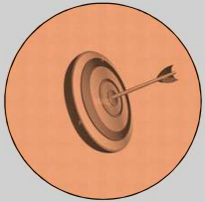
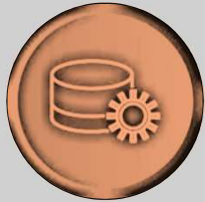


**Implémentez un modèle  
de scoring**

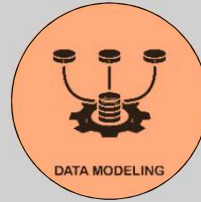
# Sommaire



**CONTEXTE  
& OBJECTIF**



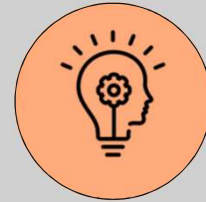
**EXPLORATION DES  
DONNEES**



**MODELISATION**

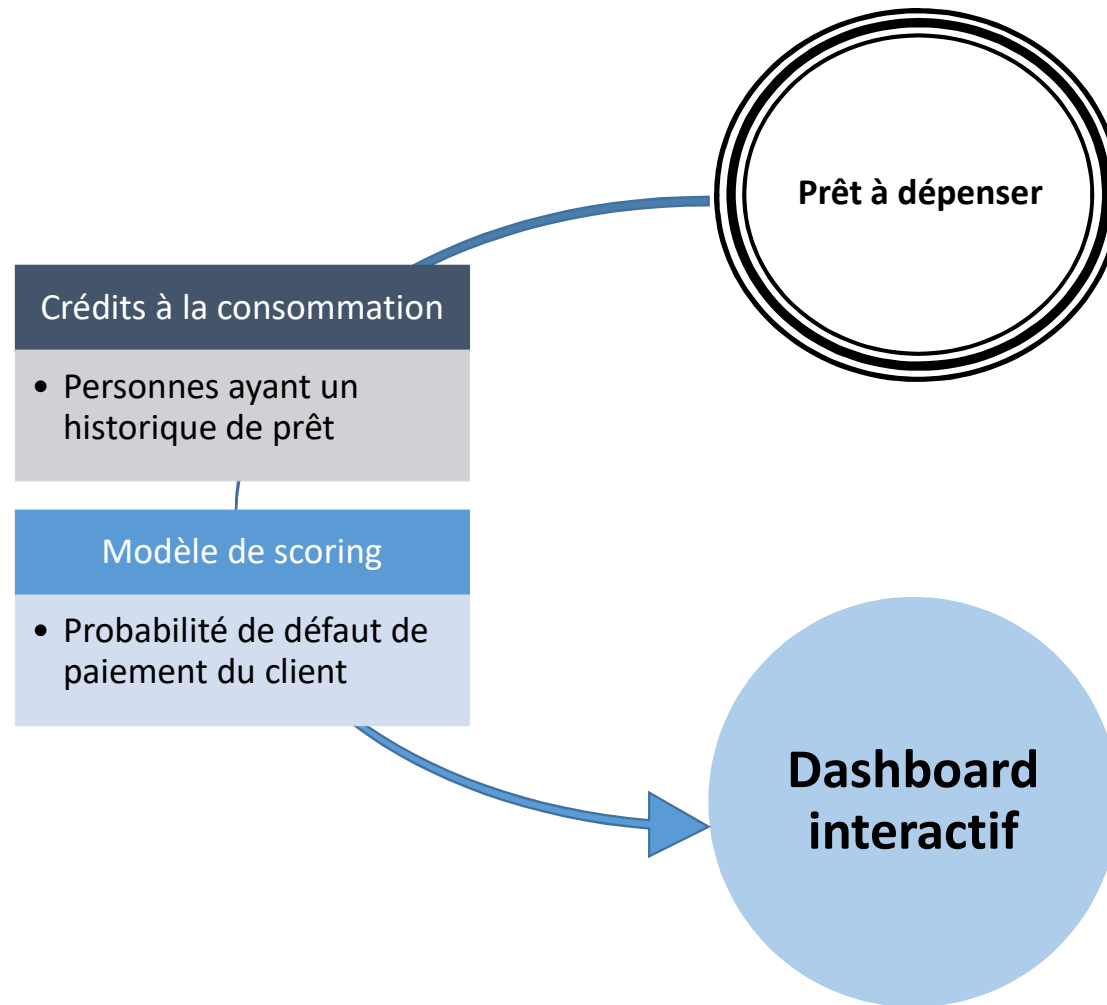


**DASHBOARD**



**CONCLUSION**

## CONTEXTE ET OBJECTIFS



# Impact marché !

LE CRÉDIT À LA CONSOMMATION  
CONCERNE PLUS D'UN MÉNAGE SUR 4\*



\*SOURCE : OBSERVATOIRE DES CRÉDITS DES MÉNAGES, JANVIER 2018

## Exploration des données

### Jeu de données principal

#### application\_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

#### bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

SK\_ID\_BUREAU

#### bureau\_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

#### POS\_CASH\_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

#### previous\_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

SK\_ID\_PREV

#### instalments\_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

#### credit\_card\_balance.csv

- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

Parametres des prêts antérieurs chez Home Credit

données comportementales

[Home Credit Default Risk | Kaggle](https://www.kaggle.com/competitions/home-credit-default-risk/data)

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

prêts que le client a obtenus auprès d'autres institutions et qui ont été signalés au bureau de crédit

