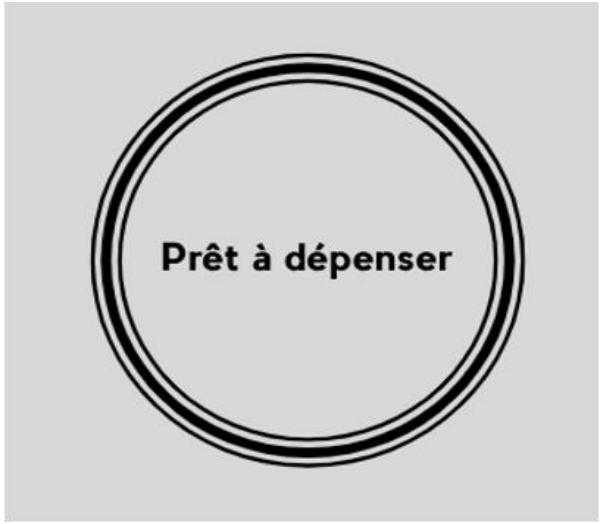
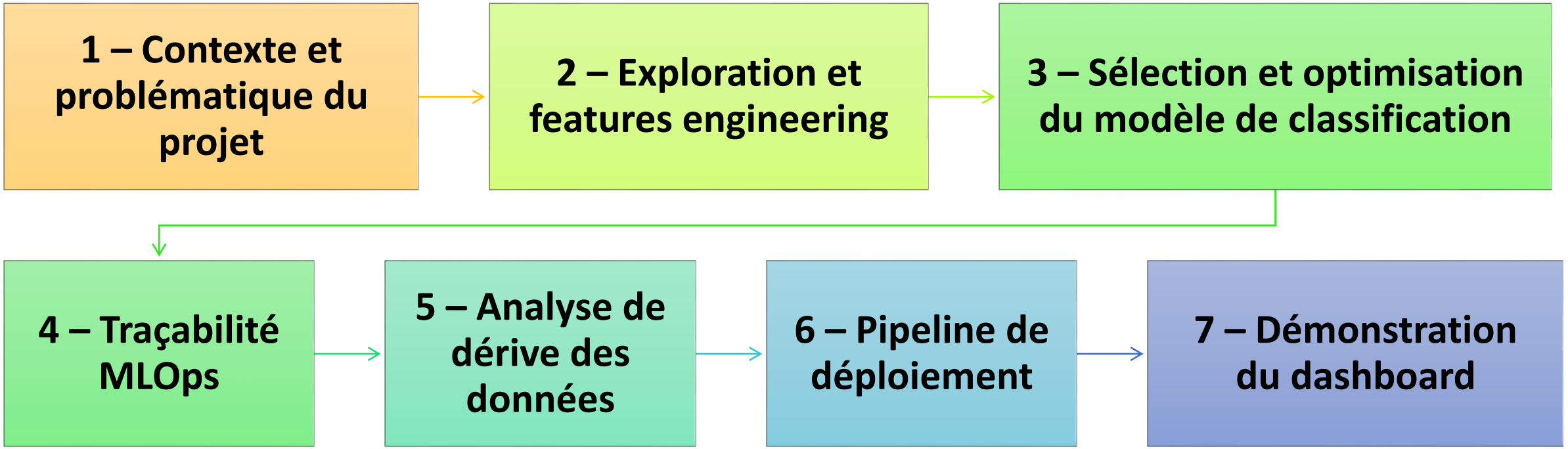


IMPLEMENTATION D'UN MODELE DE SCORING

Etude et Présentation: **Haïtem**

16/02/2024

PLAN DE LA PRESENTATION



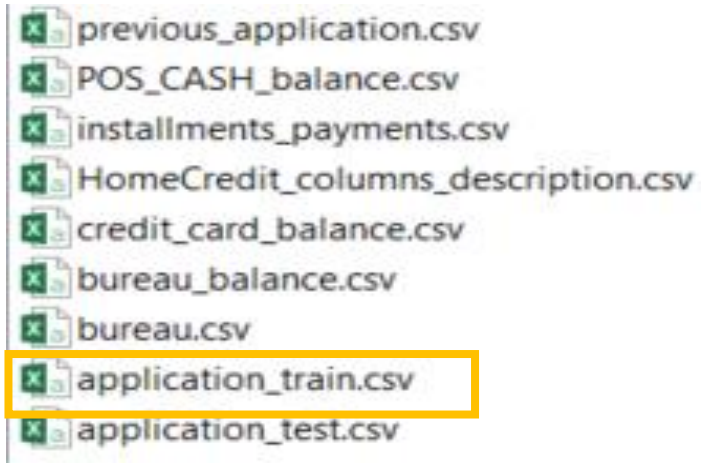
CONTEXTE ET PROBLEMATIQUE DU PROJET

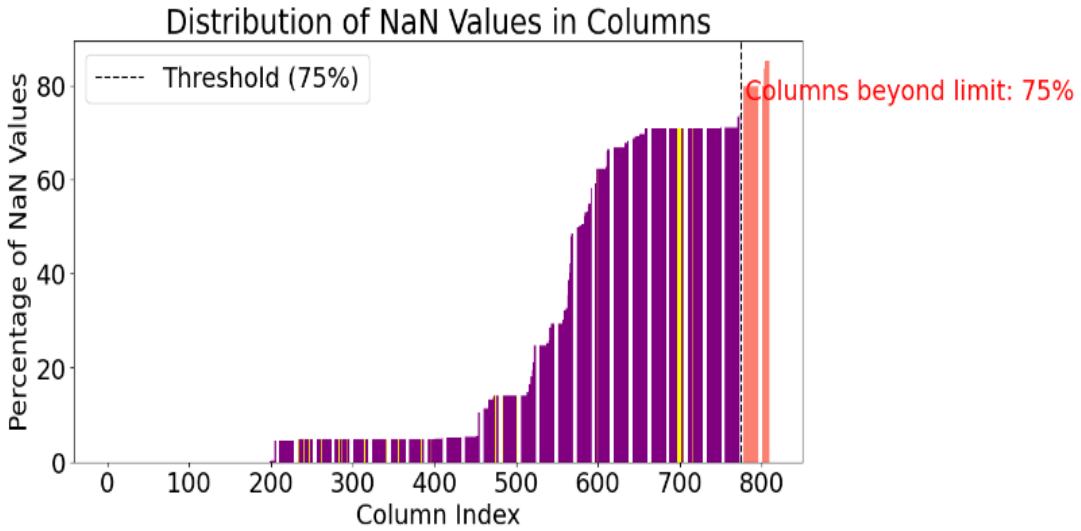
- **Contexte**: La société financière « Prêt à dépenser » souhaite offrir ses offres de crédit à une large clientèle.
- **Défi 1**: Clientèle avec un historique de crédit limité ou inexistant.
- **Défi 2:** Manque de transparence sur les motivations d'octroi ou de refus de demande de crédit.
- **Objectif** : Fournir aux conseillers financiers un outil interactif (un tableau de bord) permettant d'évaluer la solvabilité des demandes de crédit.
  - **Système de notation** visant à fournir une prédiction automatisée quant au remboursement ou au défaut d'un prêt.
  - **Des outils d'interprétation** de la décision générées par le modèle,
  - **Répondre au besoin de transparence** concernant les décisions de financement.

