

---

# PROJET N°7

**“IMPLEMENTEZ UN MODELE DE SCORING”**

# SOMMAIRE

## Avant Propos

**Etape 1** : Données fournies et data preprocessing

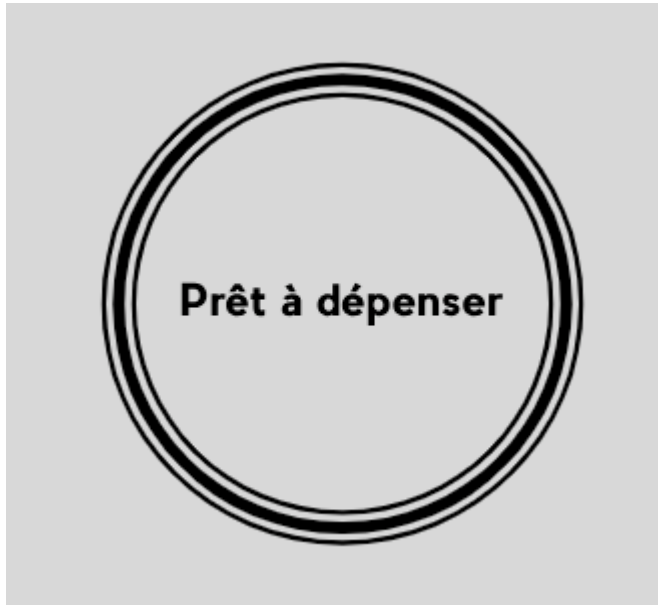
**Etape 2** : Modélisation

**Etape 3** : Dashboard

**Etape 4** : Pipeline de déploiement

**Etape 5** : Conclusion

# AVANT PROPOS



## CONTEXTE & OBJECTIFS :

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

L'entreprise souhaite

- ✓ mettre en œuvre un **outil de scoring crédit** qui calcule la probabilité qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé.
- ✓ - développer **un dashboard interactif** pour plus de transparence afin que les chargés de relation client puissent à la fois expliquer de façon la plus transparente possible les décisions d'octroi de crédit, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

## **ETAPE 1 :**

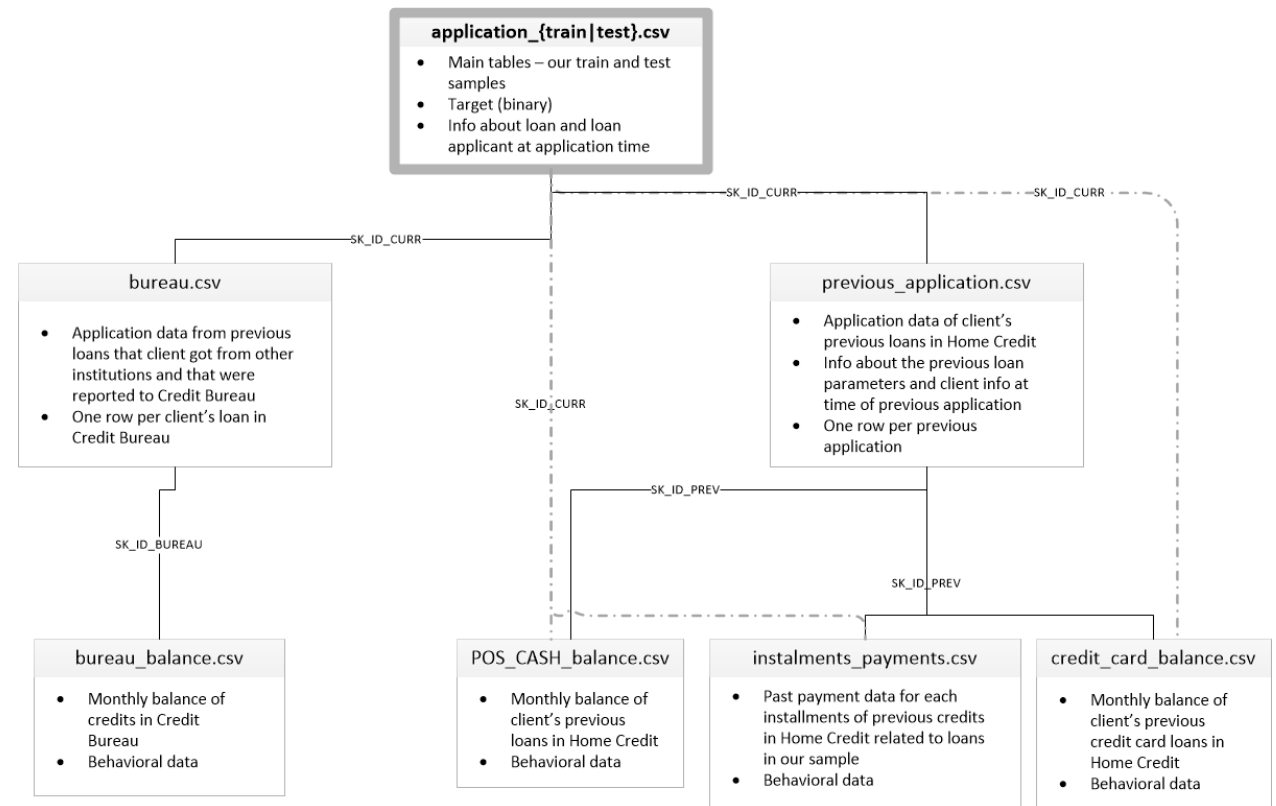
DONNÉES FOURNIES &  
DATA PREPROCESSING

# DONNEES FOURNIES



## 7 Datasets:

- ✓ Reprenant des informations personnelles du client et sur le crédit souhaité
- ✓ Des historiques des crédits dans l'établissement bancaire
- ✓ Des historiques des crédits demandés dans d'autres institutions financières