## Description des données

application\_train - rows: 307511 columns: 122

application\_test - rows: 48744 columns: 121

bureau - rows: 1716428 columns: 17

bureau\_balance - rows: 27299925 columns: 3


credit\_card\_balance - rows: 3840312 columns: 22

installments\_payments - rows: 13605401 columns: 7

previous\_application - rows: 1670214 columns: 37

POS\_CASH\_balance - rows: 10001358 columns: 7

sample\_submission - rows: 48744 columns: 2



float64	65
int64	41
object	16

## Prétraitements

Valeurs manquantes : imputation

Variables numériques et catégorielles

Outliers : Valeurs atypiques

L'âge du client, durée d'emploi

Corrélation : avec la cible

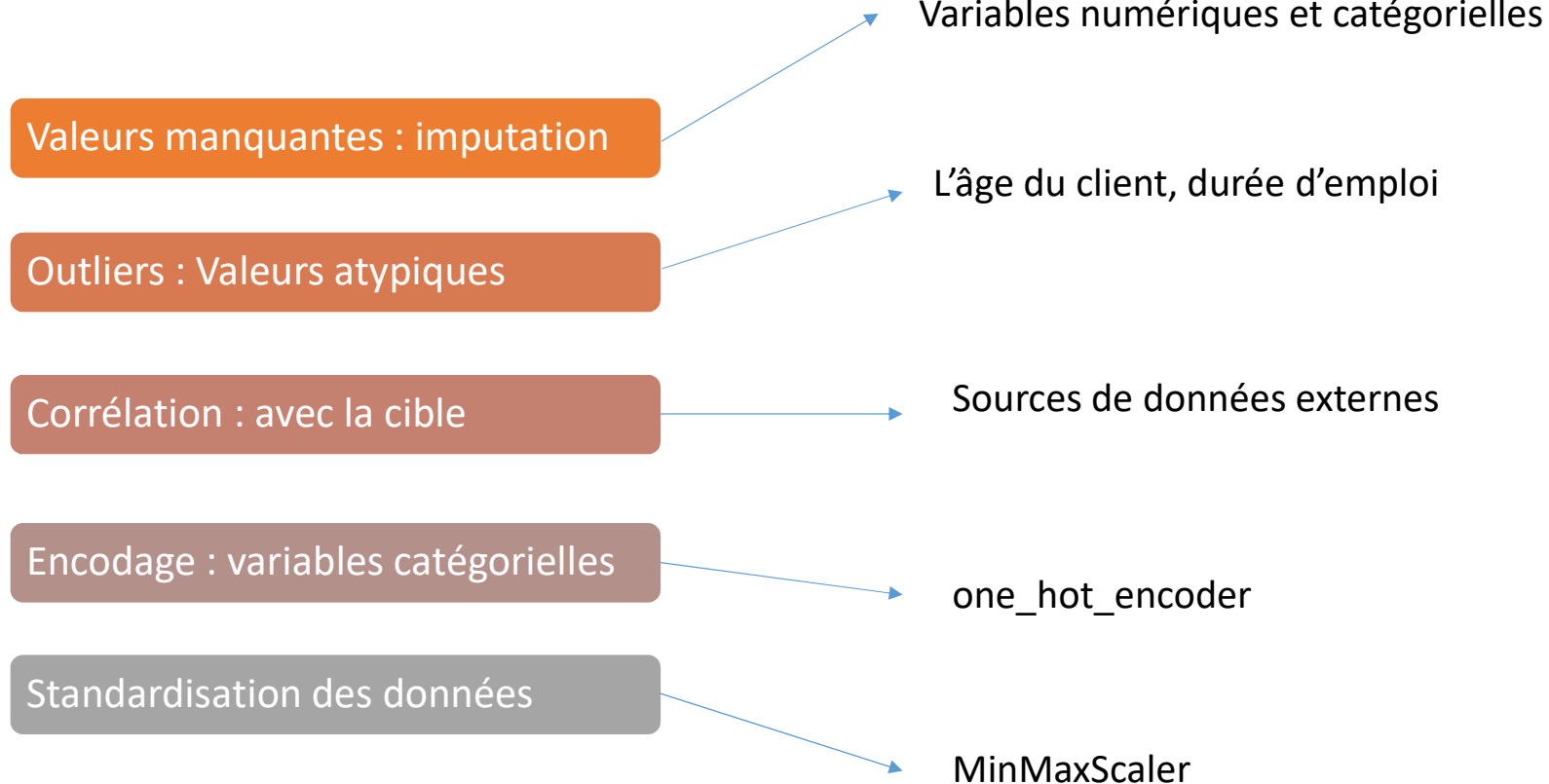
Sources de données externes

