` pour les variables à deux valeurs uniques
  - `OneHotEncoder` pour les autres
- Imputation à l'aide d'un `SimpleImputer`

Type de colonnes	Nombre de colonnes
Booléenne	131
Float	65
Int	44

# Feature Engineering

◆ DAYS\_EMPLOYED\_PERCENT

→

$$\frac{\text{DAYS\_EMPLOYED}}{\text{DAYS\_BIRTH}}$$

◆ INCOME\_CREDIT\_PERCENT

→

$$\frac{\text{AMT\_INCOME\_TOTAL}}{\text{AMT\_CREDIT}}$$

◆ INCOME\_PER\_PERSON

→

$$\frac{\text{AMT\_INCOME\_TOTAL}}{\text{CNT\_FAM\_MEMBERS}}$$

◆ ANNUITY\_INCOME\_PERCENT

→

$$\frac{\text{AMT\_ANNUITY}}{\text{AMT\_INCOME\_TOTAL}}$$

◆ PAYMENT\_RATE

→

$$\frac{\text{AMT\_ANNUITY}}{\text{AMT\_CREDIT}}$$

# Création de variables

# Features polynomiales

Création de features polynomiales de degré 3 avec les variables suivantes :

DAYS\_BIRTH

EXT\_SOURCE\_1

EXT\_SOURCE\_2

EXT\_SOURCE\_3

Exemple de nouvelles features

- $\text{EXT\_SOURCE\_1}^2$  ,  $\text{EXT\_SOURCE\_1}^3$
- $\text{EXT\_SOURCE\_1} * \text{EXT\_SOURCE\_2}$  ,  $\text{EXT\_SOURCE\_3} * \text{DAYS\_BIRTH}$
- $\text{EXT\_SOURCE\_2} * \text{EXT\_SOURCE\_3} * \text{DAYS\_BIRTH}$

Nombre total de features créées → 35

# Preprocessing

- Échantillonnage du jeu d'entraînement (50%) avec conservation des proportions de la cible
- Entraînement réalisé avec et sans les features polynomiales
- Méthodes de scaling testées →  $\begin{cases} \text{Standard Scaler} \\ \text{Min Max Scaler} \end{cases}$

Dimensions des jeux d'entraînement		
Jeu	Lignes	Colonnes
X_train_sampled	153755	244
X_train_poly_sampled	153755	275

# Modélisation

- Modèles testés
  - Logistic Regression
  - Random Forest Classifier
  - XGBoost Classifier
- Utilisation de GridSearchCV pour déterminer les hyperparamètres optimaux avec
  - Validation croisée à 5 plis
  - Score à optimiser : ROC AUC Score
- Stockage des résultats dans MlFlow



## Mlflow UI

## Expérimentations

### Experiments

Search Experiments

- ☐ XGBOOST Classifier
- ☐ Random Forest Classifier
- ☒ Logistic Regression

### Logistic Regression

Provide Feedback Add Description

Runs Evaluation **Experiments** Teams **Experiments**

Time created Static: Active Datasets Sort: Created Columns Group by New run

metrics

Run Name	Created	Duration	Accuracy	CV Accuracy	CV ROC AUC	CV Recall	F1 Score	ROC AUC	Recall Score	Score multiplier
randmlflow-experiment-641	9 days ago	25.7s	0.919	0.919	0.719	0.5	0	0.717	0	0.991
randmlflow-experiment-293	9 days ago	57.1s	0.919	0.919	0.744	0.501	0.007	0.738	0.004	0.993
randmlflow-experiment-485	9 days ago	1.86s	0.919	0.919	0.637	0.5	0	0.634	0	0.991
randmlflow-experiment-585	9 days ago	23.2s	0.919	0.919	0.719	0.5	0	0.717	0	0.991
randmlflow-experiment-912	9 days ago	56.4s	0.919	0.919	0.744	0.501	0.007	0.738	0.004	0.993
randmlflow-experiment-491	9 days ago	1.86s	0.919	0.919	0.637	0.5	0	0.634	0	0.991

Show more columns (8 total)