# DATA PREPROCESSING

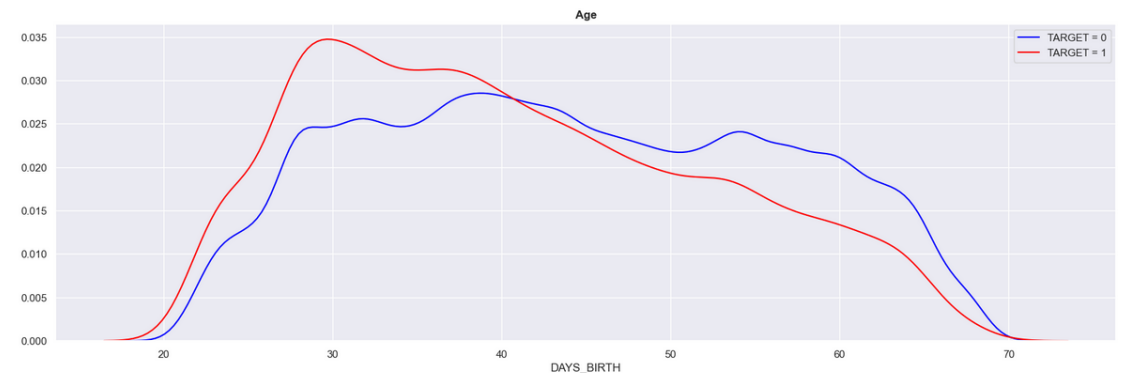
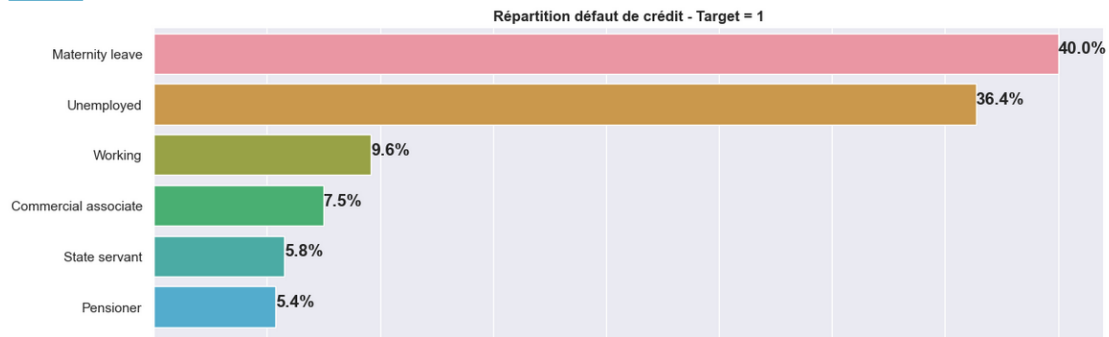
1

## Analyse des 7 datasets séparément

- ✓ suppression des features non pertinentes
- ✓ Analyse des variables manquantes ou aberrantes
- ✓ Suppression de features fortement corrélées

2

## Analyse exploratoire



3

## Agrégation des data

- ✓ Sur une data globale (train + test), intégration features brut ou agréé des autres fichiers
- ✓ Traitement des valeurs manquantes (Avec simpleimputer – Médiane)

----> A ce stade : 170 features

# DATA PREPROCESSING

## 4

### Préprocessing des données

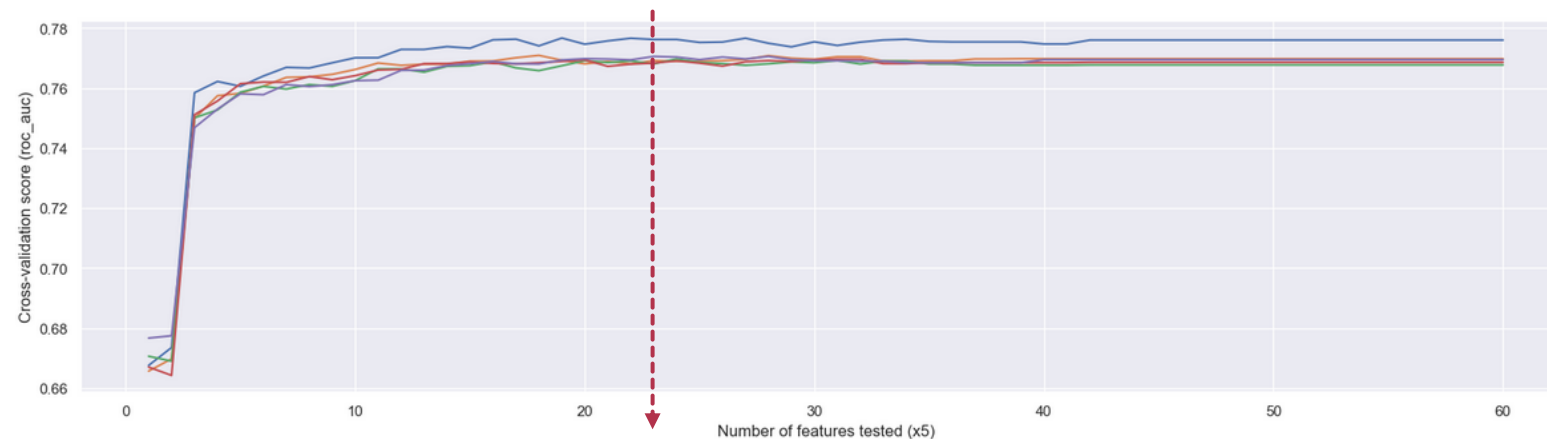
- ✓ Création de 4 nouvelles variables métier
- ✓ Encodage des variables catégorielles (Labelencoder) + dummies
- ✓ Standardisation des données (MinMaxScaler)