Encodage : variables catégorielles

one\_hot\_encoder

Standardisation des données

MinMaxScaler



ENRECHISSEMENT

## Opération de fusion

Echantillon de travail principal initial :

356255, 123

Combinaison des 7 jeux de données :

### Merging et agrégations de données

- PREVIOUS\_LOANS\_COUNT
- MONTHS\_BALANCE\_MEAN
- PREVIOUS\_APPLICATION\_COUNT

Features engineering :

### Ratios explicatifs

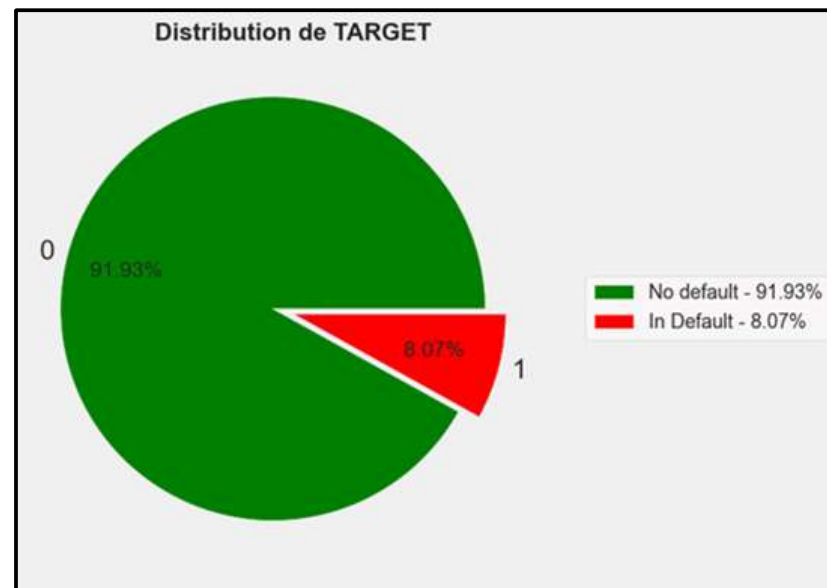
- CREDIT\_INCOME\_PERCENT : montant du crédit / revenu du client
- ANNUITY\_INCOME\_PERCENT : la rente du crédit / revenu du client
- DAYS\_EMPLOYED\_PERCENT : jours d'emploi / l'âge du client
- CREDIT\_TERM : la rente du crédit / montant du prêt

Echantillon de travail final:

