A series of thin, black, overlapping lines forming various geometric shapes like triangles and polygons, creating a complex, abstract pattern in the upper left portion of the slide.

PROJET 7 OPENCLASSROOMS : IMPLÉMENTEZ UN MODÈLE DE SCORING

Dabidin Keshika

26/08/2023



SOMMAIRE

1. Introduction
2. Prétraitement des données
3. La méthodologie d'entraînement du modèle
4. La fonction coût métier et l'algorithme d'optimisation
5. Tableau de synthèse des résultats
6. L'interprétabilité du modèle
7. Pipeline de déploiement
8. L'analyse du Data Drift
9. Conclusion

INTRODUCTION

ETUDE D'UN MODÈLE DE SCORING

- **Prêt à dépenser** souhaite développer un modèle de Scoring de la probabilité de défaut de paiement du client pour étayer la décision **d'accorder ou non un prêt à un client potentiel**.
- Proposition avec 3 modèles de **machine learning de gradient boosting** (LightBoost Classifier, XGBoost Classifier et CatBoost Classifier)

DEMANDES DU MANAGER

- Partir d'un **kernel Kaggle** pour faciliter l'étude et la préparation des données.
- Les données sont récupérables sur le lien suivant :
[Home Credit Default Risk | Kaggle](#)
- Réaliser une **note méthodologique** expliquant en détails la construction du modèle.
- **Déploiement** du dashboard sur le Cloud.

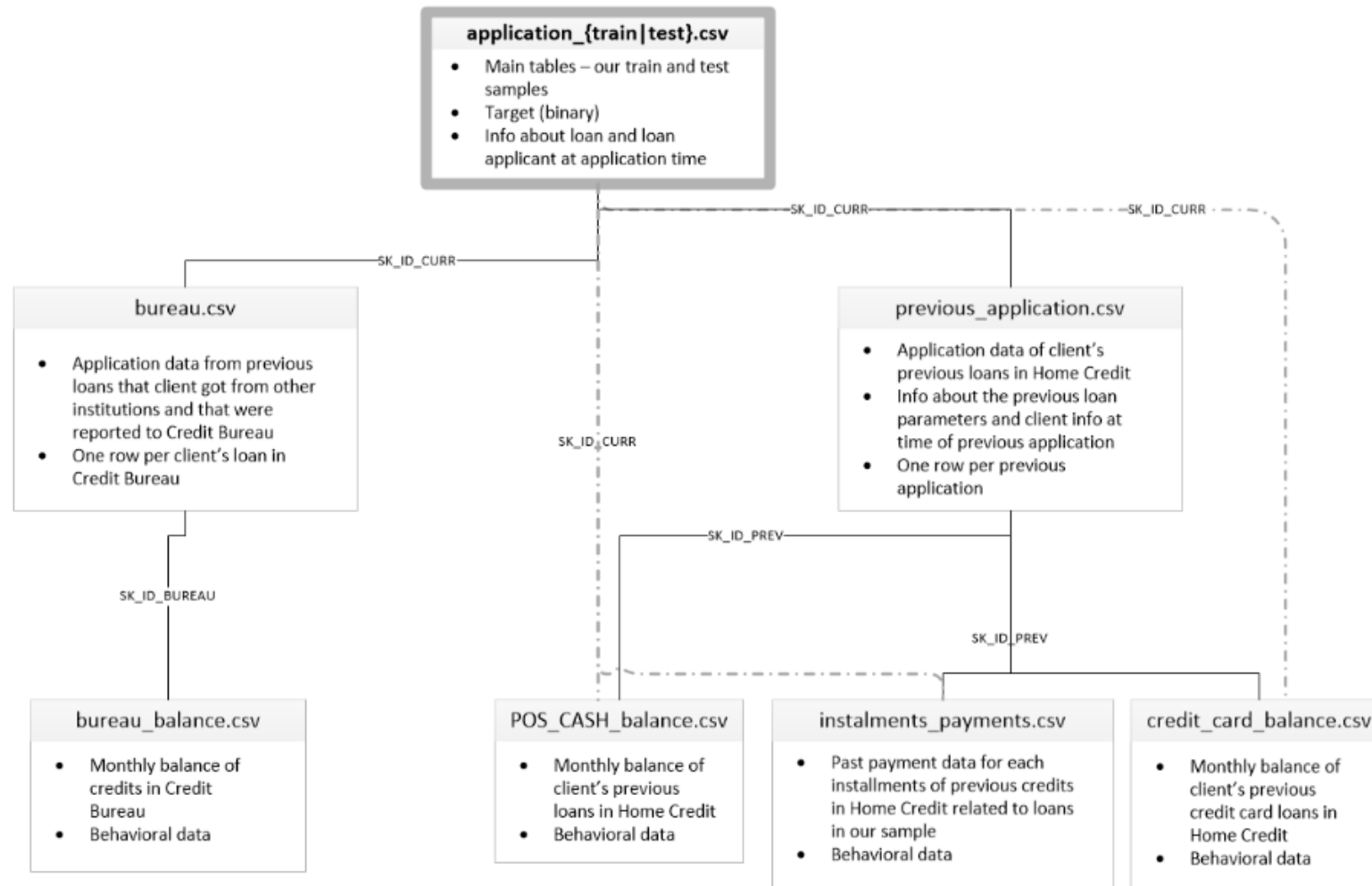
DÉVELOPPEMENT D'UN DASHBOARD

- Développement d'un **Dashboard interactif** pour que les chargés de relation client puissent à la fois **expliquer de façon la plus transparente possible les décisions d'octroi de crédit**.

Le dashboard doit permettre de :

- Visualiser le **score** pour chaque client.
- Visualiser des **informations descriptives** relatives à un client.
- **Comparer les informations** descriptives relatives à un client à l'ensemble des clients ou à un **groupe de clients similaires**.

PRÉSENTATION DES DONNÉES



Application train :