-----> A ce stade : 294 features

## 5

### Réduction de dimension

- ✓ Utilisation de la technique d'élimination des caractéristiques récursives avec validation croisée ( RFECV ).



**Seuil Optimal AUC/Features : 112**



**ETAPE 2 :**

MODELISATION

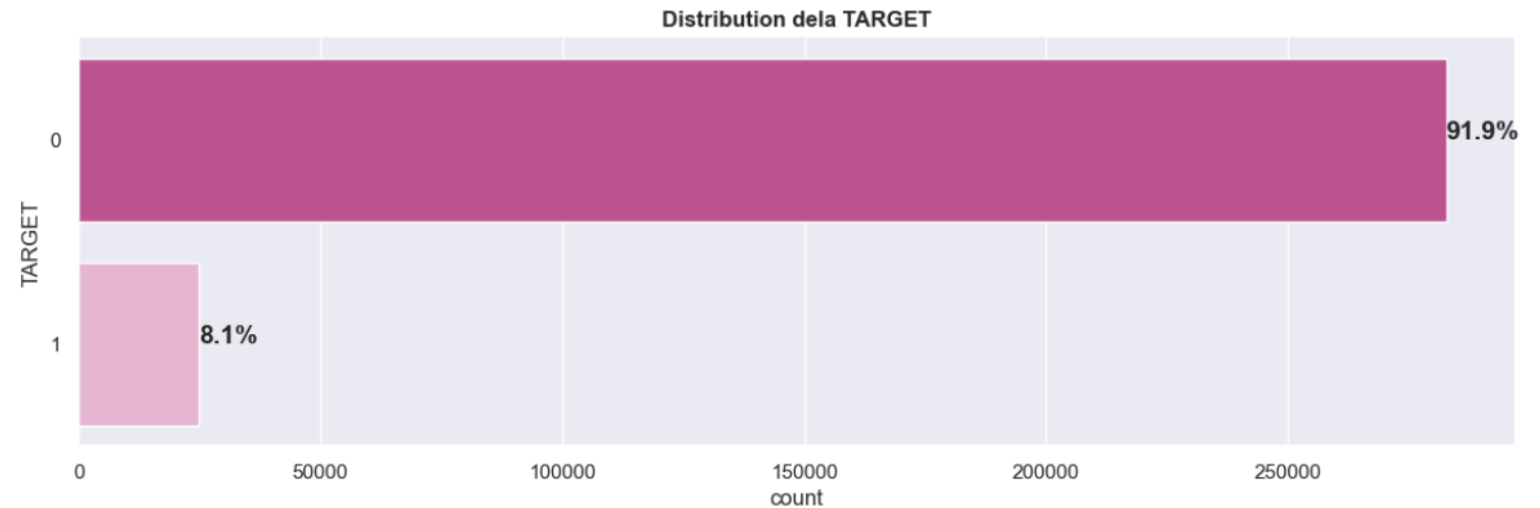




# MODELISATION

## TRAITEMENT DU DESEQUILIBRE DES CLASSES

**PROBLEME** : Les clients en difficulté de paiement sont largement sous-représentés (8.1%) dans les données d'entraînement  
→ Mauvais résultat ou induire en erreur avec des scores trop optimistes



### SOLUTIONS :

- ✓ **Méthodes de Data Level** qui consistent en une modification des données d'entraînement (comme over sampling, under sampling ou SMOTE (Suréchantillonnage))
- ✓ **Les Méthodes Algorithm-level** qui ne modifient pas les données, mais qui pénalisent les mauvaises prédictions sur la classe minoritaire en donnant plus de poids à la fonction de perte

# MODELISATION

## METRIQUES & SCORE METIER

### La matrice de confusion



### Les métriques classiques

#### Accuracy

% de bonnes prédictions

- Affecté par le déséquilibre des classes
- Ne distingue pas les erreurs commises

#### Rappel

% de la classe positive détectée

- Reflète la capacité à éviter les faux positifs

#### Précision

% des vrais positifs dans les positifs détectés

- Reflète la capacité à détecter tous les cas positifs

#### F1-score

Moyenne harmonique de Précision / Rappel

- Combinaison du rappel et de la précision
- A privilégier si déséquilibre des classes

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

**Le score métier :** d'un point de vue métier, le risque de refuser un crédit à tort (faux négatifs) n'a évidemment pas le même poids que d'accorder un crédit à un client n'ayant pas honoré son engagement (faux positifs).