JEU DE DONNEES: EXPLORATION ET FEATURES ENGINEERING

Kernel Kaggle de J. Aguiar(https://www.kaggle.com/code/jsaguiar/lightgbm-with-simple-features/script)

	Traitement
Fusion des fichiers CSV	Fichiers de base: Application_tain/test
Variables catégorielles	Imputation des valeurs manquantes par la valeur la plus fréquente
	OneHotEncoding
Variables numériques	Agrégation: Min, Max, mean, sum
	Imputation des valeurs manquantes par la médiane
	Imputation des valeurs Infinies par +/- MAX(ABS)
	Standardisation
Colonnes avec 75% des valeurs nulles	Suppression
Colonnes avec des valeurs uniques	Suppression

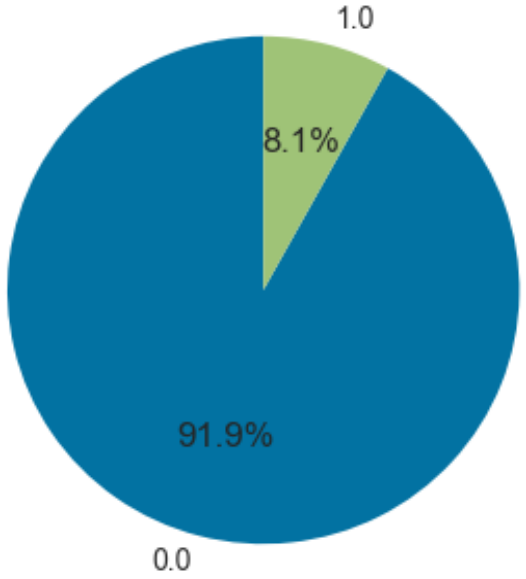




Taille du dataset d'entrainement: (307507, 741)

JEU DE DONNEES: EXPLORATION - VARIABLE CIBLE ‘TARGET’

Distribution of good (0) & faulty (1) clients in the train set



Classe 0

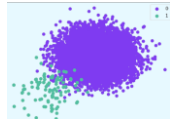
Clients solvables

Classe 1

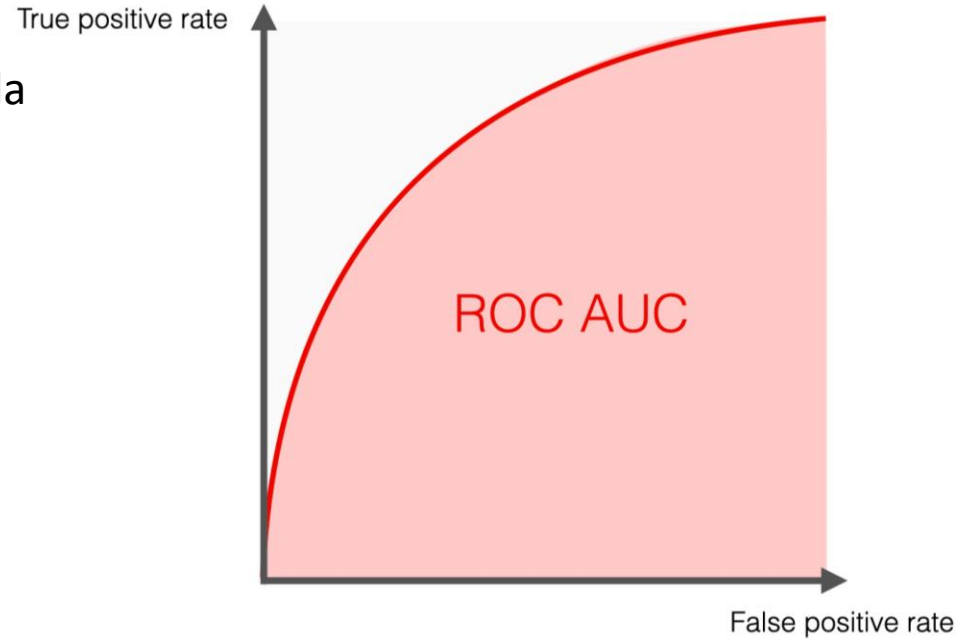
Clients insolvables

- **Effectif des clients : 307 507**

- **Problématique de déséquilibre de classes.**



MODELISATION: METRIQUES PERTINENTES



- **AUC-ROC** : L'aire sous la courbe ROC est une mesure synthétique de la performance d'un modèle de classification binaire

- **Matrice de confusion**: TN, TP, FN, FP

- **Fonction coût métier** $Coût\ métier = \frac{10FN + FP}{LEN(\text{Dataset d'entraînement})}$

TN: Prédiction réussie d'un client fiable	**0**	**-1**	**FP: Client fiable mal classé**
FN: Client non solvable et non identifié par le modèle de prédiction	**-10**	**0**	**TP: Client unifiable bien identifié**

CHOIX DU MODELE: PRESELECTION

Sélection préliminaire de 3 modèles candidats

Valider les modèles sélectionnés sur le dataset entier

Présélection de modèles candidats sur un échantillon de 60000 individus

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9198	0.7697	0.0317	0.5209	0.0596	0.0508	0.1128	426.751
lightgbm	Light Gradient Boosting Machine	0.9192	0.7645	0.0415	0.4683	0.0760	0.0639	0.1204	23.345
lda	Linear Discriminant Analysis	0.9156	0.7516	0.0832	0.3868	0.1368	0.1114	0.1492	14.641
ada	Ada Boost Classifier	0.9188	0.7514	0.0498	0.4588	0.0896	0.0750	0.1301	87.503
rf	Random Forest Classifier	0.9196	0.7055	0.0003	0.1000	0.0006	0.0005	0.0052	123.017
et	Extra Trees Classifier	0.9196	0.7005	0.0003	0.1000	0.0006	0.0004	0.0043	58.341
nb	Naive Bayes	0.8302	0.6778	0.2882	0.1829	0.2110	0.1282	0.1366	3.592
dt	Decision Tree Classifier	0.8525	0.5354	0.1576	0.1369	0.1465	0.0662	0.0664	56.784
qda	Quadratic Discriminant Analysis	0.2878	0.5119	0.7787	0.0828	0.1490	0.0051	0.0156	8.178
knn	K Neighbors Classifier	0.9163	0.5017	0.0071	0.1313	0.0134	0.0050	0.0121	11.510
dummy	Dummy Classifier	0.9196	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	2.937
lr	Logistic Regression	0.9196	0.4917	0.0000	0.0000	0.0000	0.0000	0.0000	3.916
ridge	Ridge Classifier	0.9196	0.0000	0.0036	0.4500	0.0070	0.0059	0.0337	3.866
svm	SVM - Linear Kernel	0.6567	0.0000	0.3139	0.0325	0.0566	0.0005	0.0002	27.206