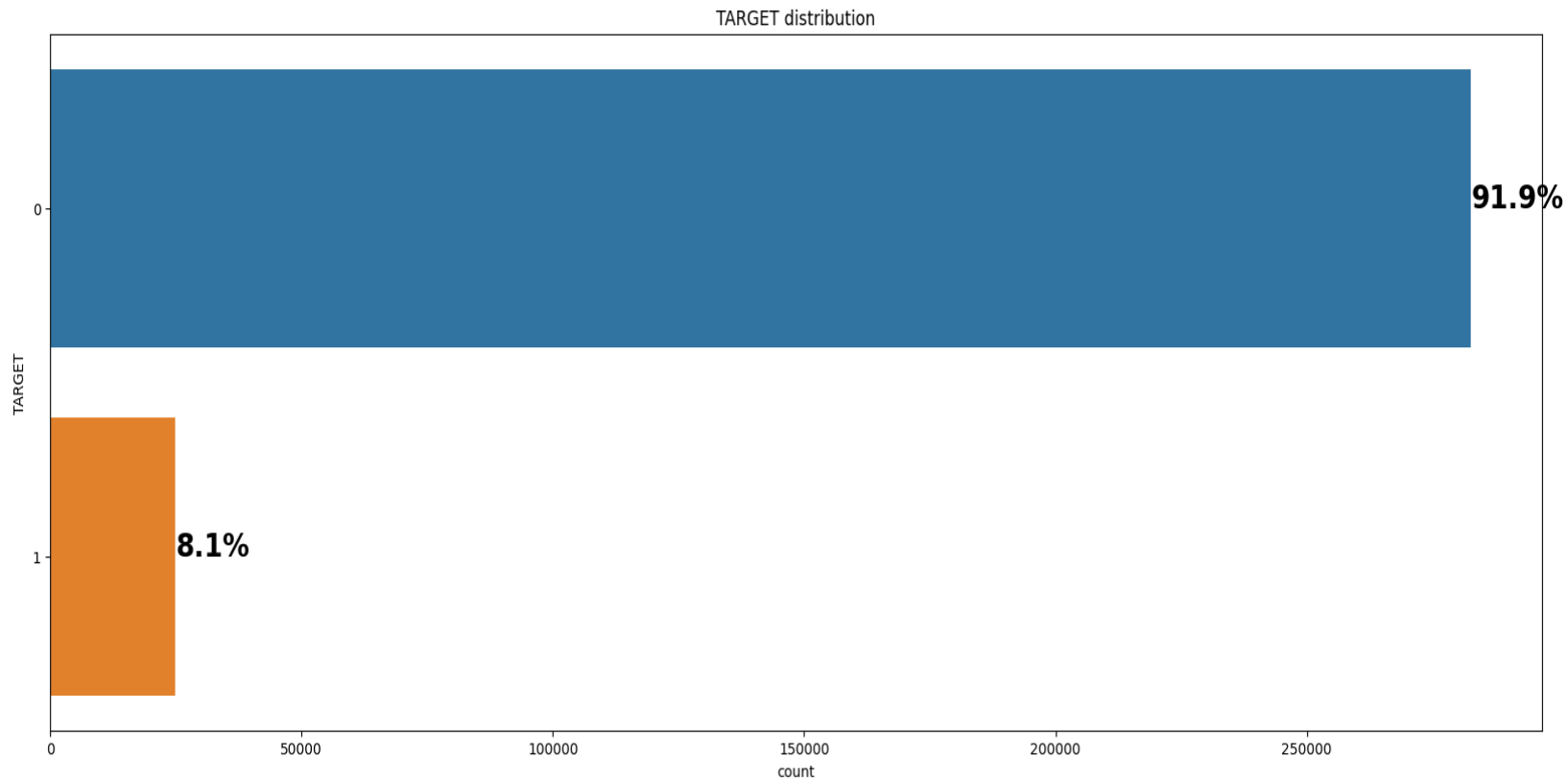
307511 lignes
122 colonnes

Application test :

48744 lignes
121 colonnes

Pas de target

ANALYSE EXPLORATOIRE DES DONNÉES

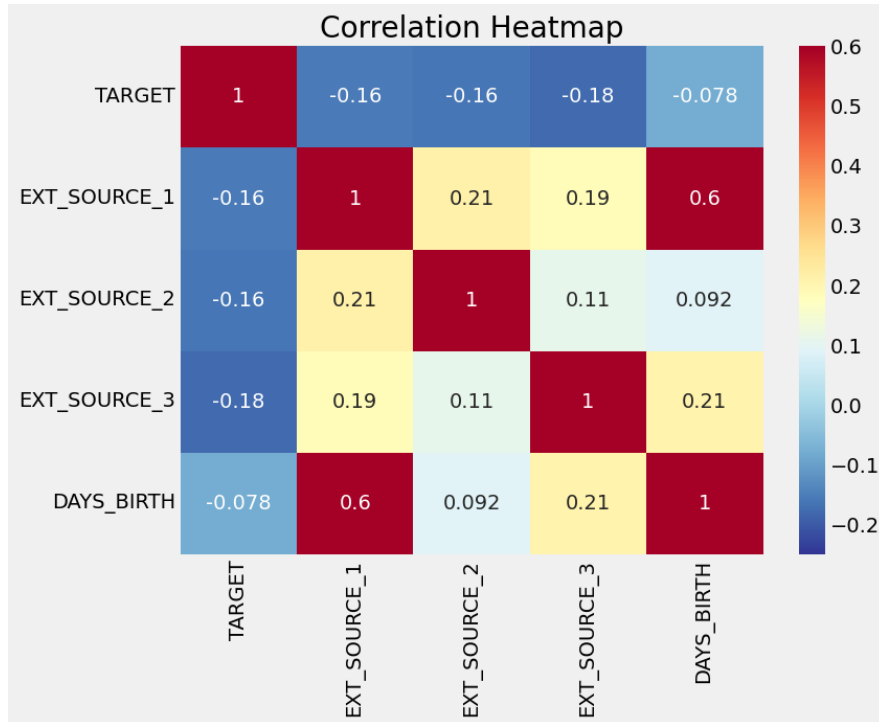


- Le target est 1 si l'individu est éligible pour le prêt, 0 sinon.
- Déséquilibre dans les données (SMOTE ou Sample Weights pour équilibrer).

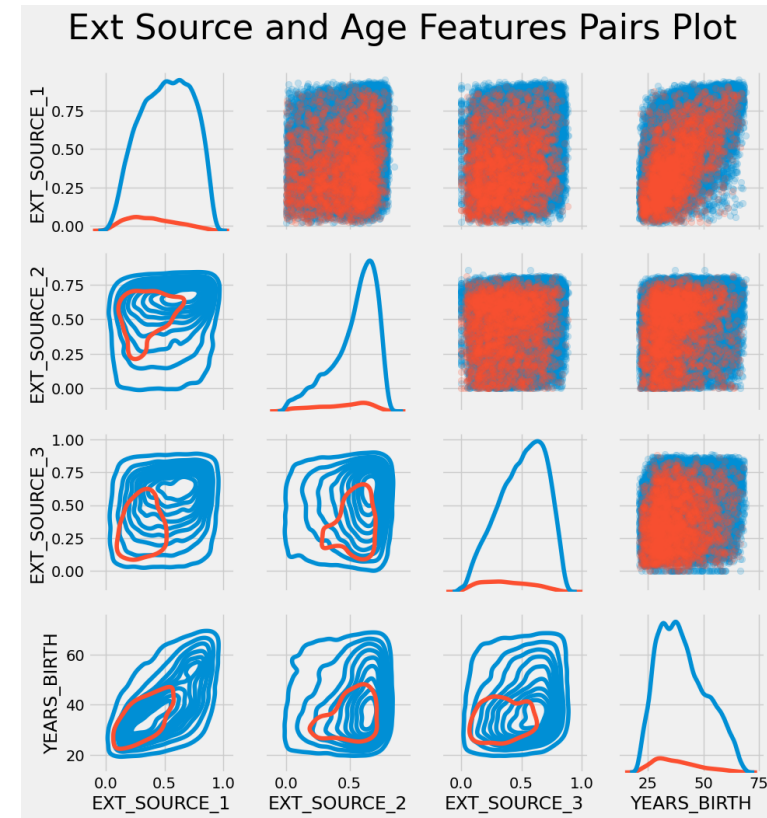
ANALYSE EXPLORATOIRE DES DONNÉES

1. Les valeurs manquantes sont plus présentes dans les caractéristiques des habitats.
2. Les prêts renouvelables ne représentent qu'une petite fraction (10%) du nombre total de prêts
3. En termes de pourcentage de non-remboursement du prêt, le mariage civil a le pourcentage le plus élevé de non-remboursement (10%), la veuve étant le plus bas.
4. Les demandeurs avec le type de revenu Congé de maternité ont un ratio de près de 40% de prêts non remboursés, suivis des chômeurs (37%). Les autres types de revenus sont inférieurs à la moyenne de 10% pour ne pas rembourser les prêts.
5. La plupart des prêts sont contractés par des ouvriers, suivis par les vendeurs/commerciaux. Le personnel informatique prend le montant de prêts le plus bas. La catégorie avec le pourcentage le plus élevé de prêts non remboursés est celle des ouvriers peu qualifiés (plus de 17%), suivis des chauffeurs et des serveurs / barmen, du personnel de sécurité, des ouvriers et du personnel de cuisine.
6. Les loueurs d'appartements (non propriétaires de leur résidence principale), ainsi que ceux qui vivent chez leurs parents, ont un taux de non-remboursement supérieur à 10%.

ANALYSE EXPLORATOIRE DES DONNÉES



- EXT_SOURCE ont des corrélations négatives avec la cible
- DAYS_BIRTH est positivement corrélé avec EXT_SOURCE_1 indiquant que l'un des facteurs de ce score est peut-être l'âge du client.