→ Création d'un **score métier normalisé** en prenant en compte un risque 10 x plus élevé de prédire un faux positif qu'un faux négatif. Un poids de 1 est attribué aux bonnes prédictions :

$$SCORE_{METIER} = \frac{(J - J_{min})}{(J_{max} - J_{min})}$$

Où

$$J = 1 * TP + 1 * TN - 1 * FP - 10 * FN$$

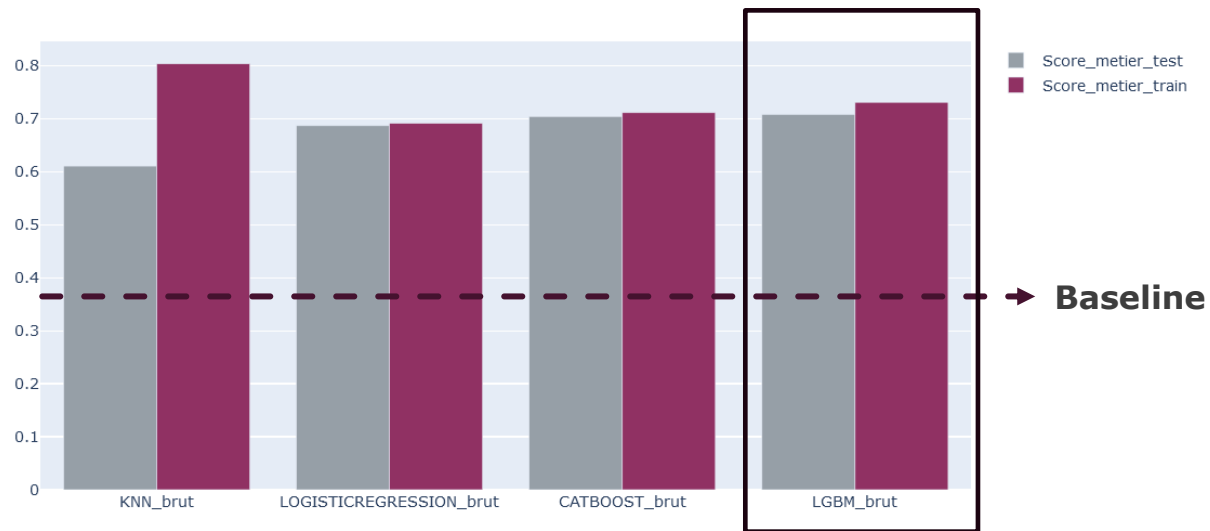
$$J_{max} = (FP + TN) * 1 + (FN + TP) * 1$$