PYCARET

Sélection finale sur l'ensemble du dataset

- **Light Gradient Boosting (LightGBM)**
- Linear Discimanant Analysis (LDA)
- Ada Boot (ADA)
- Dummy Classifier

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.9197	0.7793	0.0384	0.5414	0.0717	0.0616	0.1282	66.517
lda	Linear Discriminant Analysis	0.9175	0.7699	0.0624	0.4236	0.1087	0.0898	0.1380	47.333
ada	Ada Boost Classifier	0.9189	0.7613	0.0323	0.4666	0.0604	0.0505	0.1060	323.691
dummy	Dummy Classifier	0.9193	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	7.257

PYCARET

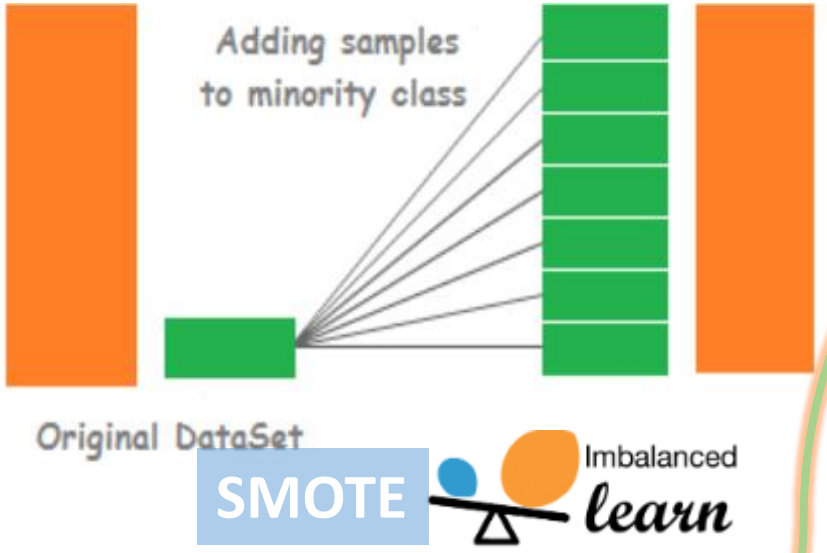
CHOIX DU MODELE: DESEQUILIBRE DE CLASSES

Sélection préliminaire de 3 modèles candidats → **Valider les modèles sélectionnés sur le dataset entier** → **Stratégie d'échantillonnage**

Problématique de déséquilibre de classes

Undersampling

Oversampling



SMOTE — Imbalanced learn

Method	Accuracy	Precision	Recall	F1_Score	AUC	Business_Cost
SMOTE	0.919645	0.534722	0.031036	0.058667	0.766572	0.520308
RandomUnderSampler	0.705408	0.173127	0.702136	0.277764	0.774056	0.508683
ImbalanceRatioAdjustment	0.733895	0.185007	0.674929	0.290409	0.780745	0.500000

Hyperparamètre LightGBM: scale_pos_weight

$$\frac{N\ classe\ négative}{N\ classe\ positive} = 11.385$$

scale_pos_weight = 11.385

CHOIX DU MODELE: OPTIMISATION

Sélection préliminaire de 3 modèles candidats → **Valider les modèles sélectionnés sur le dataset entier** → **Stratégie d'échantillonnage** → **Optimisation des hyperparamètres**

LightGBM + scale_pos_weight

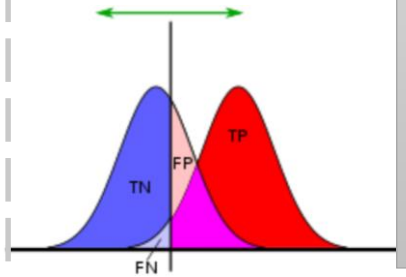
OPTUNA

- Optimisation bayésienne: Algorithme « Tree-Structured Parzen Estimator »
- Validation croisée: 5 folds (par défaut)
- Nombre d'itérations: 90
- Minimsation de la fonction coût