- Le rouge indique les prêts qui n'ont pas été remboursés et le bleu les prêts qui sont payés.
- Relation linéaire positive modérée entre EXT_SOURCE_1 et DAYS_BIRTH

PRÉTRAITEMENT DES DONNÉES

- 1, Les données ont été fusionnées et retravaillées à partir des fichiers train, test, fichiers bureaux, fichiers liés au cash balance, etc.
- 2, Ajout de nouvelles variables pertinentes telles que :
 - PREVIOUS_LOANS_COUNT de bureau.csv qui indique le nombre total des précédents crédits pris par chaque client.
 - CREDIT_INCOME_PERCENT : Pourcentage du montant du crédit par rapport au revenu d'un client.
- 3, Splitting en Train(80%) et Test(20%)
- 4, Encodage des features catégorielles avec le label encoder et get_dummies.
- 5, Imputation des valeurs manquantes par la moyenne ou la médiane et traitement des valeurs aberrantes.
- 6, Standardisation des données avec le MinMaxScaler.
- 7, Une version de données avec les valeurs manquantes a également été conservée car certains algorithmes tels que le light gradient boost peuvent effectuer des prédictions même avec des valeurs manquantes.

TRAITEMENT DU DÉSÉQUILIBRE DES CLASSES

1, SMOTE (Synthetic Minority Oversampling Technique) :

SMOTE permet de créer des données synthétiques à partir des données existantes.

	Classe 0	Classe 1
Avant Rééquilibrage	197845	17412
Après Rééquilibrage	197845	197845

2, Sample Weights (scale_pos_weight X) permet de modifier les poids associés aux observations de sorte qu'une observation mal classée dans la classe minoritaire pénalise davantage la fonction de perte qu'une observation mal classée dans la classe majoritaire.

ENTRAÎNEMENT DU MODÈLE

1, Classification binaire

La classification consiste à identifier les classes d'appartenance de nouveaux objets à partir d'exemples antérieurs connus. Dans le contexte métier du projet, la classification est binaire représentée par une variable de sortie à deux classes, à savoir acceptation du crédit (1) ou refus du crédit (0).

2, Modèles utilisés

Le modèle de baseline (Régression Logistique) ainsi que trois algorithmes de gradient boosting (CatBoost, XGBoost et LightBoost) ont été testés sous différentes conditions :

- sans équilibrage de données,
- avec équilibrage en utilisant SMOTE et,
- avec équilibrage en testant plusieurs sample weights.

ENTRAÎNEMENT DU MODÈLE

3, Suivi avec MLFlow

```
with mlflow.start_run():

    trained_model=model.fit(X_train, y_train)
    predicted_qualities=trained_model.predict(X_test)

    #Evaluation metrics
    (rmse,mae,r2)=eval_metrics(y_test,predicted_qualities)
    accuracy=accuracy_score(y_test, predicted_qualities)
    precision=precision_score(y_test,predicted_qualities)
    recall=recall_score(y_test,predicted_qualities)
    f_one=metrics.f1_score(y_test,predicted_qualities)

    #print("Model used:" % str(model))
    print(" RMSE: %s" % rmse)
    print(" MAE:%s" % mae)
    print(" R2:%s" % r2)
    print(" Accuracy: %s" % accuracy)
    print(" Precision:%s" % precision)
    print(" Recall:%s" % recall)
    print(" F-1 Score: %s" % f_one)

    roc_auc = roc_auc_score(y_test, trained_model.predict_proba(X_test)[:,:1])
    print('AUC : %0.4f' %roc_auc)
    print(classification_report(y_test, trained_model.predict(X_test)))
    cf_matrix_roc_auc(model, y_test, trained_model.predict(X_test), trained_model)

    mlflow.log_metric("rmse",rmse)
    mlflow.log_metric("r2",r2)
    mlflow.log_metric("mae",mae)
    mlflow.log_metric("accuracy",accuracy)
    mlflow.log_metric("precision",precision)
    mlflow.log_metric("recall",recall)
    mlflow.log_metric("F-1 score",f_one)
    mlflow.log_metric("AUC",roc_auc)

    mlflow.sklearn.log_model(model,"model")
```

Gradient-Boosting-models > LightBoost_baseline_with_smote >

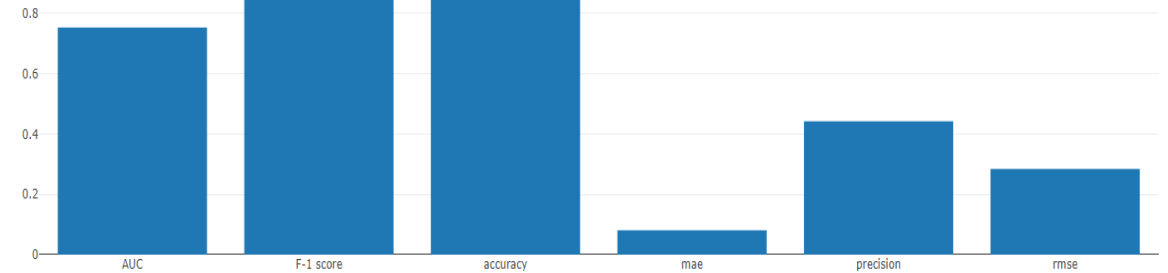
Metrics

Y-axis:

AUC X F-1 score X accuracy X mae X precision X
rmse X

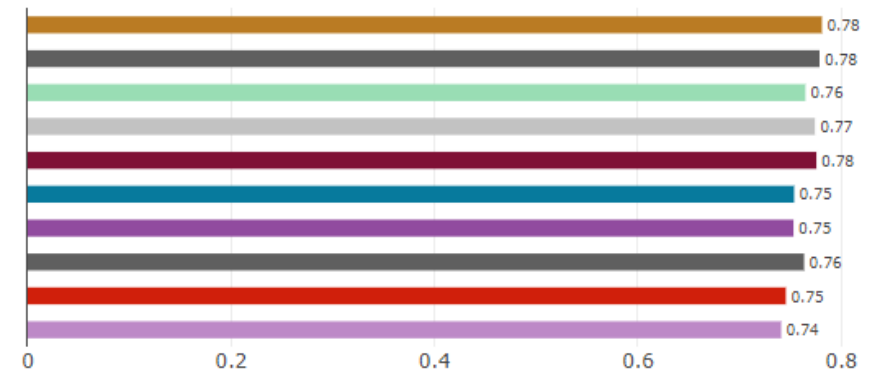
Y-axis Log Scale: ☐

Download CSV



AUC

Comparing first 10 runs



ENTRAÎNEMENT DU MODÈLE

3, Suivi avec MLFlow

mlflow2.4.1

ExperimentsModels

Experiments

+

-

Search Experiments

☐ Default

☒ Optimized_lightboost_with_rfe_and_custom...

☒ Recursive_feature_selection_experiments

☒ Gradient-Boosting-models

☒ Linear-Regression-SMOTE

☒ Linear-Regression

Displaying Runs from 5 Experiments

Table viewChart viewArtifact view

metrics.rmse < 1 and params.model = "tree"