$$J_{min} = (FP + TN) * (-1) + (FN + TP) * (-10)$$

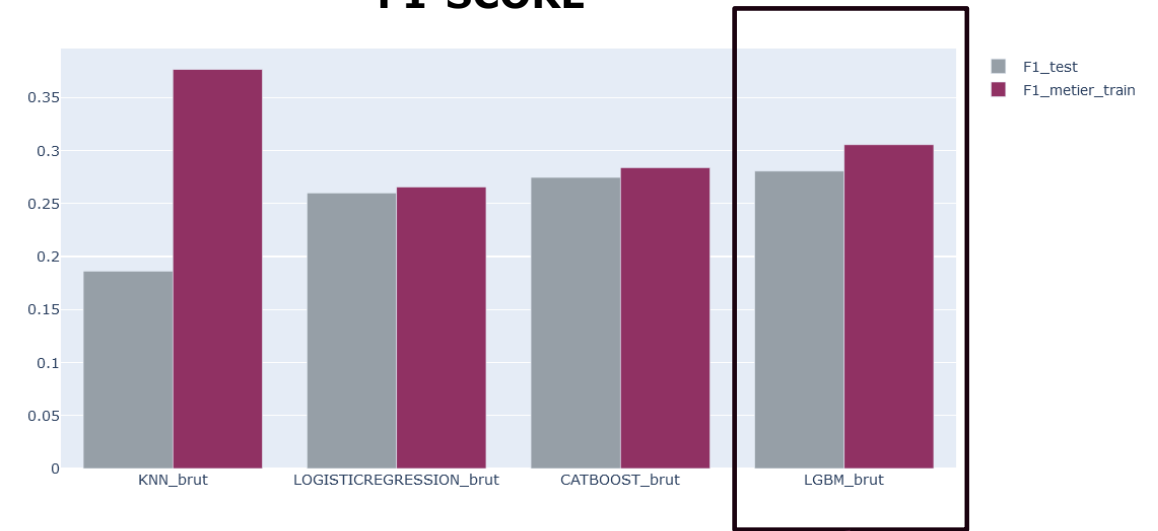
# MODELISATION

## MODELISATION

### SCORE METIER



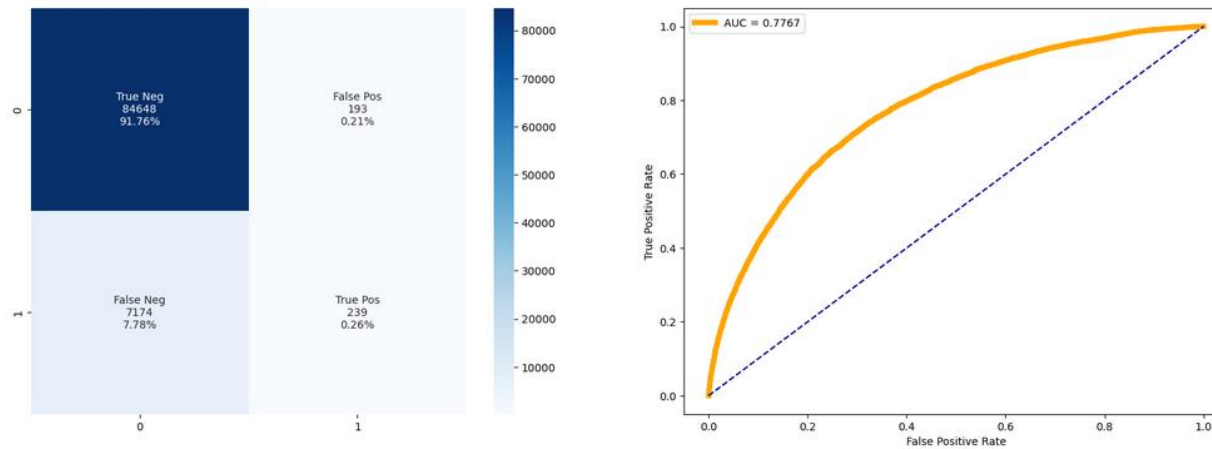
### F1-SCORE



Meilleurs résultats pour le Score métier comme pour le F1\_Score avec LGBM

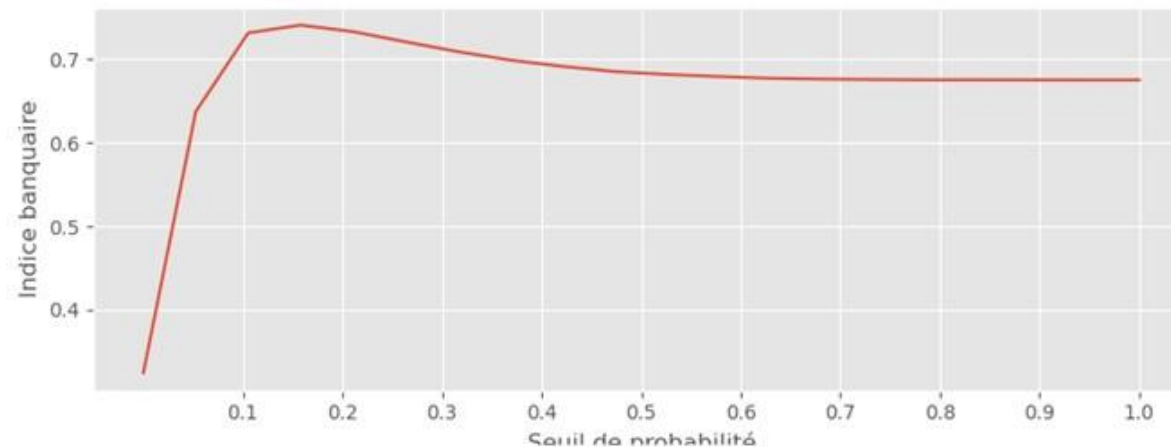
# MODELISATION

## MODELISATION



Ci-contre la matrice de confusion et la courbe ROC, qui montre un AUC de 0.7767.

On peut également noter que le seuil optimal du seuil de probabilité est aux alentours de 0,10.