### Experiments

Search Experiments

- ☐ XGBOOST Classifier
- ☒ Random Forest Classifier
- ☐ Logistic Regression

### Random Forest Classifier

Provide Feedback Add Description

Runs Evaluation **Experiments** Teams **Experiments**

Time created Static: Active Datasets Sort: Accuracy Columns Group by New run

metrics

Run Name	Created	Duration	Accuracy	CV Accuracy	CV ROC AUC	CV Recall	F1 Score	ROC AUC	Recall Score	Score multiplier
randmlflow-experiment-641	8 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.731	0	0.991
randmlflow-experiment-293	9 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.731	0	0.991
randmlflow-experiment-245	9 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.732	0	0.991
randmlflow-experiment-77	9 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.732	0	0.991
randmlflow-experiment-734	9 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.732	0	0.991
randmlflow-experiment-338	9 days ago	1.3s	0.919	0.919	0.739	0.5	0	0.731	0	0.991

Show more columns (8 total)

### Experiments

Search Experiments

- ☒ XGBOOST Classifier
- ☐ Random Forest Classifier
- ☐ Logistic Regression

### XGBOOST Classifier

Provide Feedback Add Description

Runs Evaluation **Experiments** Teams **Experiments**

Time created Static: Active Datasets Sort: Accuracy Columns Group by New run

metrics

Run Name	Created	Duration	Accuracy	CV Accuracy	CV ROC AUC	CV Recall	F1 Score	ROC AUC	Recall Score	Score multiplier
randmlflow-experiment-796	8 days ago	21.6ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999
randmlflow-experiment-238	8 days ago	21.7ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999
randmlflow-experiment-243	8 days ago	21.6ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999
randmlflow-experiment-18	8 days ago	21.6ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999
randmlflow-experiment-715	8 days ago	22.8ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999
randmlflow-experiment-338	8 days ago	25.8ms	0.919	0.919	0.750	0.506	0.031	0.753	0.016	0.999

Show more columns (10 total)

# Synthèse des résultats

Modèle	Paramètres	Scaler	Poly	ROC
Logistic	<ul style="list-style-type: none"><li>• max_iter : 500</li><li>• penalty : l2</li><li>• solver : liblinear</li><li>• C : 0.001</li></ul>	STD	NON	0,738
RFC	<ul style="list-style-type: none"><li>• n_estimators : 1500</li><li>• min_samples_leaf : 5</li><li>• max_depth : 20</li></ul>	STD	NON	0,732
XGBoost	<ul style="list-style-type: none"><li>• n_estimators : 1000</li><li>• booster : gbtree</li><li>• learning_rate : 0.01</li></ul>	STD	NON	0,753