Time createdState Active

Refresh

Sort: CreatedColumns

							Metrics						
<input type="checkbox"/>	<input type="checkbox"/>	Run Name	Created	<div>↕</div>	Duration	Source	Models	AUC	F-1 score	accuracy	mae	precision	r2
<input type="checkbox"/>	<input type="checkbox"/>	<div>LightBoost_optimized_2</div>	<div>✓ 19 days ago</div>		28.7s	C:\Users\...	sklearn	0.781	0.887	0.92	0.08	0.567	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Lightboost_optimized_1</div>	<div>✓ 19 days ago</div>		25.5s	C:\Users\...	sklearn	0.779	0.886	0.92	0.08	0.565	
<input type="checkbox"/>	<input type="checkbox"/>	<div>XGBoost_baseline_with_rfe</div>	<div>✓ 20 days ago</div>		9.3s	C:\Users\...	sklearn	0.765	0.888	0.919	0.081	0.456	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Lightboost_baseline_with_rfe</div>	<div>✓ 20 days ago</div>		10.7s	C:\Users\...	sklearn	0.774	0.886	0.92	0.08	0.542	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Catboost_baseline_with_rfe</div>	<div>✓ 20 days ago</div>		20.1s	C:\Users\...	sklearn	0.775	0.886	0.92	0.08	0.562	
<input type="checkbox"/>	<input type="checkbox"/>	<div>XGBoost_baseline_with_smote</div>	<div>✓ 20 days ago</div>		16.2s	C:\Users\...	sklearn	0.754	0.888	0.918	0.082	0.433	
<input type="checkbox"/>	<input type="checkbox"/>	<div>LightBoost_baseline_with_smote</div>	<div>✓ 20 days ago</div>		17.3s	C:\Users\...	sklearn	0.753	0.884	0.919	0.081	0.442	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Catboost_baseline_with_smote</div>	<div>✓ 20 days ago</div>		1.1min	C:\Users\...	sklearn	0.763	0.885	0.92	0.08	0.522	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Ir_c_1.0_max_iter_100_smote</div>	<div>✓ 20 days ago</div>		26.4s	C:\Users\...	sklearn	0.745	-	-	0.3	-	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Ir_c_0.001_max_iter_100_smote</div>	<div>✓ 20 days ago</div>		28.8s	C:\Users\...	sklearn	0.741	-	-	0.306	-	
<input type="checkbox"/>	<input type="checkbox"/>	<div>XGBoost_scale_pos_weight_5</div>	<div>✓ 20 days ago</div>		15.9s	C:\Users\...	sklearn	0.762	0.305	0.869	0.131	0.266	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Lightboost_scale_pos_weight_5</div>	<div>✓ 20 days ago</div>		11.6s	C:\Users\...	sklearn	0.776	0.32	0.865	0.135	0.269	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Catboost_scale_pos_weight_5</div>	<div>✓ 20 days ago</div>		20.4s	C:\Users\...	sklearn	0.778	0.324	0.872	0.128	0.282	
<input type="checkbox"/>	<input type="checkbox"/>	<div>XGBoost_scale_pos_weight_3</div>	<div>✓ 20 days ago</div>		10.9s	C:\Users\...	sklearn	0.764	0.265	0.901	0.099	0.328	
<input type="checkbox"/>	<input type="checkbox"/>	<div>LGBMBoost_scale_pos_weight_3</div>	<div>✓ 20 days ago</div>		11.8s	C:\Users\...	sklearn	0.776	0.272	0.905	0.095	0.354	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Catboost_scale_pos_weight_3</div>	<div>✓ 20 days ago</div>		18.4s	C:\Users\...	sklearn	0.777	0.269	0.906	0.094	0.361	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Lightboost_baseline</div>	<div>✓ 20 days ago</div>		12.2s	C:\Users\...	sklearn	0.767	0.096	0.919	0.081	0.45	
<input type="checkbox"/>	<input type="checkbox"/>	<div>XGBoost_baseline</div>	<div>✓ 20 days ago</div>		17.5s	C:\Users\...	sklearn	0.774	0.059	0.92	0.08	0.531	
<input type="checkbox"/>	<input type="checkbox"/>	<div>Catboost_baseline</div>	<div>✓ 20 days ago</div>		23.0s	C:\Users\...	sklearn	0.776	0.057	0.92	0.08	0.56	

24 matching runs

PARAMÈTRES D'ÉVALUATION DU MODÈLE

1, Precision : $Precision = \frac{vrais\ positifs}{frais\ positifs + faux\ positifs}$

2, Recall : $Recall = \frac{vrais\ positifs}{frais\ positifs + faux\ négatifs}$

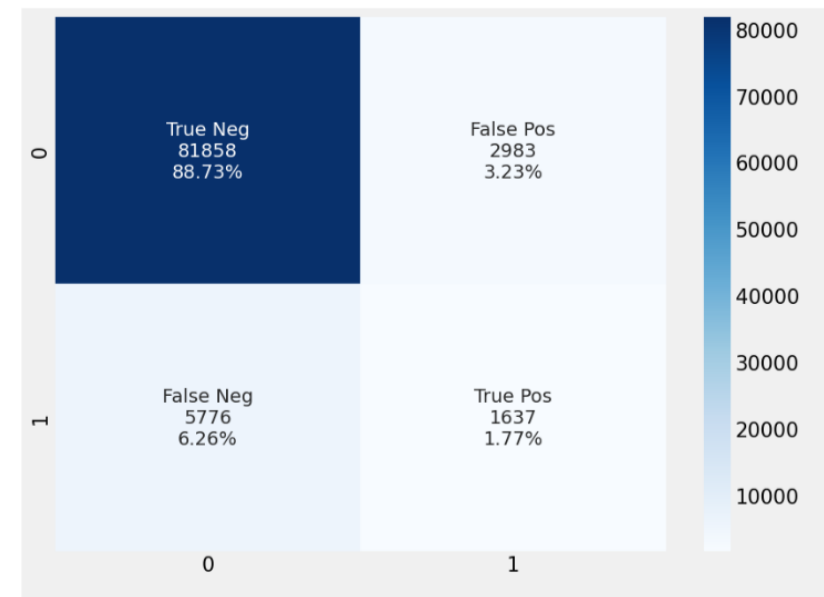
3, F1Score : $F_1 = \frac{2}{recall^{-1} + precision^{-1}} = 2 \frac{precision \cdot recall}{precision + recall} = \frac{2tp}{2tp + fp + fn}$

4, Temps d'entraînement du modèle

5, Matrice de confusion :

Notre modèle idéal serait de retrouver 100% de TP, car ce sont les individus qui ne remboursent pas leur prêt et donc, d'éventuellement détecter les FN aussi.

	Model	AUC	Accuracy	Precision	Recall	F1	Time
0	CatBoostClassifier	0.775818	0.920166	0.56	0.030217	0.05734	16.268135
1	LGBMClassifier	0.773816	0.919938	0.530752	0.031431	0.059348	11.755588
2	XGBClassifier	0.76713	0.918681	0.449944	0.053959	0.096362	9.122981



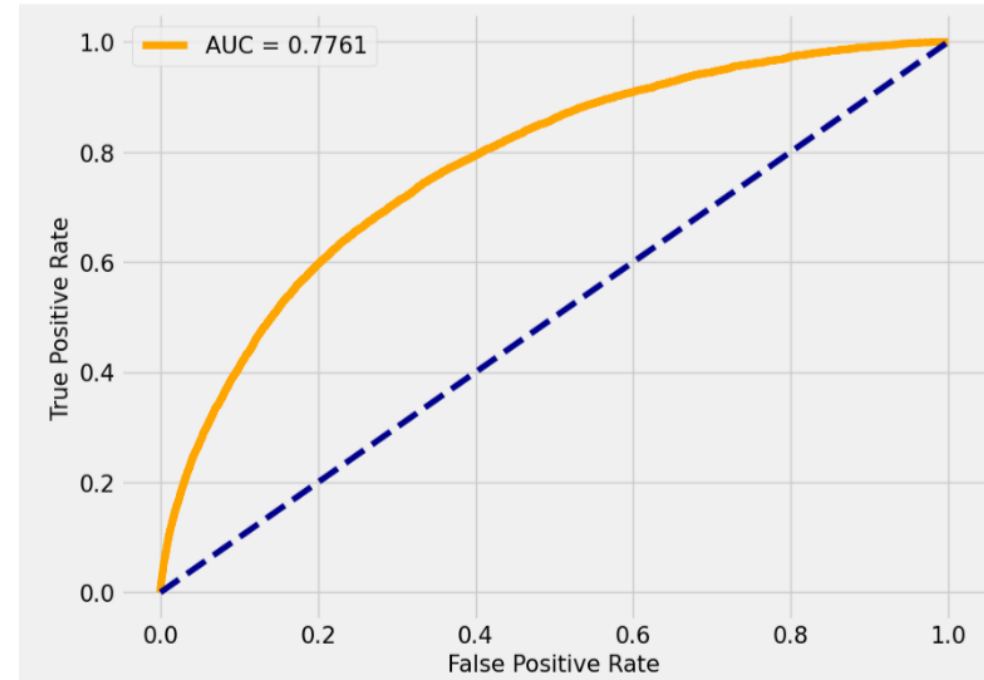
PARAMÈTRES D'ÉVALUATION DU MODÈLE

6, ROC et AUC Score

La courbe ROC (Receiver Operating Characteristic) est un outil utilisé avec les classifieurs binaires. Elle croise le taux de TP avec le taux de FP.

Sur la figure, la ligne en pointillée représente la courbe ROC d'un classifieur purement aléatoire. Un bon classifieur s'en écarte autant que possible (vers le coin supérieur gauche).