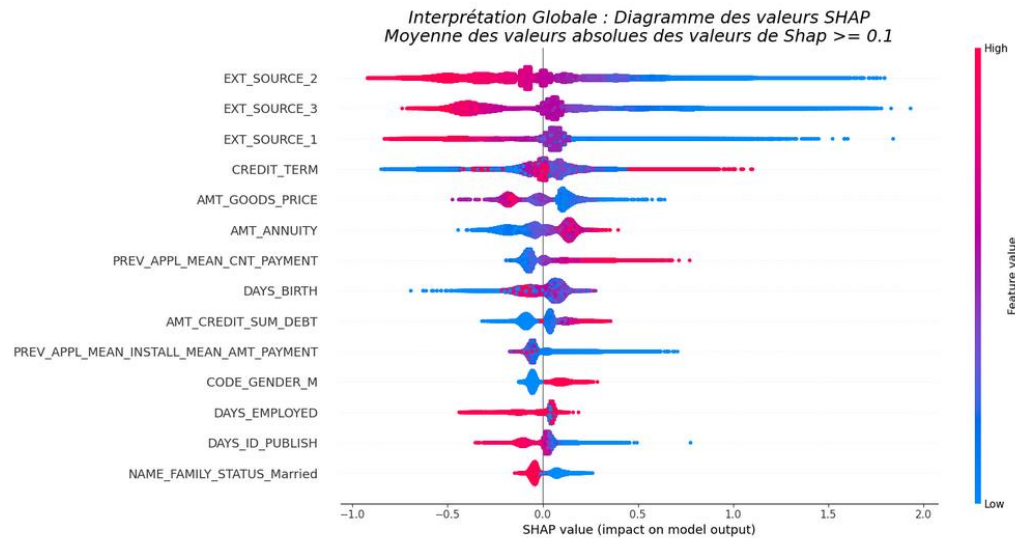


# MODELISATION

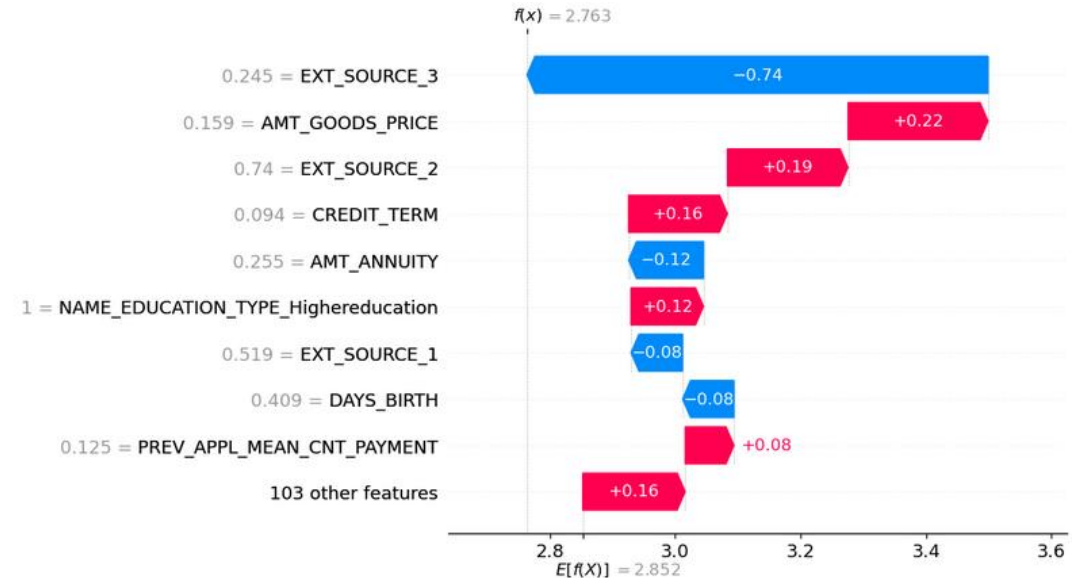
## INTERPRETATION DU MODELE

**LES VALEURS DE SHAP** : elles calculent l'importance d'une variable en comparant ce qu'un modèle prédit avec et sans cette variable. Cependant, étant donné que l'ordre dans lequel un modèle voit les variables peut affecter ses prédictions, cela se fait dans tous les ordres possibles, afin que les fonctionnalités soient comparées équitablement.

### INTERPRETATION GLOBALE



### INTERPRETATION LOCALE



# ETAPE 3 :

## DASHBOARD



# DASHBOARD

## INTERPRETATION DU MODELE

### → Dashbord permet :

- De sélectionner un client
- D'afficher des éléments clés du portefeuille source ayant servi à la modélisation
- Affichage des informations pertinentes du clients et positionnement par rapport au portefeuille source
- Affichage de la probabilité de risque de défaut de crédit pour le client sélectionner et représentation locale de SHAP



```
FROM python:3.10.11-slim-buster

# Récupération fichiers
COPY . /app1

# Répertoire de travail
WORKDIR /app1

# Dépendance pour LightGBM
RUN apt-get update
RUN apt-get install -y libgomp1

# Using pip:
RUN python -m pip install -r requirements.txt

# Déclaration du port d'entrée à l'app depuis l'extérieur du container
EXPOSE 8501

HEALTHCHECK CMD curl --fail http://localhost:8501/_stcore/health

# Déplacement des fichiers de configuration de streamlit
RUN cp config.toml ~/config.toml
RUN cp credentials.toml ~/credentials.toml
WORKDIR /app1

# Lance streamlit
CMD ["streamlit", "run", "app1.py"]
```

```
Windows PowerShell
Copyright (C) Microsoft Corporation. Tous droits réservés.

Installer la dernière version de PowerShell pour de nouvelles fonctionnalités et améliorations : https://aka.ms/PSWindows

PS C:\Users\Valere> cd C:\Users\Valere\Documents\OPENCLASIMON\PMOLIST T:\PMOLIST_API
PS C:\Users\Valere\Documents\OPENCLASIMON\PMOLIST T:\PMOLIST_API> streamlit run app1.py

You can now view your Streamlit app in your browser.

URL: http://localhost:8501
Network URL: http://192.168.1.12:8501

A new version of Streamlit is available.
See what's new at https://discuss.streamlit.io/announcements

version 1.1.3 when using version 1.1.2. This might lead to breaking code or invalid results.
See https://discuss.streamlit.io/announcements for more details.
Warning: The 'st.sidebar.selectbox' method is deprecated. Use 'st.selectbox' instead.
Warning: The 'st.sidebar.markdown' method is deprecated. Use 'st.markdown' instead.
Warning: The 'st.sidebar.text' method is deprecated. Use 'st.text' instead.
```

```
html_temp = """
<div style="background-color: #054773; padding:10px; border-radius:10px">
<div style="color: white; text-align:center">Dashboard Scoring Credit</div>
</div>

"""
st.markdown(html_temp, unsafe_allow_html=True)

#Customer ID selection
st.sidebar.header("**INFORMATION GENERAL**")

#Loading selector
chk_id = st.sidebar.selectbox("Client ID", id_client)

#Loading general info
nb_credits, rev_moy, credits_moy = load_infos_gen(data)

## Display of information in the sidebar ##
#Number of loans in the sample
st.sidebar.markdown("<u>NOMBRE DE CREDIT :</u>", unsafe_allow_html=True)
st.sidebar.text(nb_credits)

#Average income
st.sidebar.markdown("<u>REVENU MOYEN DATA:</u>", unsafe_allow_html=True)
st.sidebar.text(rev_moy)

#RAT CREDIT
st.sidebar.markdown("<u>MONTANT MOYEN DU CREDIT DATA :</u>", unsafe_allow_html=True)
st.sidebar.text(credits_moy)

# HOME PAGE - MAIN CONTENT
#####

#Customer information display : Customer Gender, Age, Family status, Children, ...
st.header(" INFORMATION CLIENT SELECTIONNE ")

if st.checkbox("AFFICHER LES INFORMATIONS SUR LE CLIENT ?"):
    info_client = identify_client(data, chk_id)
    st.write(" SEXE : ", info_client["COM_GENDER"].values[0])
    st.write(" AGE : ", info_client["DAYS_BIRTH"][-365])
    st.write("SITUATION DE FAMILLE : ", info_client["NAME_FAMILY_STATUS"].values[0])
```