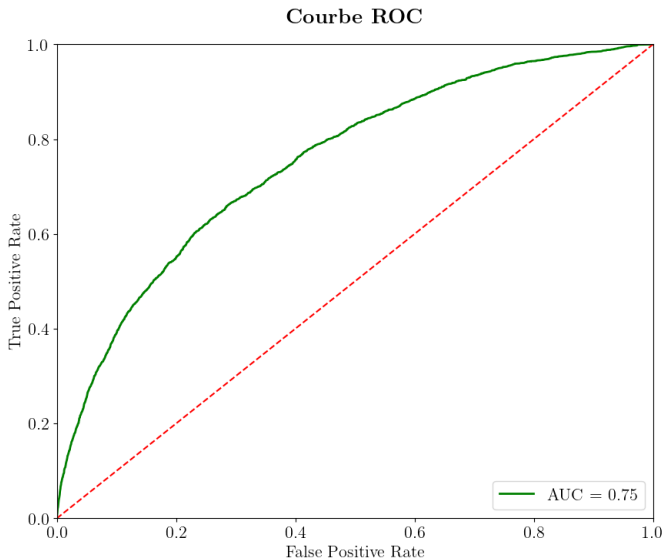
# Modèle retenu

## XGBoost Classifier

<div> <div>angry-ape-715</div> <div>Model registered</div> </div>																							
Overview	Model metricsSystem metricsAirStacks																						
<div> <div>Description</div> <div>No description</div> </div>																							
<div> <div>Details</div> <table> <tr> <td>Created at</td><td>2024-08-30 18:22:36</td></tr> <tr> <td>Created by</td><td>adrien</td></tr> <tr> <td>Experiment ID</td><td>88f248793779357967</td></tr> <tr> <td>Status</td><td>Finished</td></tr> <tr> <td>Run ID</td><td>2446518a660470c9f1a3bd792b4041d7</td></tr> <tr> <td>Duration</td><td>22 mins</td></tr> <tr> <td>Dataset used</td><td>-</td></tr> <tr> <td>Tags</td><td>AI4</td></tr> <tr> <td>Source</td><td>ipykernel_launcher.py</td></tr> <tr> <td>Logged models</td><td>sklearn</td></tr> <tr> <td>Registered models</td><td>sklearn-xgboost-origines-131</td></tr> </table> </div>		Created at	2024-08-30 18:22:36	Created by	adrien	Experiment ID	88f248793779357967	Status	Finished	Run ID	2446518a660470c9f1a3bd792b4041d7	Duration	22 mins	Dataset used	-	Tags	AI4	Source	ipykernel_launcher.py	Logged models	sklearn	Registered models	sklearn-xgboost-origines-131
Created at	2024-08-30 18:22:36																						
Created by	adrien																						
Experiment ID	88f248793779357967																						
Status	Finished																						
Run ID	2446518a660470c9f1a3bd792b4041d7																						
Duration	22 mins																						
Dataset used	-																						
Tags	AI4																						
Source	ipykernel_launcher.py																						
Logged models	sklearn																						
Registered models	sklearn-xgboost-origines-131																						
<div> <div>Parameters (4)</div> <table> <tr> <th>Parameter</th><th>Value</th></tr> <tr> <td>objective</td><td>binary:logistic</td></tr> <tr> <td>n_estimators</td><td>1008</td></tr> <tr> <td>booster</td><td>gbtree</td></tr> <tr> <td>learning_rate</td><td>0.01</td></tr> </table> </div>		Parameter	Value	objective	binary:logistic	n_estimators	1008	booster	gbtree	learning_rate	0.01												
Parameter	Value																						
objective	binary:logistic																						
n_estimators	1008																						
booster	gbtree																						
learning_rate	0.01																						
<div> <div>Metrics (8)</div> <table> <tr> <th>Metric</th><th>Value</th></tr> <tr> <td>ROC AUC</td><td>0.733</td></tr> <tr> <td>CV Recall</td><td>0.906</td></tr> <tr> <td>CV Accuracy</td><td>0.819</td></tr> <tr> <td>Score-matrix</td><td>0.899</td></tr> <tr> <td>CV ROC AUC</td><td>0.758</td></tr> <tr> <td>Recall Score</td><td>0.816</td></tr> <tr> <td>F1 Score</td><td>0.821</td></tr> <tr> <td>Accuracy</td><td>0.819</td></tr> </table> </div>		Metric	Value	ROC AUC	0.733	CV Recall	0.906	CV Accuracy	0.819	Score-matrix	0.899	CV ROC AUC	0.758	Recall Score	0.816	F1 Score	0.821	Accuracy	0.819				
Metric	Value																						
ROC AUC	0.733																						
CV Recall	0.906																						
CV Accuracy	0.819																						
Score-matrix	0.899																						
CV ROC AUC	0.758																						
Recall Score	0.816																						
F1 Score	0.821																						
Accuracy	0.819																						

La pipeline contenant StandardScaler & XGBClassifier  
est enregistrée sous format .pkl

# Courbe ROC



# **OPTIMISATION DU MODÈLE**

# Score de performance

## Définition

- Données  $\rightarrow \begin{cases} \text{Coût d'un FP} = 1 \\ \text{Coût d'un FN} = 10 \end{cases}$
- Définition du score de performance

$$\text{SCORE\_METIER} = \frac{\text{FP} \times 1 + \text{FN} \times 10}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

- Normalisation du score de performance après calculs

Légende	
TP : Vrais positifs	FP : Faux positifs
TN : Vrais négatifs	FN : Faux négatifs

# Procédure

- Normalisation du score de performance

$$\text{SCORE MÉTIER NORMALISÉ} = 1 - \frac{\text{SCORE MÉTIER}}{\text{Max (SCORE MÉTIER)}}$$