Une autre façon de comparer des classifieurs consiste à mesurer l'aire sous la courbe (Area Under the Curve ou AUC). Un classifieur parfait aurait un score AUC égal à 1, tandis qu'un classifieur purement aléatoire aurait un score AUC de 0.5.



	Model	AUC	Accuracy	Precision	Recall	F1	Time
0	CatBoostClassifier	0.775818	0.920166	0.56	0.030217	0.05734	16.268135
1	LGBMClassifier	0.773816	0.919938	0.530752	0.031431	0.059348	11.755588
2	XGBClassifier	0.76713	0.918681	0.449944	0.053959	0.096362	9.122981

PARAMÈTRES D'ÉVALUATION DU MODÈLE

7, Sélection des variables pertinentes - technique d'élimination des caractéristiques récursives avec validation croisée (RFECV).

RFECV conserve les features avec un Rank 1 > True.

8, La fonction Coût

Les erreurs de prédiction doivent être minimisées, dans cette logique une fonction coût ayant pour objectif de pénaliser les Faux Positifs et les Faux Négatifs a été implémentée. Hypothèse à Beta=10 :

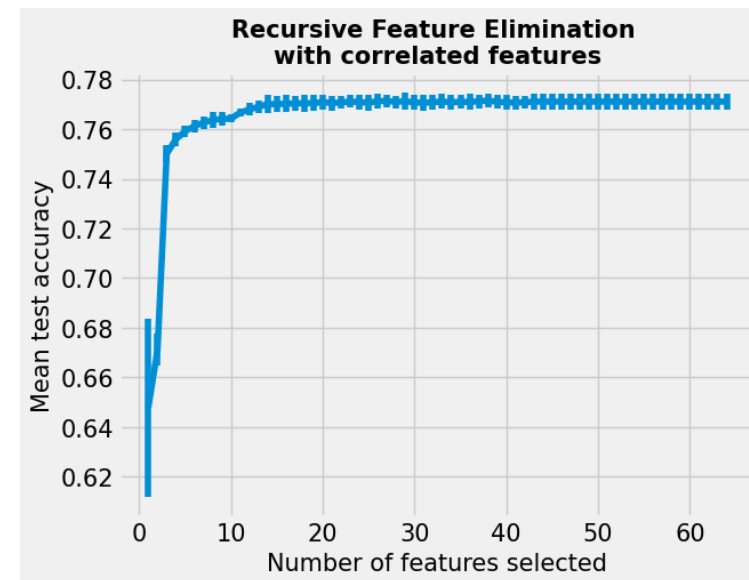
- Défaut de paiement 100% du montant du crédit en pertes et autres recouvrements.
- 10% de chance d'obtenir un crédit pour un client lambda qui souhaite emprunter.

Optimal number of features : 182

Selected Features: [True True True True True True True True True True True True False

False True False True False True True True]

Feature Ranking : [1 1 1 1 1 1 1 1 1 1 7 7 1 7 1 3 1 1 1]



$$Fscore = \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

$$Beta = \frac{coef Recall}{coef Precision}$$

OPTIMISATION DU MODÈLE

Avec Hyperopt, on peut facilement analyser notre modèle de Boosting tout en variant les hyperparamètres. LightGBM couvre plus de 100 hyperparamètres.

```
#Parameter space
space = {
    'n_estimators': hp.quniform('n_estimators', 100, 600, 100),
    'learning_rate': hp.uniform('learning_rate', 0.001, 0.03),
    'max_depth': hp.quniform('max_depth', 3, 7, 1),
    'subsample': hp.uniform('subsample', 0.60, 0.95),
    'colsample_bytree': hp.uniform('colsample_bytree', 0.60, 0.95),
    'reg_lambda': hp.uniform('reg_lambda', 1, 20)
}
```

On a choisi les hyperparamètres suivants :

- `n_estimators` : nombre d'arbres séquentiels.
- `learning_rate` : détermine l'impact de chaque arbre sur le résultat final.
- `max_depth` : profondeur maximale d'un arbre.
- `subsample` : fraction de samples des données train à sélectionner pour chaque arbre.
- `colsample_bytree` : fraction de features à sélectionner pour chaque arbre.

ANALYSE DES RÉSULTATS

LightBoost est celui qui est le plus performant en termes de temps et de AUC score suivi du XGBoost. Les résultats avec équilibrage des données sont plus satisfaisants.

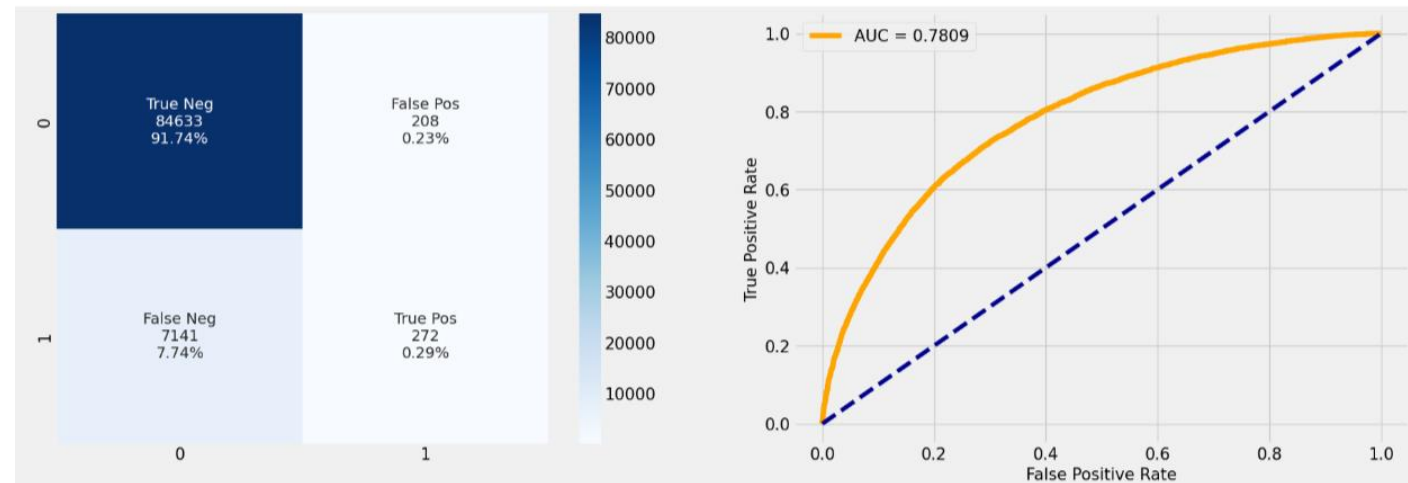
Voici les résultats obtenus avec le LightBoost optimisé :

Accuracy: 0.9203394974743643
Precision: 0.5666666666666667
Recall: 0.03669229731552678
F-1 Score: 0.8869172021298004