## **ETAPE 6 :**

## PIPELINE DE DEPLOIEMENT



# PIPELINE DE DÉPLOIEMENT

## GITHUB



Site de partage de code, sur lequel on peut publier des projets dont le code est géré avec le système de gestion de version Git.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. Tous droits réservés.

Installez la dernière version de PowerShell pour de nouvelles fonctionnalités et améliorations ! https://aka.ms/PSWindows

PS C:\WINDOWS\system32> cd C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git init
Initialized empty Git repository in C:/Users/helen/Documents/OPENCLASSROOM/PROJET 7/.git/
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git add fonctions_EDA
fatal: pathspec 'fonctions_EDA' did not match any files
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git add fonctions_EDA.py
warning: in the working copy of 'fonctions_EDA.py', LF will be replaced by CRLF the next time Git touches it
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git commit -m "Fichier fonctions_EDA"
[master (root-commit) 00602b2] Fichier fonctions_EDA
 1 file changed, 88 insertions(+)
   create mode 100644 fonctions_EDA.py
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git branch -M master
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git remote add origin git@github.com:helene1219/PROJET_7.git
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> git push -u origin master
Enumerating objects: 3, done.
Counting objects: 100% (3/3), done.
Delta compression using up to 16 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (3/3), 1.33 KiB | 1.33 MiB/s, done.
Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
To github.com:helene1219/PROJET_7.git
 * [new branch]      master -> master
branch 'master' set up to track 'origin/master'.
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7>
```

## PYTEST



Bibliothèque Python pour tester son code et assurer sa qualité et son bon fonctionnement

```
PS C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7> pytest test_Projet7.py
===== test session starts =====
platform win32 -- Python 3.10.11, pytest-7.4.4, pluggy-1.4.0
rootdir: C:\Users\helen\Documents\OPENCLASSROOM\PROJET 7
plugins: anyio-4.2.0
collected 5 items

test_Projet7.py ..... [100%]