LightGBM

Dataset

Training | **Testing**

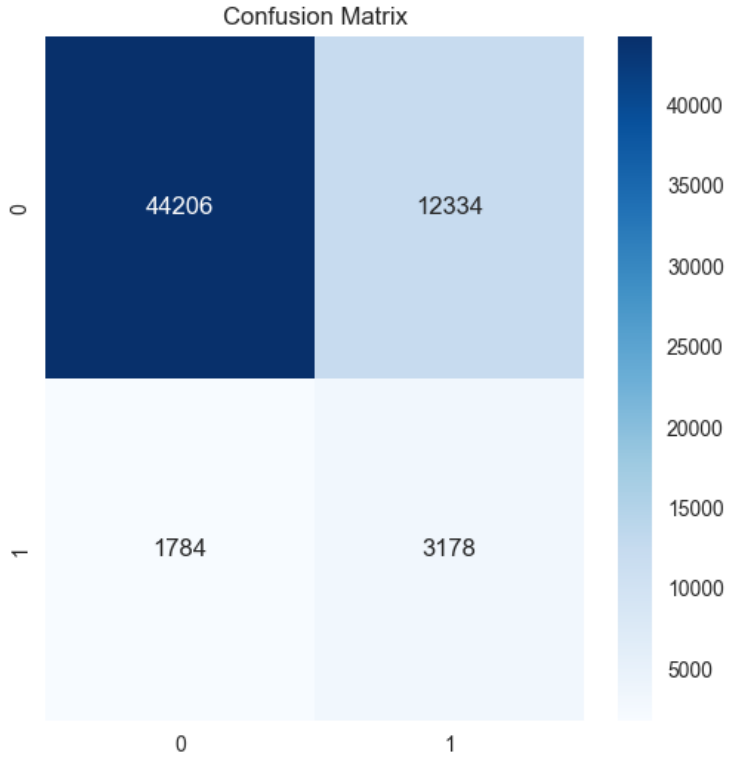


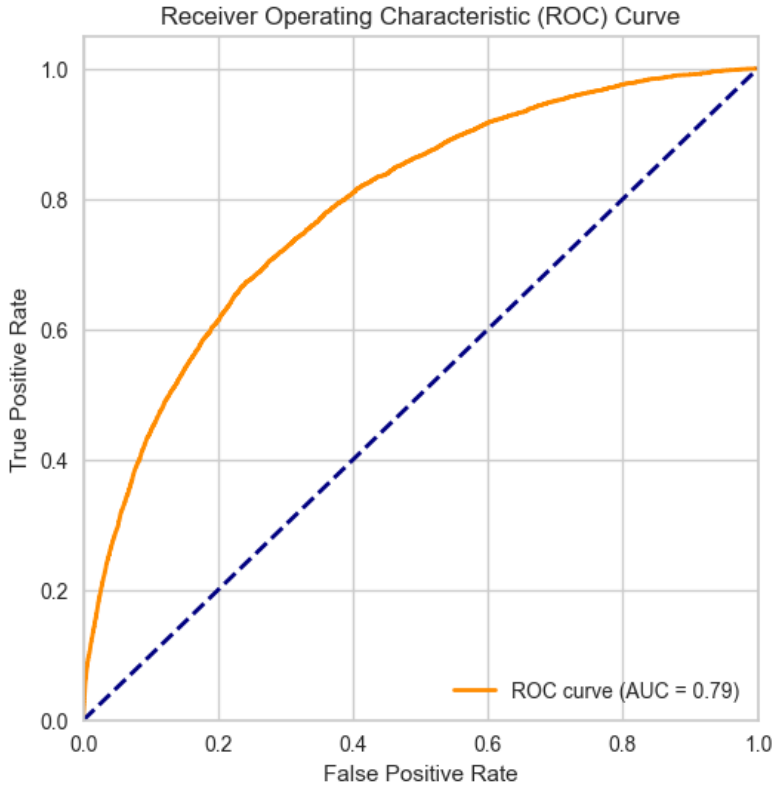
scale_pos_weight: $\frac{N\ classe\ négative}{N\ classe\ positive}$ **- 2,** $\frac{N\ classe\ négative}{N\ classe\ positive}$ **+ 2**
learning_rate: 0.001, 0.01
n_estimators: 100, 4000
min_child_samples: 2, 256
reg_alpha: 1e-8, 1e-3
reg_lambda: 1e-8, 1e-3

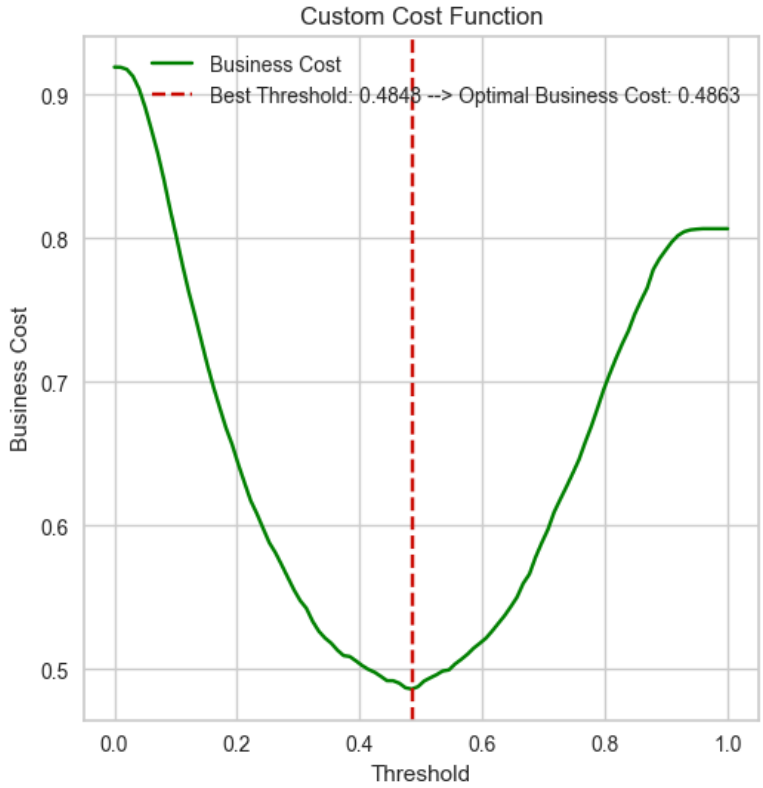
$\frac{N\ classe\ négative}{N\ classe\ positive}$ = 11.385

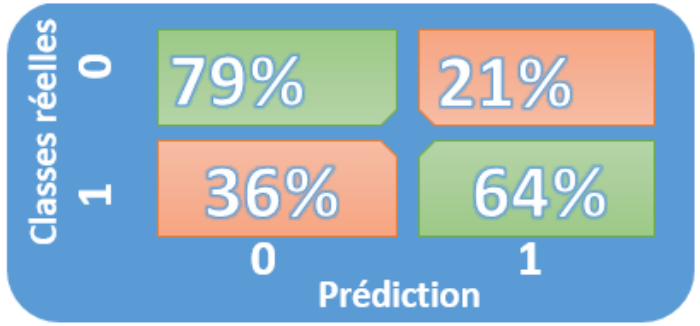
scale_pos_weight:	10.532
learning_rate :	0.0082
n_estimators:	3155
min_child_samples :	183
reg_alpha:	0.000153
reg_lambda:	6.8e-06