356255, 192



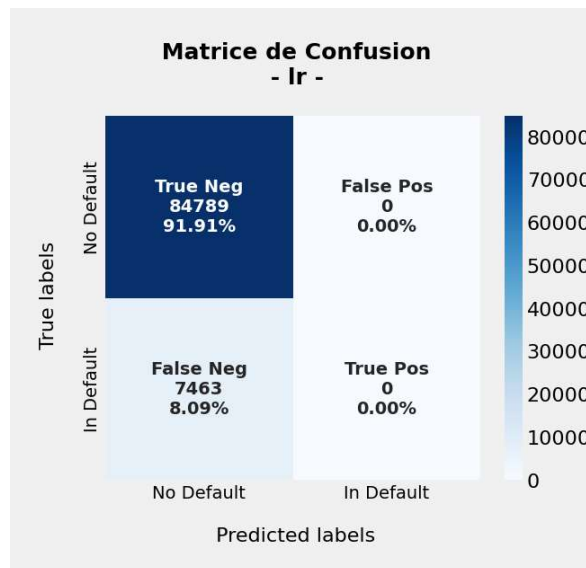
### Distribution de TARGET



Déséquilibre de classe nécessitant un rééchantillonnage (Resampling)

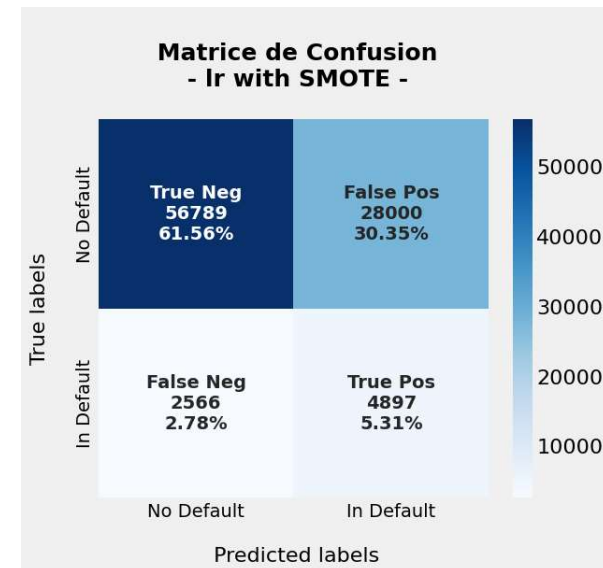
# Modélisation

Baseline model fixé par **Logistic Regression**



**AUC : 0.69**

Sans resampling



**AUC : 0.72**

Avec Oversampling SMOTE

## Elaboration d'un modèle à base d'un algorithme de Gradient Boosting

### LightGBM

Rééchantillonnage des données d'entraînement

Imblearn

Undersampling

Oversampling

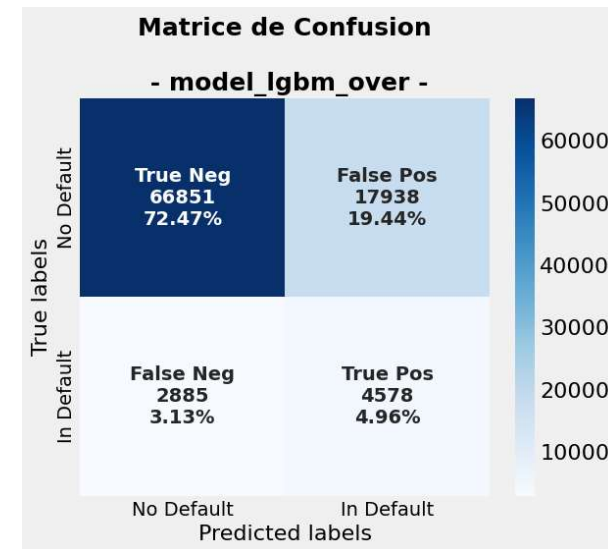
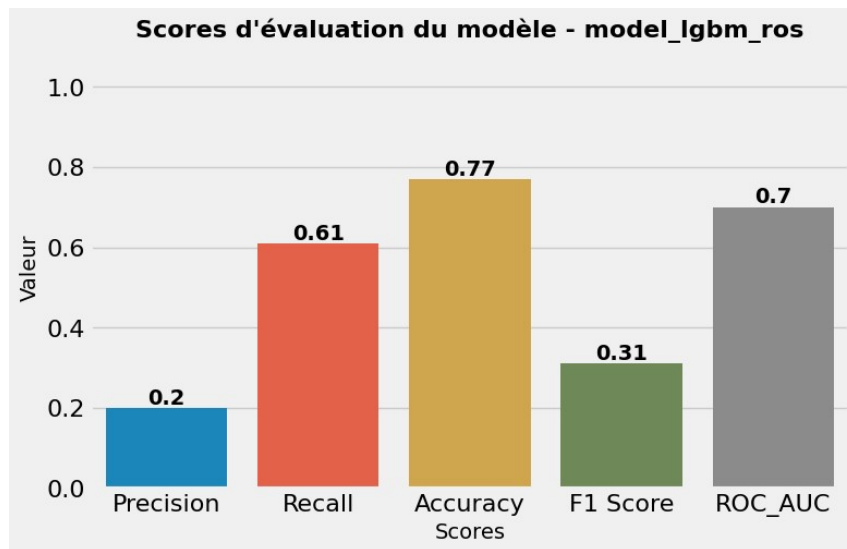
SMOTE

```
params = {  
    'objective': 'binary',  
    'metric': 'auc',  
    'n_estimators': 2000,  
    'learning_rate': 0.02,  
    'num_leaves': 34,  
    'colsample_bytree': 0.9497036,  
    'subsample': 0.8715623,  
    'max_depth': 8,  
    'reg_alpha': 0.041545473,  
    'reg_lambda': 0.0735294,  
    'min_split_gain': 0.0222415,  
    'min_child_weight': 39.3259775  
}
```