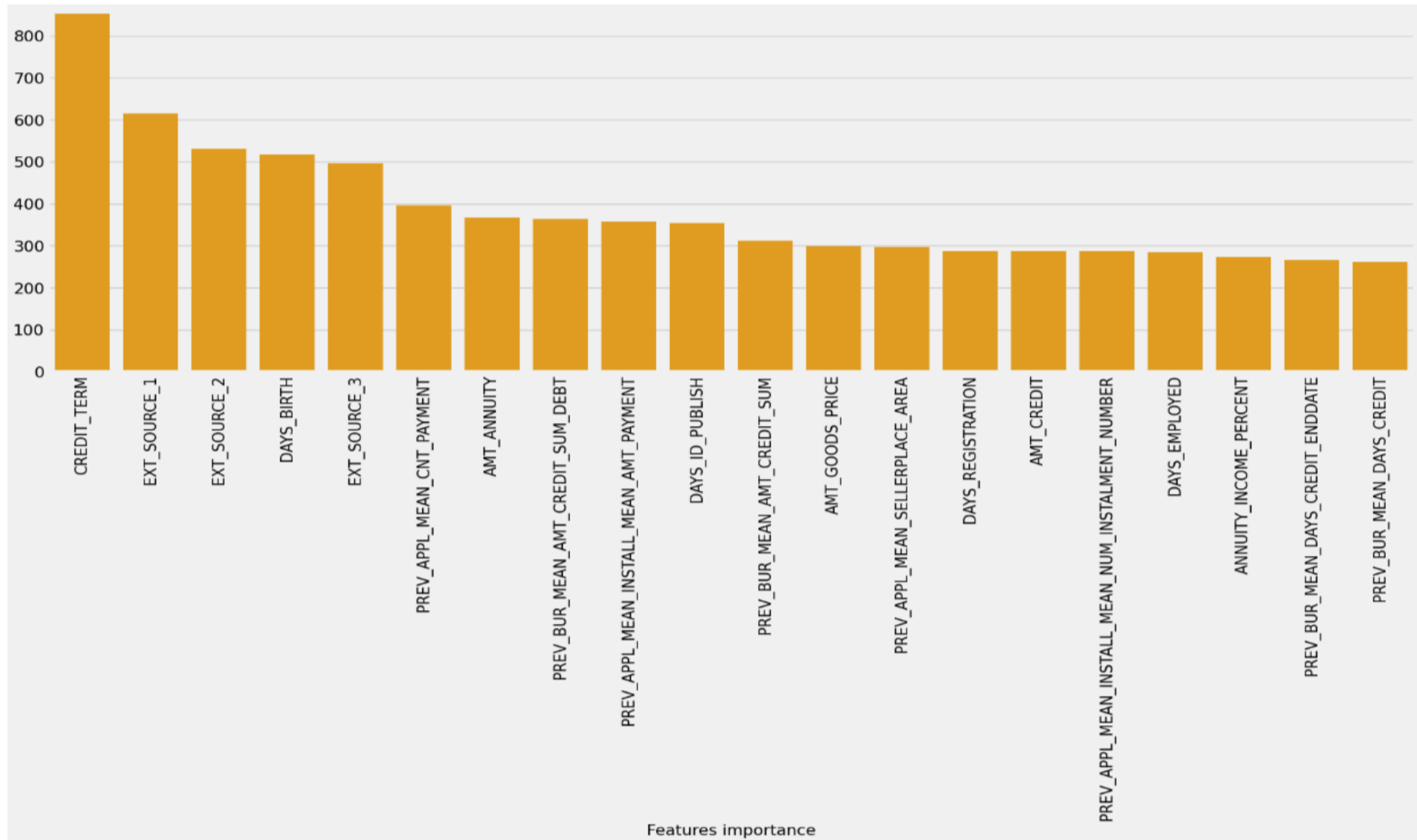
La fonction coût permet de pénaliser les erreurs de prédiction qui peuvent coûter cher à l'entreprise. Au final la métrique métier permet de pénaliser légèrement mieux les erreurs du modèle.

	Model	AUC	Accuracy	Precision	Recall	F1	Time
3	CatBoostClassifier	0.777736	0.872103	0.281574	0.381357	0.323956	16.992299
0	CatBoostClassifier	0.77731	0.906508	0.360626	0.21152	0.266644	16.448041
4	LGBMClassifier	0.77625	0.865112	0.26886	0.394712	0.319851	10.059992
1	LGBMClassifier	0.776072	0.905045	0.354252	0.220828	0.272062	9.614727
2	XGBClassifier	0.763585	0.900872	0.327901	0.222582	0.265167	8.230987
5	XGBClassifier	0.762459	0.869491	0.266478	0.356131	0.30485	9.089757

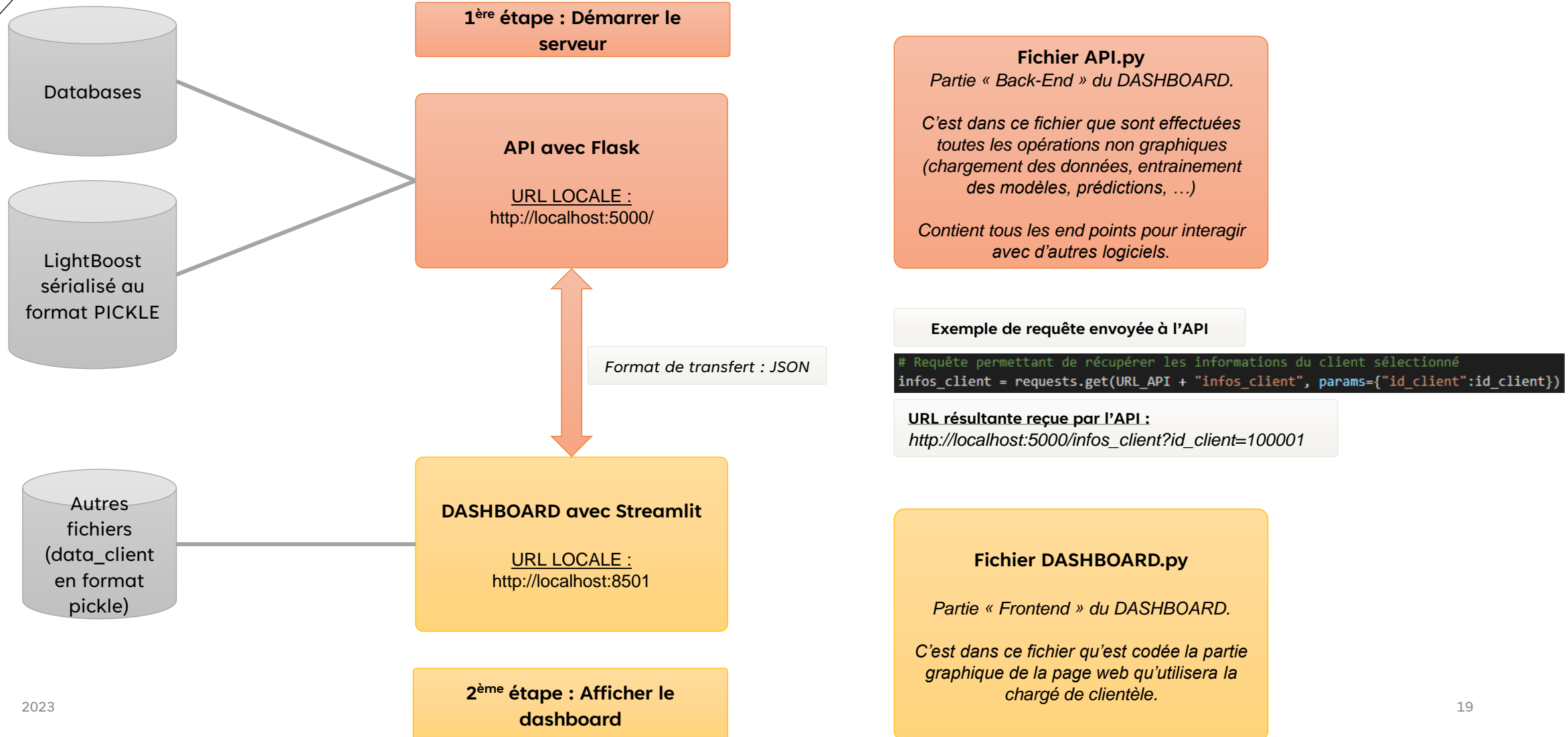
Avec SMOTE



ANALYSE DES RÉSULTATS



PIPELINE DE DÉPLOIEMENT APP_LOCAL



PIPELINE DE DÉPLOIEMENT APP_DEPLOY

The screenshot shows a GitHub repository page for 'keshika-dabidin-audam/projet-7'. The repository is public and has 39 commits ahead of main. The main branch is selected. The repository contains several files and folders, including 'Notebook Files', 'app_deploy', 'app_local', 'tests flask', 'tests_streamlit', 'Note méthodologique.docx', 'README.md', '~Ste méthodologique.docx', and '~WRL2889.tmp'. The 'README.md' file is highlighted. The right sidebar shows the repository's activity, including 0 stars, 1 watching, and 0 forks. The repository is titled 'Openclassrooms projet 7 2023'.

github.com/keshika-dabidin-audam/projet-7/tree/master

Product Solutions Open Source Pricing

Search or jump to... Sign in Sign up

keshika-dabidin-audam / projet-7 Public

Code Issues Pull requests Actions Projects Security Insights

master 3 branches 0 tags Go to file Code

This branch is 39 commits ahead, 1 commit behind main.