- Création d'un score général

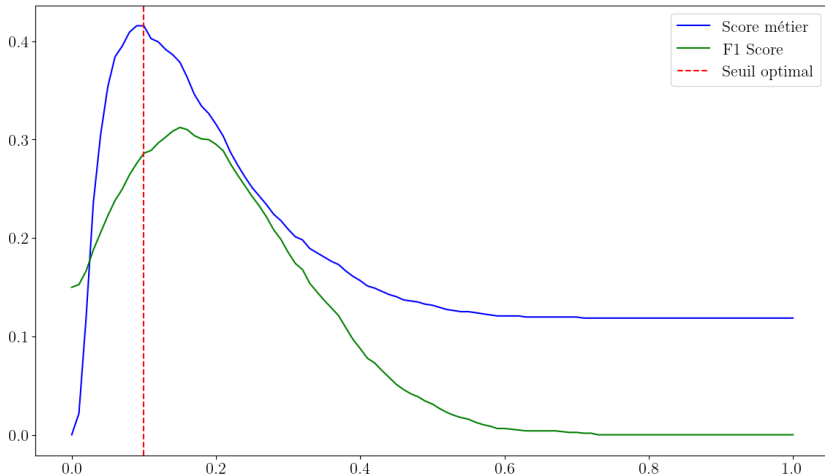
$$\text{SCORE GÉNÉRAL} = \text{SCORE MÉTIER NORMALISÉ} + \text{F1 SCORE}$$