"Bayesian optimization" sur

<https://www.kaggle.com/tili7/olivier-lightgbm-parameters-by-bayesian-opt/code>

## Evaluation & Scores



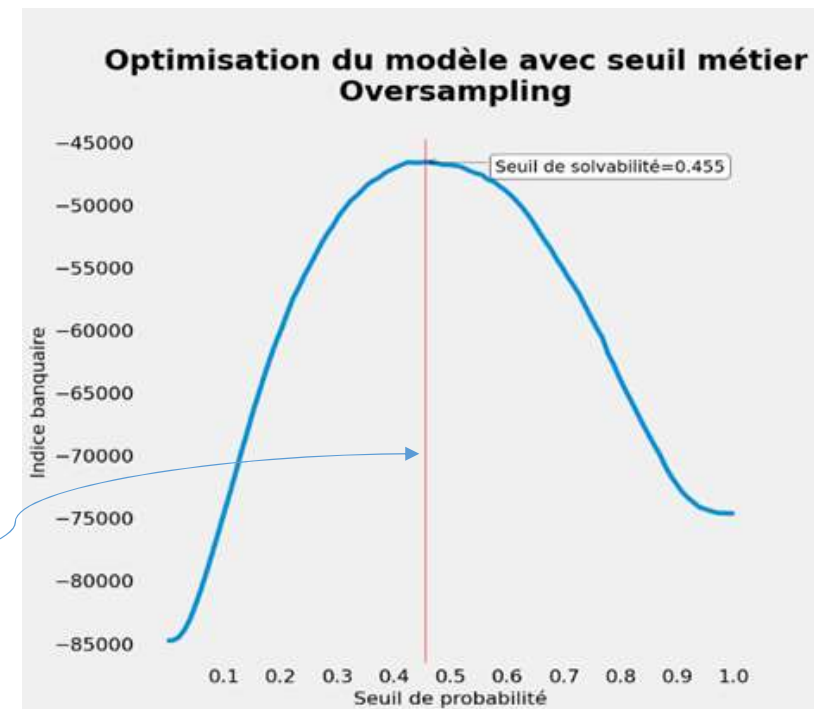
## Optimisation du modèle Fonction Coût

Limiter les **risques de perte financière** en pénalisant les Faux Négatifs et les Faux Positifs