===== 5 passed in 2.74s =====
```

# PIPELINE DE DÉPLOIEMENT



Il est essentiel de suivre un modèle en production et notamment :

- ✓ La dérive et qualité des données .
- ✓ La dérive de la cible et la performance du modèle

## DERIVE ET QUALITE DES DONNEES

### Dataset Drift

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

113

Columns

15

Drifted Columns

0.133

Share of Drifted Columns

### Data Drift Summary

Drift is detected for 13.274% of columns (15 out of 113).

Search					
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test
> prediction	num			Not Detected	Wasserstein distance (normed)
> STATUS	num			Detected	Wasserstein distance (normed)

15

Tests

15

Success

0

Warning

0

Fail

0

Error

All tests



Number of Columns

DETAILS

The number of columns is 113. The test threshold is eq=113.



Number of Rows

The number of rows is 5000. The test threshold is eq=5e+09 a 500.



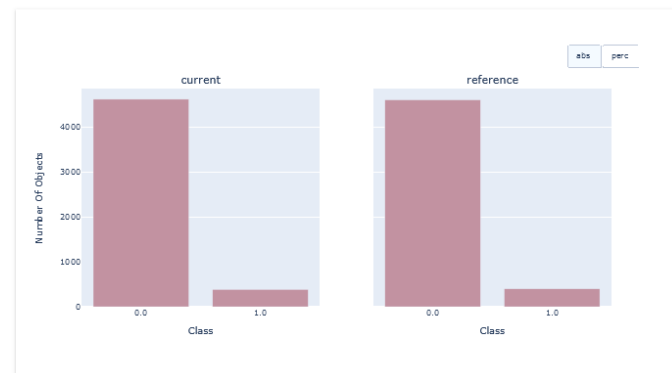
The Number of Missing Values

DETAILS

The number of missing values is 0. The test threshold is lt=0 a 1e-12.

## PERFORMANCE DU MODELE

### Class Representation



### Classification Model Performance. Target: 'target'

Current: Model Quality Metrics

0.926

Accuracy

0.727

Precision

0.042

Recall

0.079

F1

Reference: Model Quality Metrics

0.921

Accuracy

0.609

Precision

0.035

Recall

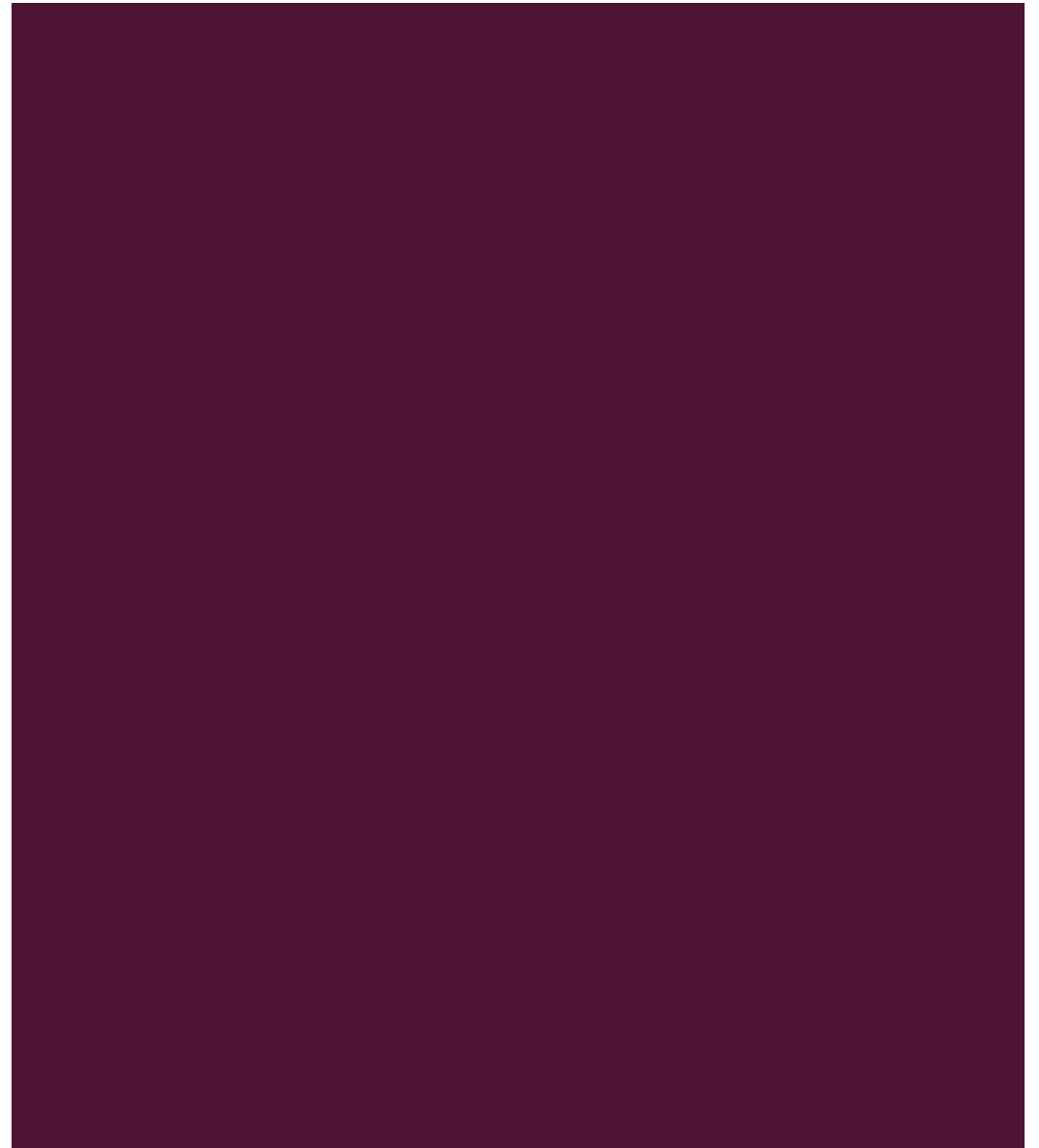
0.067

F1

---

## **ETAPE 6 :**

## CONCLUSION



# CONCLUSION

Les données mise à disposition ont permis de mettre en place un algorithme de classification assez efficace avec une modélisation optimale avec l'algorithme LGBM. Cependant, afin de pouvoir améliorer ces résultats, il serait pertinent :

- ✓ **De disposer d'une meilleure méconnaissance du milieu bancaire** ce qui permettrait de vérifier / améliorer le processus de traitement des données
- ✓ D'avoir une meilleure compréhension des variables essentielles dans l'explication du modèle (EXT\_SOURCE)
- ✓ De définir plus finement la **métrique d'évaluation** et la **fonction de coût** en collaboration avec les équipes métier
- ✓ D'améliorer les performances de la modélisation en intégrant de nouveaux hyperparamètres et / ou augmentant les valeurs testées
- ✓ D'avoir des connaissances de développement plus robuste afin de sécuriser / automatiser parfaitement le déploiement