## CONCLUSION

Seuil Optimal = 0,10

# Visualisation

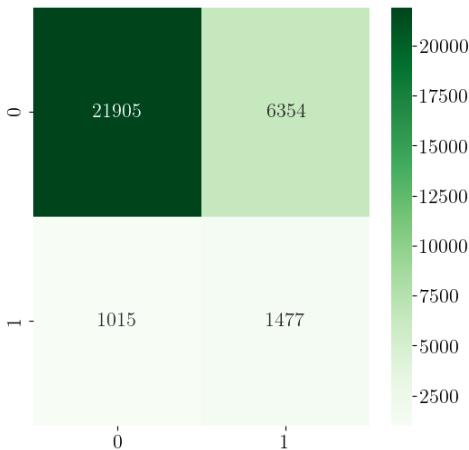
Scores en fonction du seuil





# Matrice de confusion

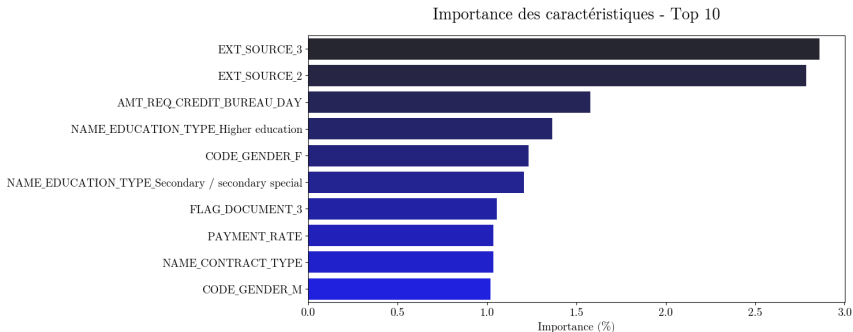
## Matrice de confusion



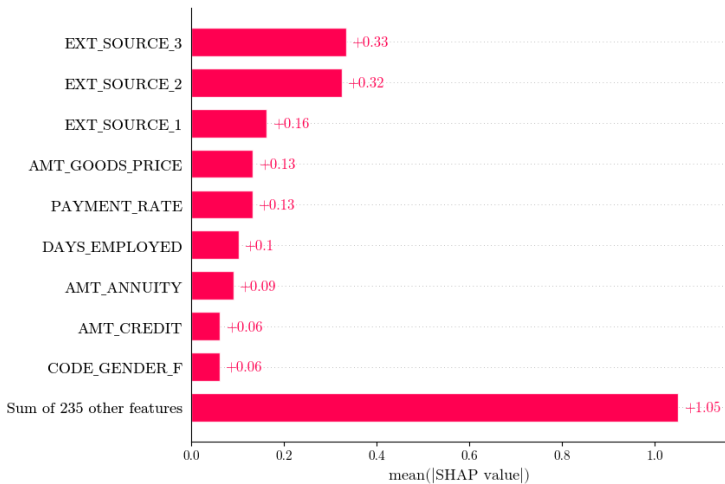
# **IMPORTANCE DES FEATURES**

# Importance Globale

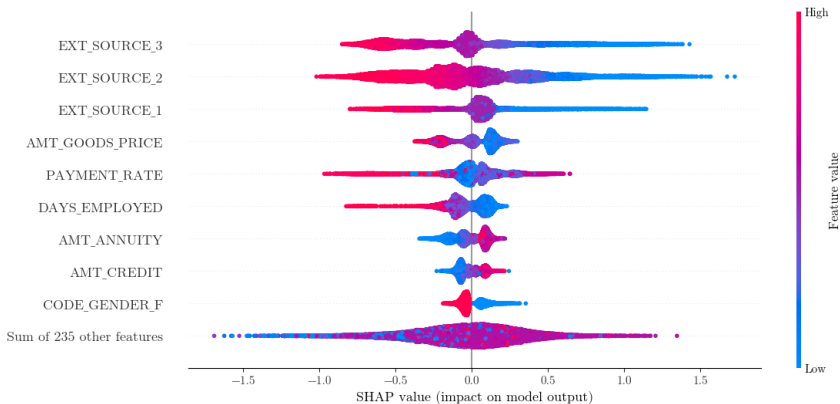
## Feature Importance



# SHAP Values



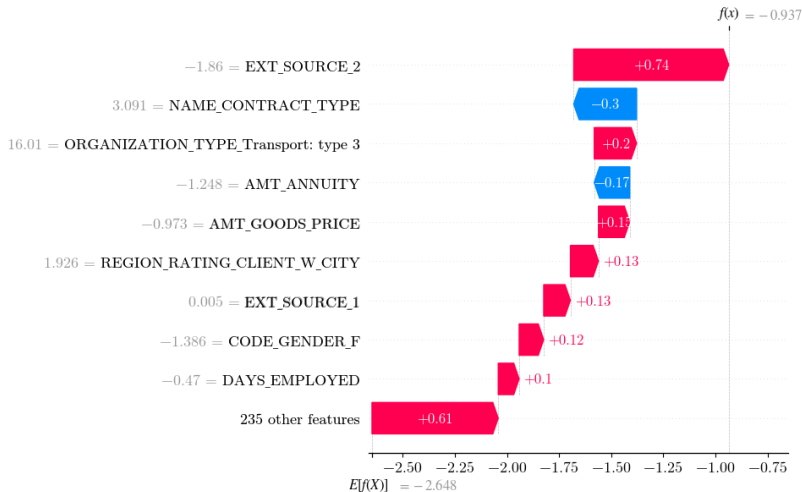
# Diagramme en abeille



# Importance locale

# Diagramme en cascade

## Observation n° 23



# DATA DRIFT

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

0.049  
Share of Drifted Columns

Drift is detected for 4.898% of columns (12 out of 245).

Search						
Column	Type	Reference Distribution	Current Distribution	Data Shift	Stat Test	Shift Score ↓
1	PAYMENT_RATE			Detected	Wasserstein distance (normed)	0.577936
2	AMT_REQ_CREDIT_BUREAU_QRT			Detected	Wasserstein distance (normed)	0.459483
3	AMT_REQ_CREDIT_BUREAU_MON			Detected	Wasserstein distance (normed)	0.204345
4	AMT_GOODS_PRICE			Detected	Wasserstein distance (normed)	0.213409
5	AMT_CREDIT			Detected	Wasserstein distance (normed)	0.207677
6	EXT_SOURCE_1			Detected	Wasserstein distance (normed)	0.130828
7	AMT_ANNUITY			Detected	Wasserstein distance (normed)	0.156061
8	NAME_CONTRACT_TYPE			Detected	Jensen-Shannon distance	0.547097
9	INCOME_CREDIT_PERCENT			Detected	Wasserstein distance (normed)	0.145547
10	AMT_REQ_CREDIT_BUREAU_WEEK			Detected	Wasserstein distance (normed)	0.145432
Rows per page: 10 rows						10 of 143 of 245