ANALYSE DES RESULTATS



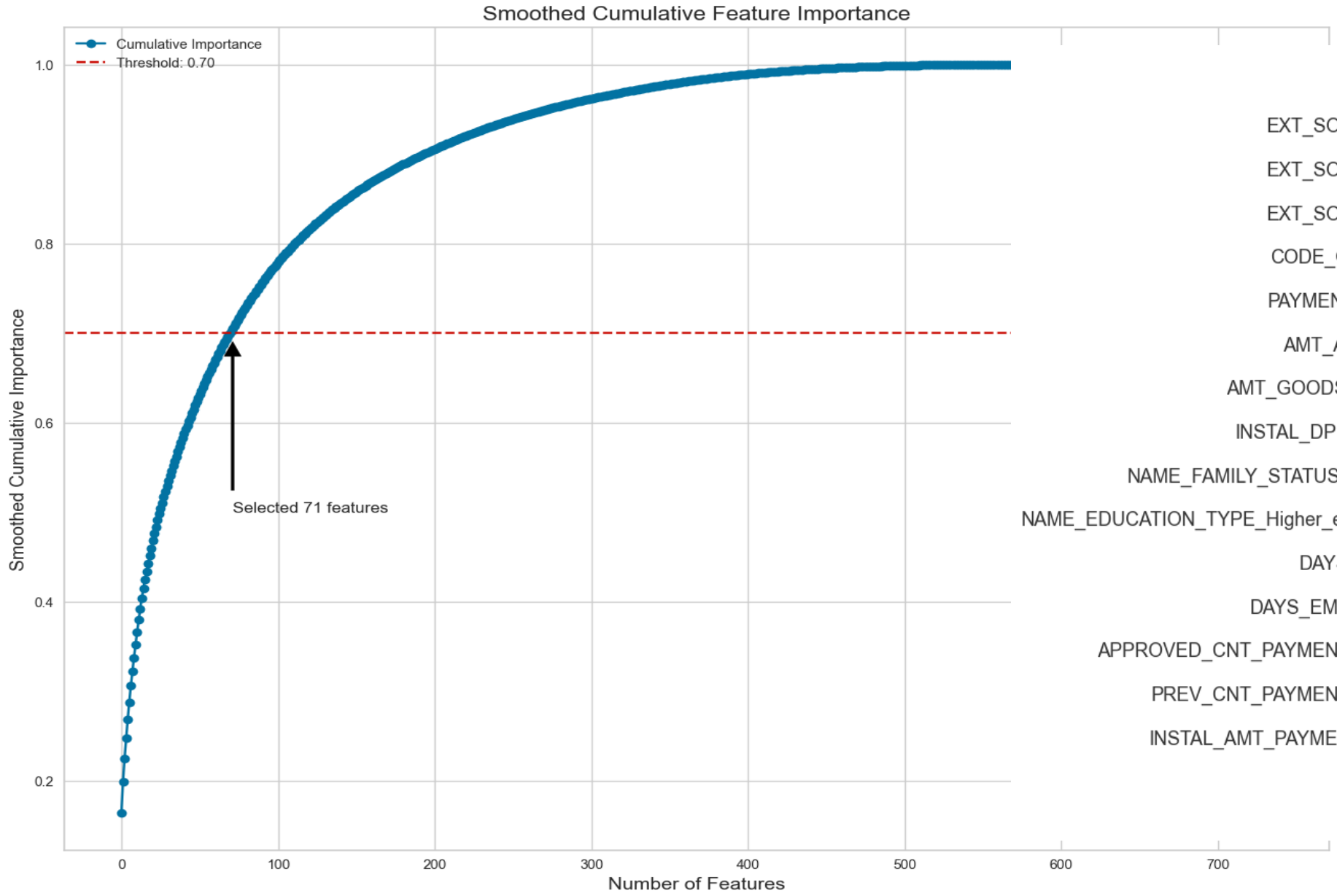


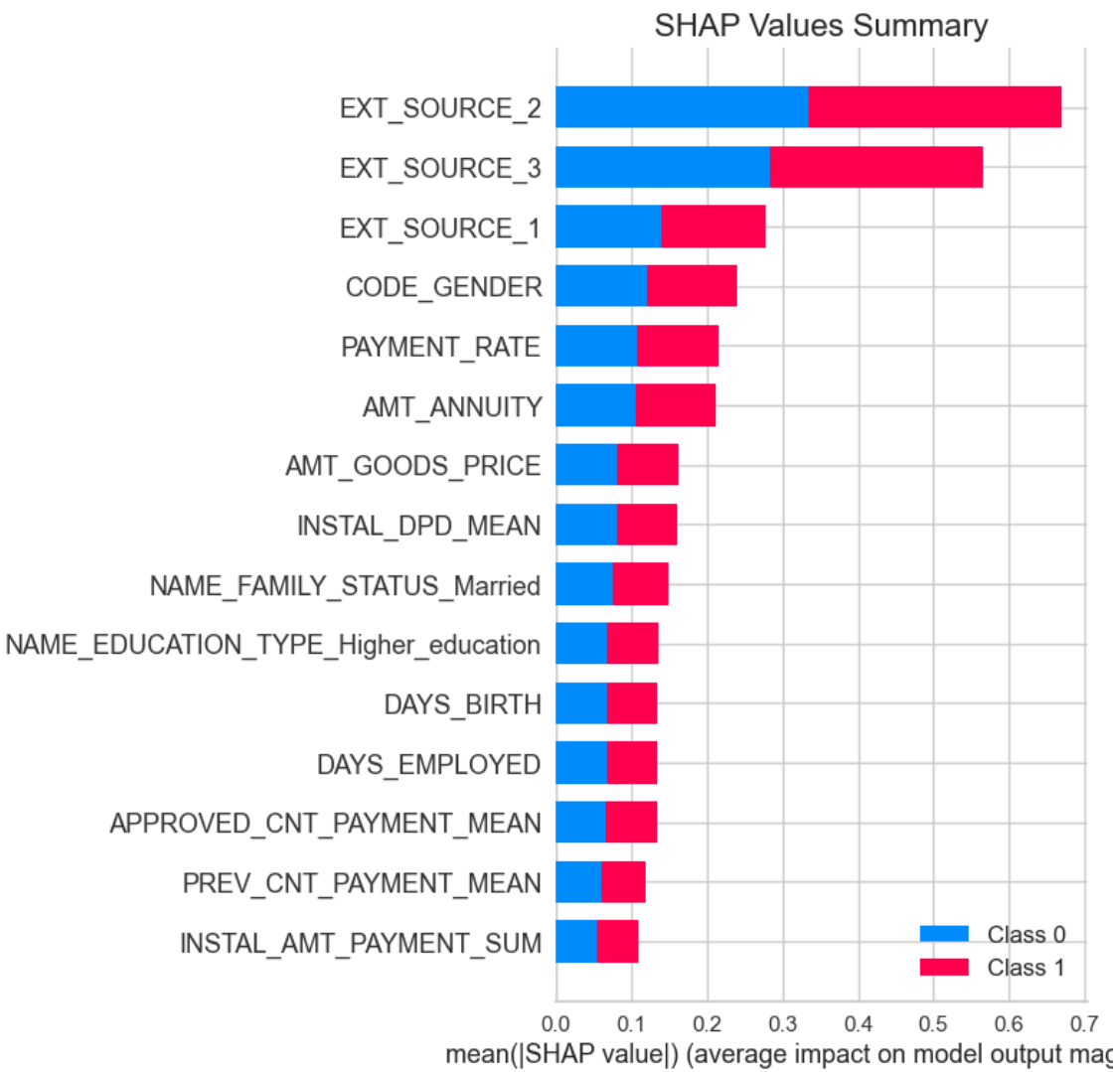




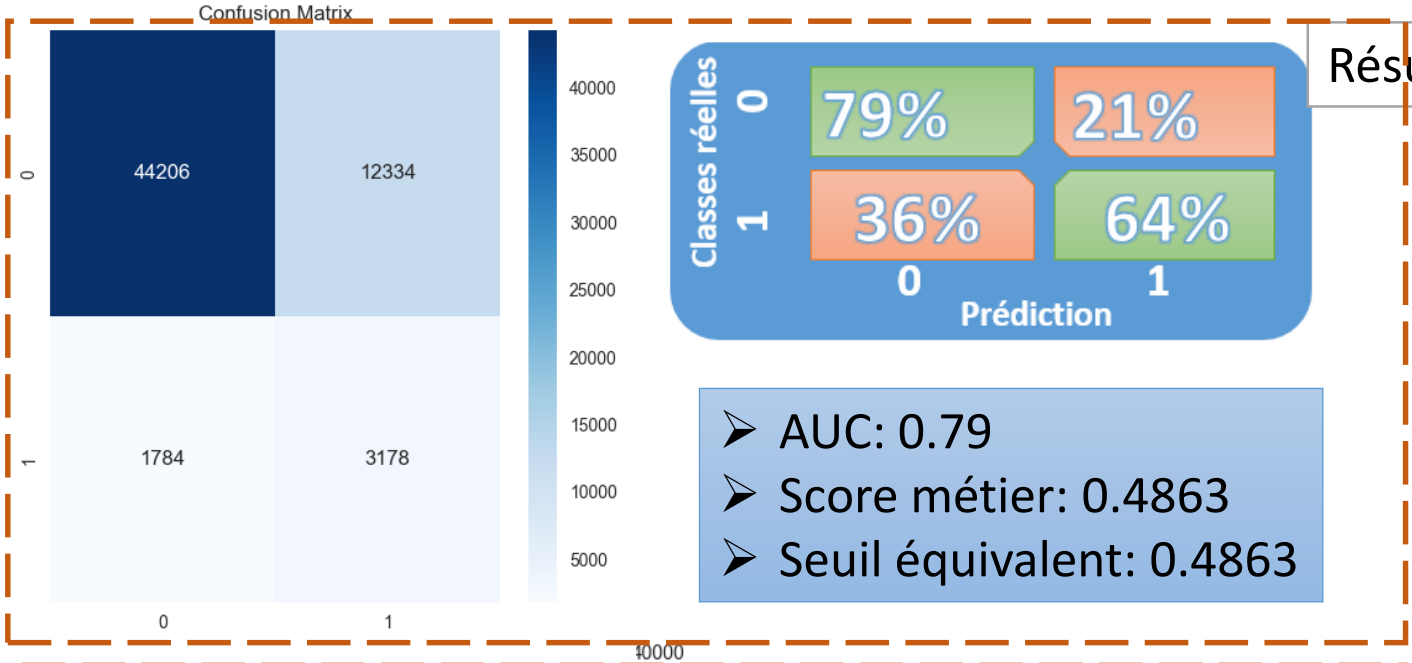
- AUC: 0.79
- Score métier: 0.4863 (Seuil équivalent): 0.4863)

ANALYSE DES RESULTATS



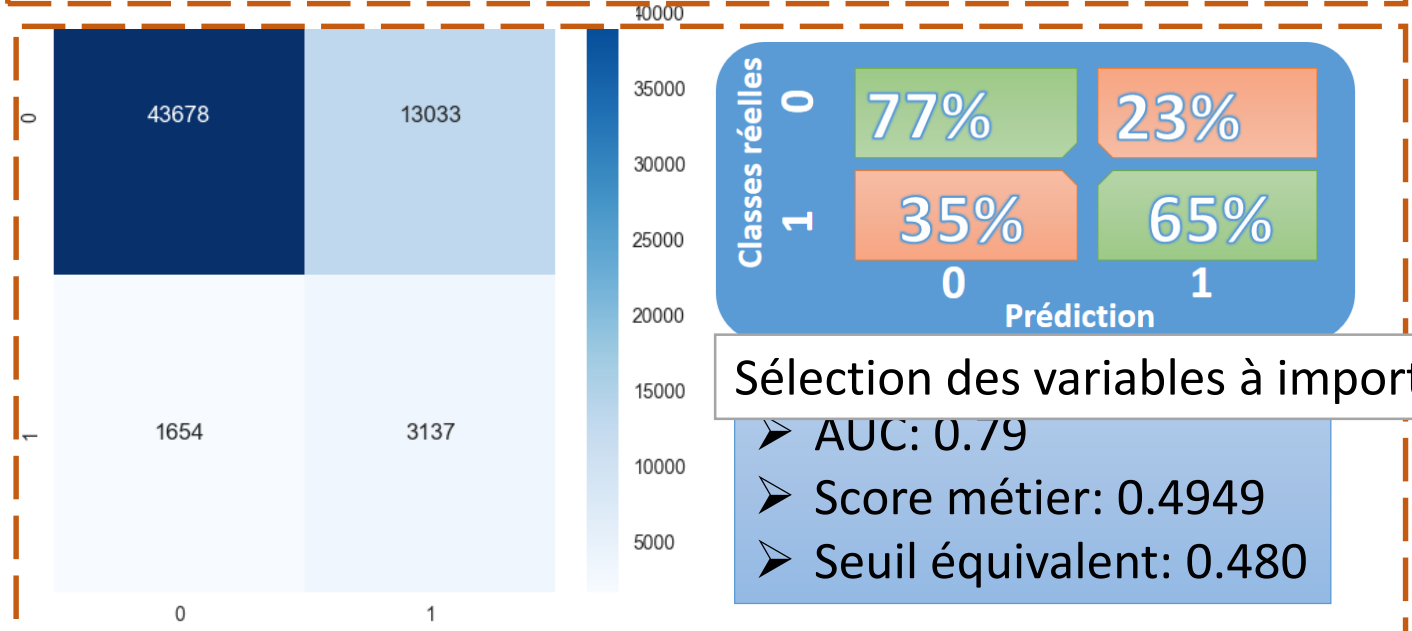


ANALYSE DES RESULTATS



Résultats initiales

- AUC: 0.79
- Score métier: 0.4863
- Seuil équivalent: 0.4863



Sélection des variables à importance cumulées à70%

- AUC: 0.79
- Score métier: 0.4949
- Seuil équivalent: 0.480

Métrique	Résultats Initiaux (741 variables)	Modèle Réduit (71 variables)
AUC	0.790	0.790
Score métier	0.486	0.495
Seuil de prédiction	0.486	0.480