keshika-dabidin-audam modification intro doc e4e4cff yesterday 39 commits

Notebook Files	api dashboard and notebooks	last week
app_deploy	Merge branch 'master' of https://github.com/keshika-dabidin-audam/pr...	yesterday
app_local	with sample app_train and app_test	2 days ago
tests flask	api dashboard and notebooks	last week
tests_streamlit	api dashboard and notebooks	last week
Note méthodologique.docx	modification intro doc	yesterday
README.md	Create README.md deployment requirements	last week
~Ste méthodologique.docx	modification intro doc	yesterday
~WRL2889.tmp	modification intro doc	yesterday

README.md

About

Openclassrooms projet 7 2023

- Readme
- Activity
- 0 stars
- 1 watching
- 0 forks

Report repository

Releases

No releases published

Packages

No packages published

PIPELINE DE DÉPLOIEMENT APP_DEPLOY

Commits

master

Commits on Aug 23, 2023

modification intro doc

keshika-dabidin-audam committed yesterday

e4e4cff <>

Merge branch 'master' of <https://github.com/keshika-dabidin-audam/pro...>

keshika-dabidin-audam committed yesterday

fc22296 <>

modification fichiers api

keshika-dabidin-audam committed yesterday

4ad0caf <>

Update Procfile

keshika-dabidin-audam committed yesterday

Verified

c160455 <>

Commits on Aug 22, 2023

with sample app_train and app_test

keshika-dabidin-audam committed 2 days ago

35b182a <>

Commits on Aug 18, 2023

dashboard added

keshika-dabidin-audam committed last week

6b054ee <>

deplacement dans dossier app deploy et api

keshika-dabidin-audam committed last week

c1f6c56 <>

edit in app flask

keshika-dabidin-audam committed last week

7c1c4f1 <>

changes for heroku

keshika-dabidin-audam committed last week

1c6172e <>

only deploy api

keshika-dabidin-audam committed last week

c2f1428 <>

Create runtime.txt

keshika-dabidin-audam committed last week

Verified

414ee1e <>

Update Procfile

Verified

35b182a <>

PIPELINE DE DÉPLOIEMENT APP_DEPLOY

The screenshot shows the Heroku dashboard for the application 'flask-api-oc-7'. The top navigation bar includes the Heroku logo, a search bar, and user profile icons. The main header shows 'Personal' and the app name 'flask-api-oc-7'. Below this, a tab bar contains 'Overview', 'Resources', 'Deploy', 'Metrics', 'Activity' (selected), 'Access', and 'Settings'. A notification banner at the top of the Activity section suggests connecting to GitHub. The Activity Feed lists several events from 'keshika.dabidin@gmail.com':

- Deployed** (d379fc83) - Yesterday at 4:58 PM - v34
- Build succeeded** - Yesterday at 4:58 PM - [View build log](#)
- Deployed** (d1c76737) - Yesterday at 4:56 PM - v33 - [Roll back to here](#)
- Build succeeded** - Yesterday at 4:55 PM - [View build log](#)
- Deployed** (bf06cac1) - Yesterday at 4:52 PM - v32 - [Roll back to here](#)
- Build succeeded** - Yesterday at 4:51 PM - [View build log](#)
- Build failed** - Yesterday at 4:48 PM - [View build log](#)
- Deployed** (4613ee60) - Yesterday at 4:44 PM - v31 - [Roll back to here](#)

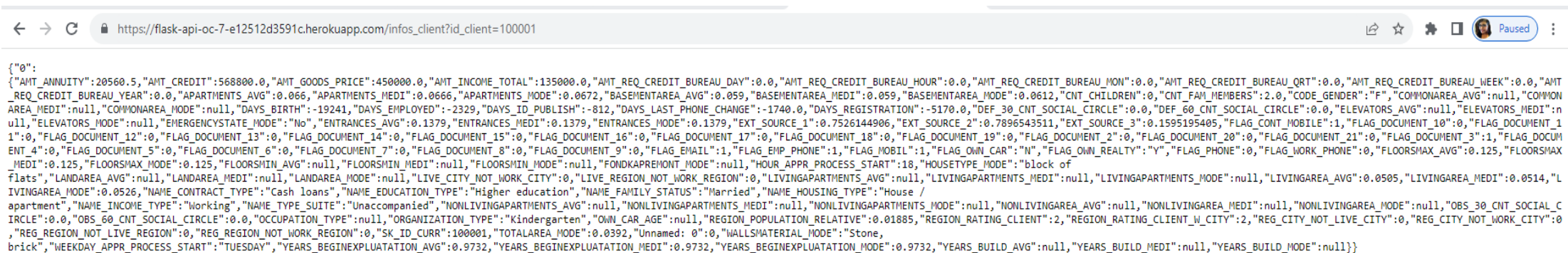
PIPELINE DE DÉPLOIEMENT APP_DEPLOY

Les liens de l'API et du Dashboard sont les Suivants :

1. API : <https://flask-api-oc-7-e12512d3591c.herokuapp.com/>

Pour accéder, par exemple aux informations relatives au client 100001, on écrira :