### **Ind\_bank :**

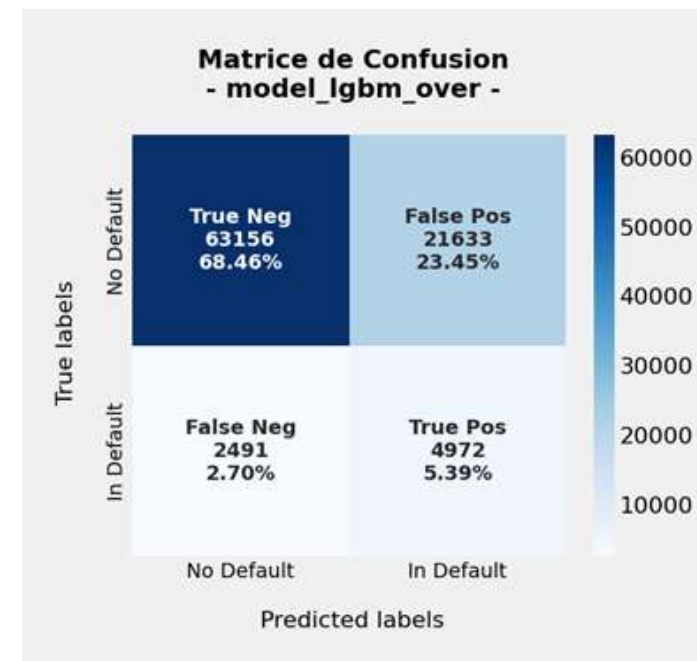
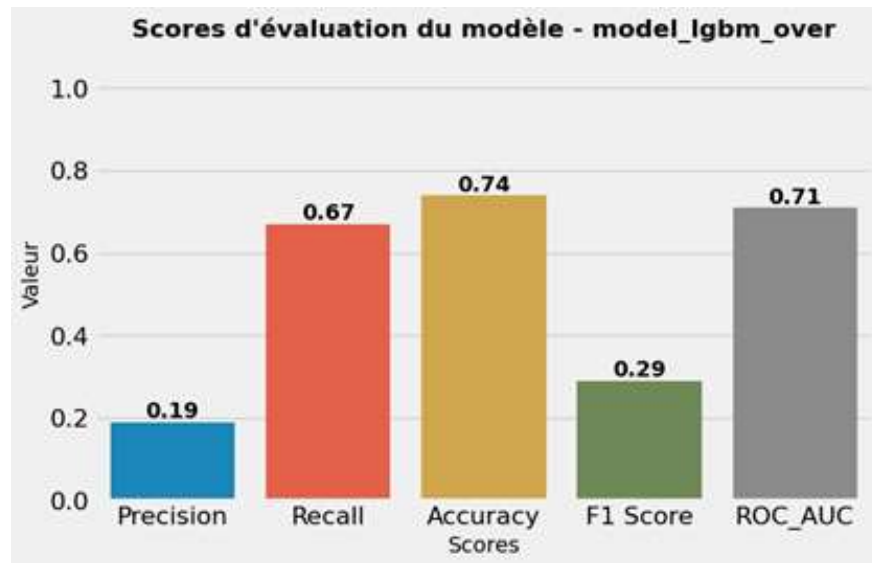
Fonction personnalisée qui permet d'évaluer les pertes financières potentielles Découlant de nos décisions de classification

**Seuil de solvabilité optimal**  
qui minimise l'indice bancaire



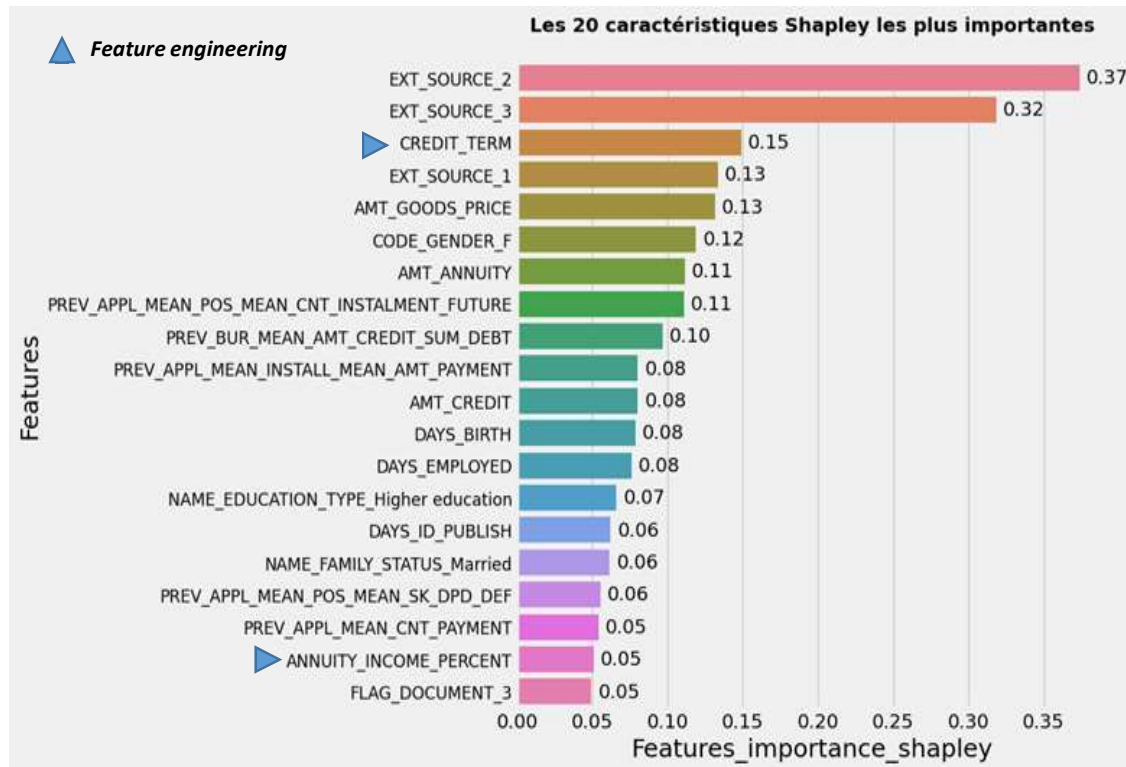
Optimisation du modèle avec seuil métier  
(fn\_value = 10 \* fp\_value)