# **TABLEAU DE BORD**

# Tableau de bord

## Présentation

- Les données utilisées sont celles du jeu de test
- Les clients sont sélectionnés via `SK_ID_CURR`
- La sélection est présentée sous forme de liste déroulante
- L'utilisateur peut afficher les informations essentielles du client sélectionné
- L'utilisateur peut prédire la probabilité de faillite du client

# Illustration

## Informations client

### Simulation de Prêt Client



Sélectionnez l'identifiant client :

Client 180065

ID du client actuel : 100065

Informations clients

Caractéristiques	Valeurs
Age	26
Nombre d'enfants	0
Revenu total	394000

Simulation

## Simulation de prêt



Sélectionnez l'identifiant client :

Client 180065

ID du client actuel : 100065

Informations clients

Simulation



PRÊT ACCORDÉ

# CONCLUSION

# Bilan

# Modélisation

- Réalisation d'une analyse exploratoire des données
- Application de techniques de feature engineering afin d'améliorer les performances du modèle
- Évaluation de plusieurs modèles
- Optimisation des hyperparamètres via une recherche en grille
- Enregistrement et suivi des entraînements des modèles avec MLflow pour une meilleure traçabilité
- Création d'une pipeline intégrant les étapes de prétraitement et le modèle optimal
- Sauvegarde de la pipeline pour une réutilisation future
- Optimisation des prédictions du modèle en minimisant les coûts grâce à un score métier adapté au contexte

# Application

- La pipeline est importée et utilisée dans le dashboard via l'ID client
- Le dashboard affiche les informations du client ainsi que sa probabilité de défaut
- L'application décide automatiquement de l'accord du prêt en fonction du seuil optimal
- L'application est déployée et accessible à tous les utilisateurs

## Liens externes



Lien vers l'application **Streamlit**








Lien vers le repository **GitHub**






# GitHub

















# Repository

## GitHub Repository

 **ocr-project7** Public






 Pin  Unwatch 1  Fork 0  Star 0

 master  1 Branch  0 Tags   Add file  Code

 <b>amaysounabe</b> <span>Ajustement du README</span> 	cc1c760 · 2 weeks ago	 27 Commits
 .github/workflows	correction librairies	2 weeks ago
 data	Version définitive	2 weeks ago
 styles	Version définitive	2 weeks ago
 README.md	Ajustement du README	2 weeks ago
 dashboard_fonctions.py	Dernière correction	2 weeks ago
 dashboard_fonctions_test.py	nouvelle correction	2 weeks ago
 dashboard_interface.py	Poussée du code avec tests unitaires	2 weeks ago
 data_drift_report.html	Version définitive	2 weeks ago
 df_test.csv	Allègement fichier csv	3 weeks ago
 fonctions.py	Poussée du code avec tests unitaires	2 weeks ago
 modelisation_notebook.ipynb	Changement de nom pour le notebook	2 weeks ago
 pytest-requirements.txt	correction librairies	2 weeks ago
 requirements.txt	Dernière correction	2 weeks ago

### About

No description, website, or topics provided.

 Readme  Activity  0 stars  1 watching  0 forks


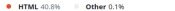

### Releases

No releases published  
[Create a new release](#)

### Packages

No packages published  
[Publish your first package](#)

### Languages

 **Jupyter Notebook** 59.1%  
 **HTML** 40.8%  **Other** 0.1%

# Tests unitaires

## GitHub Actions

### test

succeeded 2 weeks ago in 39s

- > Set up job
- > Checkout code
- > Set up Python
- > Create a virtual environnement
- > Install dependencies
- ▼ Run Pytest

```
1 ▶ Run pytest -p no:warnings
11 ===== test session starts =====
12 platform linux -- Python 3.12.5, pytest-8.3.3, pluggy-1.5.0
13 rootdir: /home/runner/work/ocr-project7/ocr-project7
14 collected 3 items
15
16 dashboard_fonctions_test.py ... [100%]
17
18 ===== 3 passed in 1.16s =====
```

- > Post Set up Python
- > Post Checkout code
- > Complete job



**FIN**