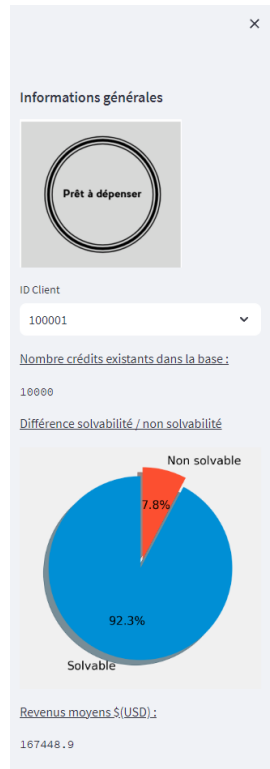
https://flask-api-oc-7-e12512d3591c.herokuapp.com/infos_client?id_client=100001



```
{
  "0": {
    "AMT_ANNUITY": 20560.5, "AMT_CREDIT": 568800.0, "AMT_GOODS_PRICE": 450000.0, "AMT_INCOME_TOTAL": 135000.0, "AMT_REQ_CREDIT_BUREAU_DAY": 0.0, "AMT_REQ_CREDIT_BUREAU_HOUR": 0.0, "AMT_REQ_CREDIT_BUREAU_MON": 0.0, "AMT_REQ_CREDIT_BUREAU_QRT": 0.0, "AMT_REQ_CREDIT_BUREAU_WEEK": 0.0, "AMT_REQ_CREDIT_BUREAU_YEAR": 0.0, "APARTMENTS_AVG": 0.066, "APARTMENTS_MEDI": 0.0666, "APARTMENTS_MODE": 0.0672, "BASEMENTAREA_AVG": 0.059, "BASEMENTAREA_MEDI": 0.059, "BASEMENTAREA_MODE": 0.0612, "CNT_CHILDREN": 0, "CNT_FAM_MEMBERS": 2.0, "CODE_GENDER": "F", "COMMONAREA_AVG": null, "COMMONAREA_MEDI": null, "COMMONAREA_MODE": null, "DAYS_BIRTH": -19241, "DAYS_EMPLOYED": -2329, "DAYS_ID_PUBLISH": -812, "DAYS_LAST_PHONE_CHANGE": -1740.0, "DAYS_REGISTRATION": -5170.0, "DEF_30_CNT_SOCIAL_CIRCLE": 0.0, "DEF_60_CNT_SOCIAL_CIRCLE": 0.0, "ELEVATORS_AVG": null, "ELEVATORS_MEDI": null, "ELEVATORS_MODE": null, "EMERGENCYSTATE_MODE": "No", "ENTRANCES_AVG": 0.1379, "ENTRANCES_MEDI": 0.1379, "ENTRANCES_MODE": 0.1379, "EXT_SOURCE_1": 0.7526144906, "EXT_SOURCE_2": 0.7896543511, "EXT_SOURCE_3": 0.1595195405, "FLAG_DOCUMENT_1": 0, "FLAG_DOCUMENT_10": 0, "FLAG_DOCUMENT_11": 0, "FLAG_DOCUMENT_12": 0, "FLAG_DOCUMENT_13": 0, "FLAG_DOCUMENT_14": 0, "FLAG_DOCUMENT_15": 0, "FLAG_DOCUMENT_16": 0, "FLAG_DOCUMENT_17": 0, "FLAG_DOCUMENT_18": 0, "FLAG_DOCUMENT_19": 0, "FLAG_DOCUMENT_2": 0, "FLAG_DOCUMENT_20": 0, "FLAG_DOCUMENT_21": 0, "FLAG_DOCUMENT_3": 1, "FLAG_DOCUMENT_4": 0, "FLAG_DOCUMENT_5": 0, "FLAG_DOCUMENT_6": 0, "FLAG_DOCUMENT_7": 0, "FLAG_DOCUMENT_8": 0, "FLAG_DOCUMENT_9": 0, "FLAG_EMAIL": 1, "FLAG_EMP_PHONE": 1, "FLAG_MOBILE": 1, "FLAG_OWN_CAR": "N", "FLAG_OWN_REALTY": "Y", "FLAG_PHONE": 0, "FLAG_WORK_PHONE": 0, "FLOORSMAX_AVG": 0.125, "FLOORSMAX_MEDI": 0.125, "FLOORSMAX_MODE": 0.125, "FLOORSMIN_AVG": null, "FLOORSMIN_MEDI": null, "FLOORSMIN_MODE": null, "FONDKAPREMONT_MODE": null, "HOUR_APPR_PROCESS_START": 18, "HOUSETYPE_MODE": "block of flats", "LANDAREA_AVG": null, "LANDAREA_MEDI": null, "LANDAREA_MODE": null, "LIVE_CITY_NOT_WORK_CITY": 0, "LIVE_REGION_NOT_WORK_REGION": 0, "LIVINGAPARTMENTS_AVG": null, "LIVINGAPARTMENTS_MEDI": null, "LIVINGAPARTMENTS_MODE": null, "LIVINGAREA_AVG": 0.0505, "LIVINGAREA_MEDI": 0.0514, "LIVINGAREA_MODE": 0.0526, "NAME_CONTRACT_TYPE": "Cash loans", "NAME_EDUCATION_TYPE": "Higher education", "NAME_FAMILY_STATUS": "Married", "NAME_HOUSING_TYPE": "House / apartment", "NAME_INCOME_TYPE": "Working", "NAME_TYPE_SUITE": "Unaccompanied", "NONLIVINGAPARTMENTS_AVG": null, "NONLIVINGAPARTMENTS_MEDI": null, "NONLIVINGAPARTMENTS_MODE": null, "NONLIVINGAREA_AVG": null, "NONLIVINGAREA_MEDI": null, "NONLIVINGAREA_MODE": null, "OBS_30_CNT_SOCIAL_CIRCLE": 0.0, "OBS_60_CNT_SOCIAL_CIRCLE": 0.0, "OCCUPATION_TYPE": null, "ORGANIZATION_TYPE": "Kindergarten", "OWN_CAR_AGE": null, "REGION_POPULATION_RELATIVE": 0.01885, "REGION_RATING_CLIENT": 2, "REGION_RATING_CLIENT_W_CITY": 2, "REG_CITY_NOT_LIVE_CITY": 0, "REG_CITY_NOT_WORK_CITY": 0, "REG_REGION_NOT_LIVE_REGION": 0, "REG_REGION_NOT_WORK_REGION": 0, "SK_ID_CURR": 100001, "TOTALAREA_MODE": 0.0392, "Unnamed: 0": 0, "WALLSMATERIAL_MODE": "Stone, brick", "WEEKDAY_APPR_PROCESS_START": "TUESDAY", "YEARS_BEGINEXPLUATATION_AVG": 0.9732, "YEARS_BEGINEXPLUATATION_MEDI": 0.9732, "YEARS_BEGINEXPLUATATION_MODE": 0.9732, "YEARS_BUILD_AVG": null, "YEARS_BUILD_MEDI": null, "YEARS_BUILD_MODE": null}}
}
```

PIPELINE DE DÉPLOIEMENT APP_DEPLOY

2. Dashboard : <https://streamlit-oc-7-5ac02169264e.herokuapp.com/>



Implémenter un modèle de scoring

API répondant aux besoins du projet 7 pour le parcours Data Scientist OpenClassRoom

Vous avez sélectionné le client : 100001

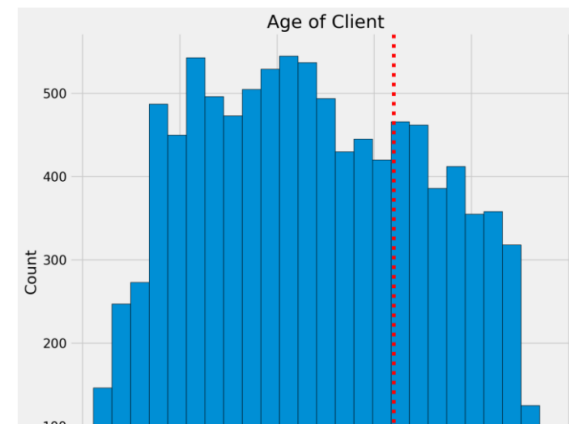
Informations client

☒ Afficher les informations du client?

Statut famille : ** Married **

Nombre d'enfant(s) : ** 0 **

Age client : 52 ans.



DATA DRIFT



La distribution des données d'entrée change au fil du temps. Le rapport Data Drift permet de détecter et d'explorer les changements dans les données d'entrée.

Voici les différents rapports obtenus à partir de l'analyse du Data Drift :

Dataset Drift		
Dataset Drift is NOT detected. Dataset drift detection threshold is 0.5		
312 Columns	3 Drifted Columns	0.00962 Share of Drifted Columns

Par défaut, la dérive de l'ensemble de données est détectée si au moins 50 % des entités dérivent. Il y a ici 3 colonnes qui dérivent mais le data drift n'est pas détecté.

Drift is detected for 0.962% of columns (3 out of 312).

Search ×						
Column	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score ↑
> PREV_APPL_MEAN_DAYS_DECISION	num			Detected	K-S p_value	0.044248
> PREV_APPL_MEAN_DAYS_LAST_DUE_1ST_VERSION	num			Detected	K-S p_value	0.079211
> PREV_APPL_MEAN_DAYS_TERMINATION	num			Detected	K-S p_value	0.113665
> PREV_APPL_MEAN_SELLERPLACE_AREA	num			Not Detected	K-S p_value	0.23649
> PREV_APPL_MEAN_DAYS_LAST_DUE	num			Not Detected	K-S p_value	0.254765
> PREV_APPL_MEAN_INSTALL_MEAN_DAYS_ENTRY_PAYMENT	num			Not Detected	K-S p_value	0.274059

DATA DRIFT

Pour les features numériques, on peut également explorer les valeurs dans un tracé. La ligne vert foncé représente la moyenne, comme on le voit dans l'ensemble de données de référence. La zone verte couvre un écart type par rapport à la moyenne.



CONCLUSION

- Après une baseline faite avec un algorithme simple de régression logistique, l'AUC score avait été estimé ≈ 0.72 avec rééquilibrage (SMOTE) des données. La suite de l'étude a été déroulée vers 3 algorithmes plus complexes de gradient boosting implémentés par LightGbm vs CatBoost vs XGBoost.
- Nous avons pu démontrer les performances de ces algorithmes par une sélection de features, à l'origine > 300 , après RFECV 182.
- LightGbm ressort comme étant le plus rapide, le plus performant sur la métrique classique de l'AUC, il a donc été optimisé et déployé en local et sur le cloud.

Limites et Améliorations :

- La modélisation effectuée dans le cadre du projet a été effectuée sur la base d'une hypothèse forte. L'axe principal d'amélioration serait de définir plus finement la métrique d'évaluation en collaboration avec les équipes métier.
- Ajouter d'autres hyperparamètres peut également permettre d'augmenter les performances actuelles.
- L'opportunité d'améliorer la modélisation en utilisant d'autres features issues de données complémentaires fournies, ainsi qu'en créant de nouvelles features en collaboration avec les équipes métier.



A series of white, thin, overlapping geometric lines on a black background, forming various polygons and intersecting points, located on the left side of the slide.

MERCI POUR VOTRE ATTENTION.