TRACABILITE MLOPS

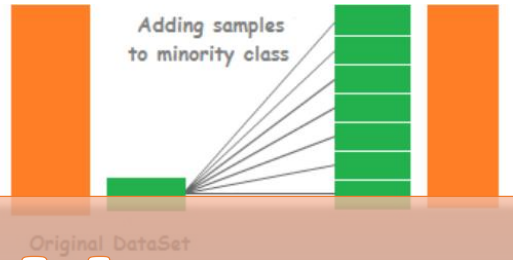
Sélection préliminaire de 3 modèles candidats → Valider les modèles sélectionnés sur le dataset entier → Stratégie d'échantillonnage → Optimisation des hyperparamètres

60000 individus

PYCARET

LightGBM, LDA, ADA

Oversampling



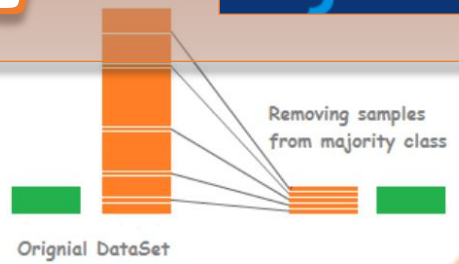
LightGBM + Weight Imbalance

MLFLOW

Undersan

mlflow

OPTUNA



- Light Gradient Boosting (LightGBM)
- Linear Discimanant Analysis (LDA)
- Ada Boot (ADA)

LightGBM

Hyperparamètre LightGBM:
Weight Imbalance

$$\frac{N\ classe\ négative}{N\ classe\ positive}$$

LightGBM

ANALYSE DE LA DERIVE DES DONNEES: BIBLIOTHEQUE

EVIDENTLY AI

Prêt à dépenser

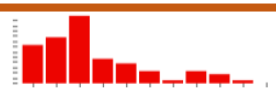
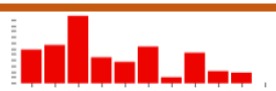
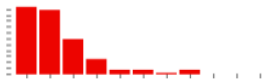
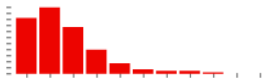
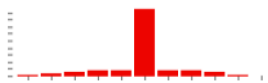
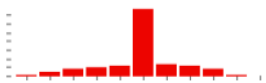
Analyse globale

Dataset Drift

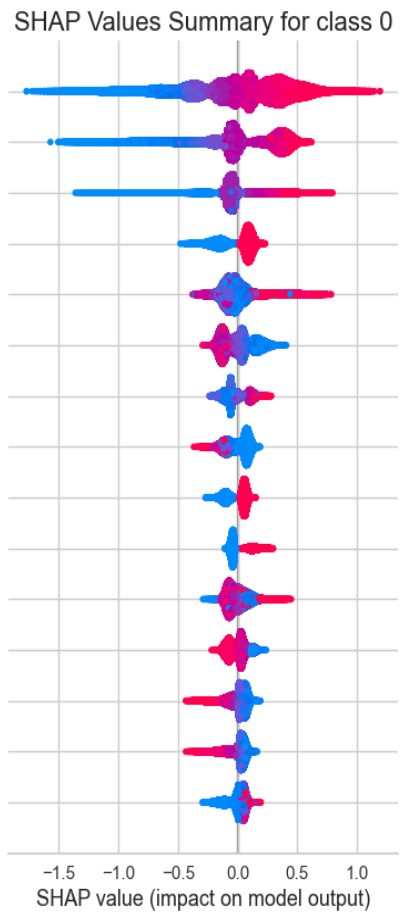
Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5

71	12	0.169
Columns	Drifted Columns	Share of Drifted Columns

Analyse des variables

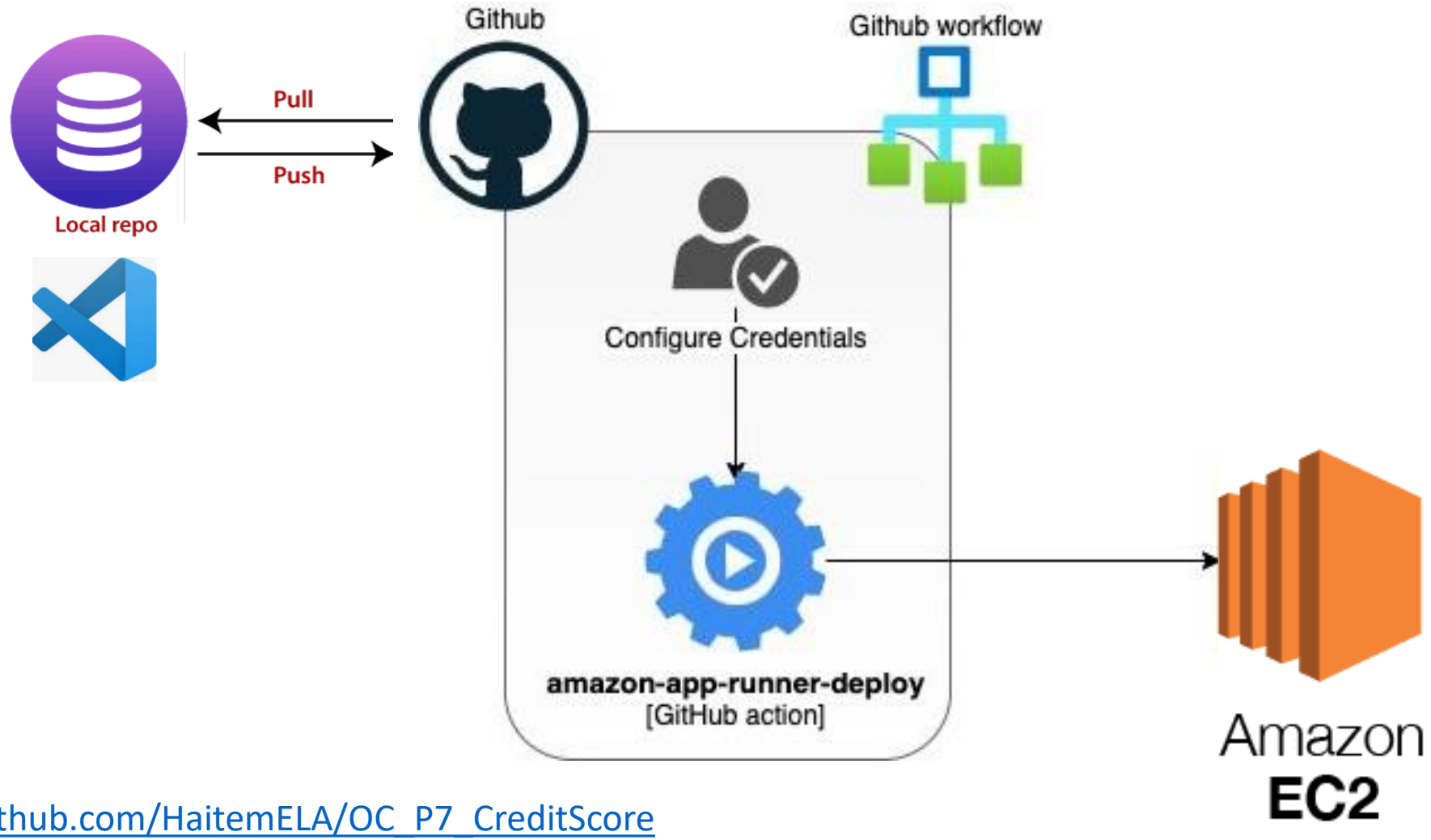
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
PAYMENT_RATE	num			Detected	Wasserstein distance (normed)	0.584121
INCOME_CREDIT_PERC	num			Detected	Wasserstein distance (normed)	0.26[illegible]504
AMT_GOODS_PRICE	num			Detected	Wasserstein distance (normed)	0.21[illegible]75
AMT_CREDIT	num			Detected	Wasserstein distance (normed)	0.206782
AMT_REQ_CREDIT_BUREAU_QRT	num			Detected	Wasserstein distance (normed)	0.179174
AMT_ANNUITY	num			Detected	Wasserstein distance (normed)	0.164308
EXT_SOURCE_1	num			Detected	Wasserstein distance (normed)	0.157793