### Scores après l'optimisation



Accuracy -----> **0,74**

## Interprétation des features



# Dashboard



Versioning GitHub :

<https://github.com/babi7777/scoring-model-credit-risk>

Application : <http://35.181.54.91:8501>

**Utilisateur**

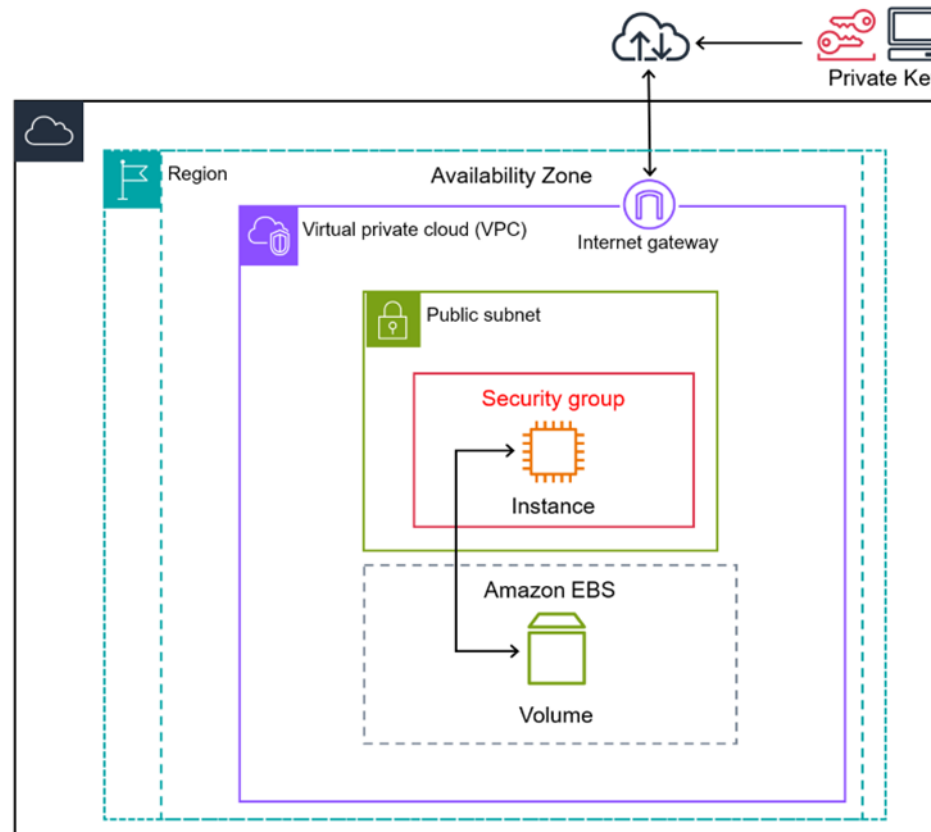
**Application Web**  
(Streamlit Dashboard)

**Modèle de Scoring**  
(Machine Learning)

**AWS**  
(Hébergement et Services Cloud)

**Mode d'Emploi de l'Application**





**Infrastructure de Déploiement du Modèle de Scoring :**  
Mise en Œuvre d'une Solution Évolutive pour la Prédiction de Crédit

## Conclusion

- Utilisation et modification d'un Kernel Kaggle.
- Entraînement d'un modèle de scoring.
- Fonction coût, optimisation et évaluation.
- Interprétabilité du modèle LightGBM avec Shapley
- Dashboard interactif.

MERCI