EXT_SOURCE_2
EXT_SOURCE_3
EXT_SOURCE_1
CODE_GENDER
PAYMENT_RATE
AMT_ANNUITY
AMT_GOODS_PRICE
INSTAL_DPD_MEAN
AMILY_STATUS_Married
_TYPE_Higher_education
DAYS_BIRTH
DAYS_EMPLOYED
_CNT_PAYMENT_MEAN
_CNT_PAYMENT_MEAN
L_AMT_PAYMENT_SUM



Drift is detected for 16.901% of columns (12 out of 71).

Pipeline de déploiement



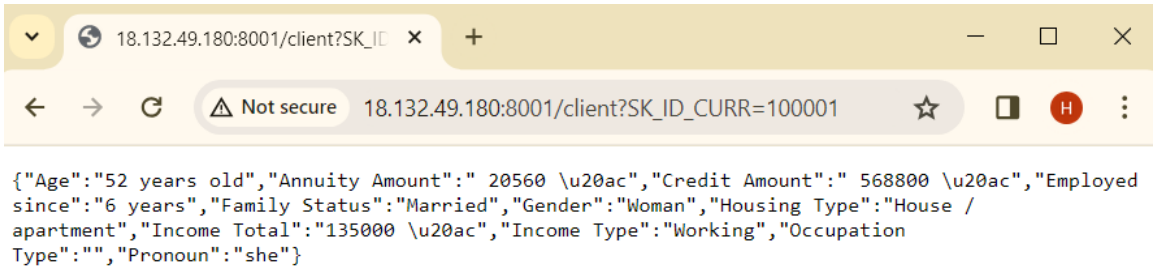
https://github.com/HaitemELA/OC_P7_CreditScore

SOLUTION FINALE

API

Flask

- Requêtes HTTP: POST - GET
- Accéder aux données des clients et aux prédictions
- Accéder à l'interprétabilité globale et locale

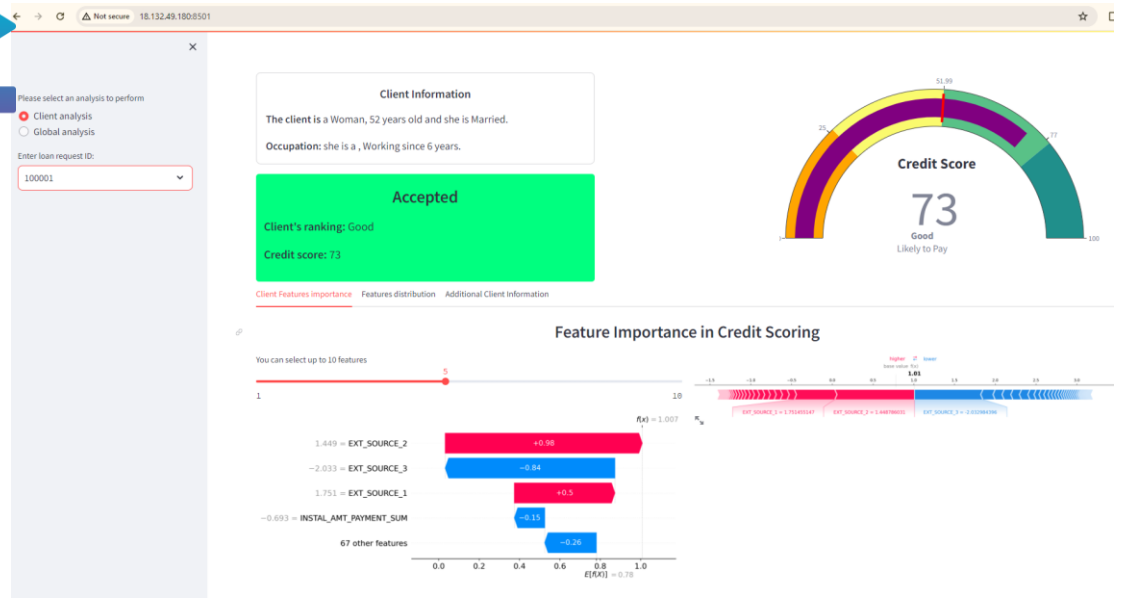


http://18.132.49.180:8001/client?SK_ID_CURR=100001

DASHBOARD

Streamlit

- Visualisation du score du client et de la décision de l'accord ou rejet de la demande de crédit.
- Visualisation des informations du client.
- Analyse des donnée du client.



http://18.132.49.180:8501/

Bilan et perspectives

1. Modélisation: LGBMCalssifier / Scale_pos_weight
2. La solution (API / Dashboard) est bien est déployée sur AWS EC2. Le pipeline de déploiement continue est également fonctionnel
3. Pour une meilleure performance:
   - Les données des clients devraient être stockées dans des bases de données
   - Optimiser le dashboard pour avoir une solution rapide à l'exécution.
   - Sécuriser l'accès au dashboard
4. La collaboration avec l'Équipe Financière est essentielle pour définir précisément les variables métiers pertinents.
5. Analyse des Risques pourrait être nécessaire à l'approche prédictive pour identifier les facteurs de risque sous-jacents et ajuster la métrique coût métier
6. Optimisation de la transparence : Documenter certaines variables importante à l'interprétation (EXT_SCOURCE_1, EXT_SCOURCE_2, et EXT_SCOURCE_3).

Prêt à dépenser

Merci pour votre attention !