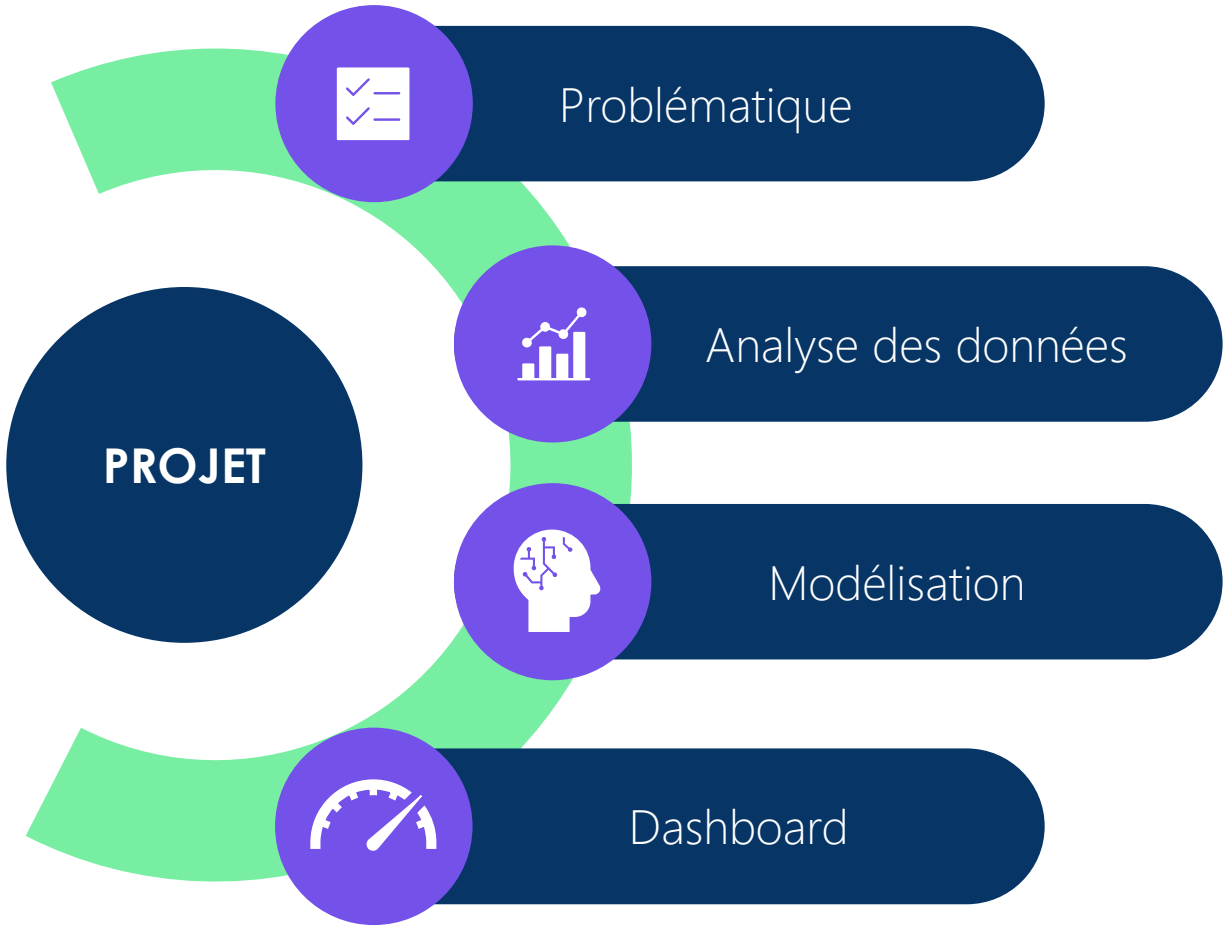
Projet N°7 :
Implémentez un modèle de scoring

Agustin Bunader (autofinancé)
Soutenance de Projet
Août 2021

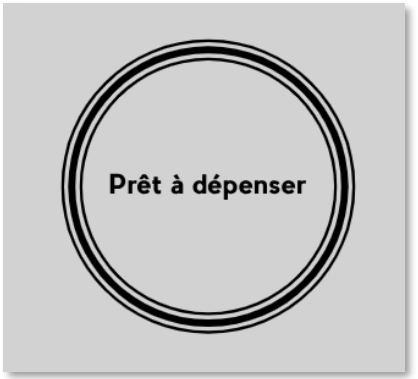
Programme


PROJET
Problématique
Analyse des données
Modélisation
Dashboard

Problématique – Présentation

Contexte :

- L'entreprise spécialisé en crédits à la consommation souhaite développer un modèle de scoring de la probabilité de défaut de paiement des clients



Objectifs :

- Développer un modèle de scoring de la probabilité de défaut de paiement du client
- Développer un dashboard interactif pour assurer une transparence sur les décisions d'octroi de crédit

Mission :

- Construire un modèle de scoring qui donnera une prédiction sur la probabilité de faillite d'un client de façon automatique
- Construire un dashboard interactif que permettra d'interpréter les prédictions faites par le modèle et d'améliorer la connaissance client des chargés de relation client

Sources :

- Base de données contenant des informations personnelles et financières des clients avec plus de 300000 clients et 120 features dans le set d'entraînement

Contraintes :

- Fonctionnement en temps réel
- Utilisation des services gratuites avec une puissance de calcul très faible et une capacitée de stockage inferieure à 1 GO

Problématique – Etapes

Interprétation :

- Exploration des données et choix des features adaptés
- Classification non-supervisée des clients avec un modèle de machine learning expliqué et réutilisable par l'équipe de relation client

Exploration des données → Préparation et classification des divers jeux de donnés → Mise en place du baseline → Assemblage du meilleur modèle et jeu de données → Dashboard

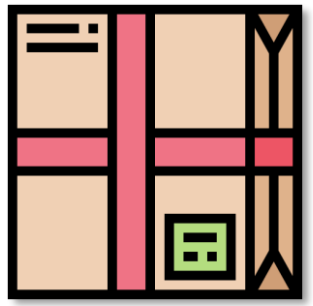




Problématique – Livrables

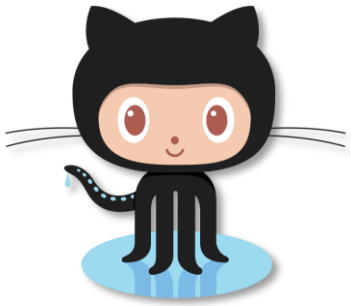
Files :

- P7_01_analyse.ipynb : EDA + Feature engineering + Feature selection + Scoring + Model evaluation
- P7_02_LIME_SHAP.ipynb : Interprétation du model
- P7_03_extras.ipynb : Dataframes extras utilises dans le dashboard
- P7_04_dashboard.py : Python contenant la version locale du dashboard crée avec Streamlit
- P7_05_note_methodologique.pdf : Fichier décrivant les détails techniques du projet
- P7_06_presentation.pdf : Présentation du projet

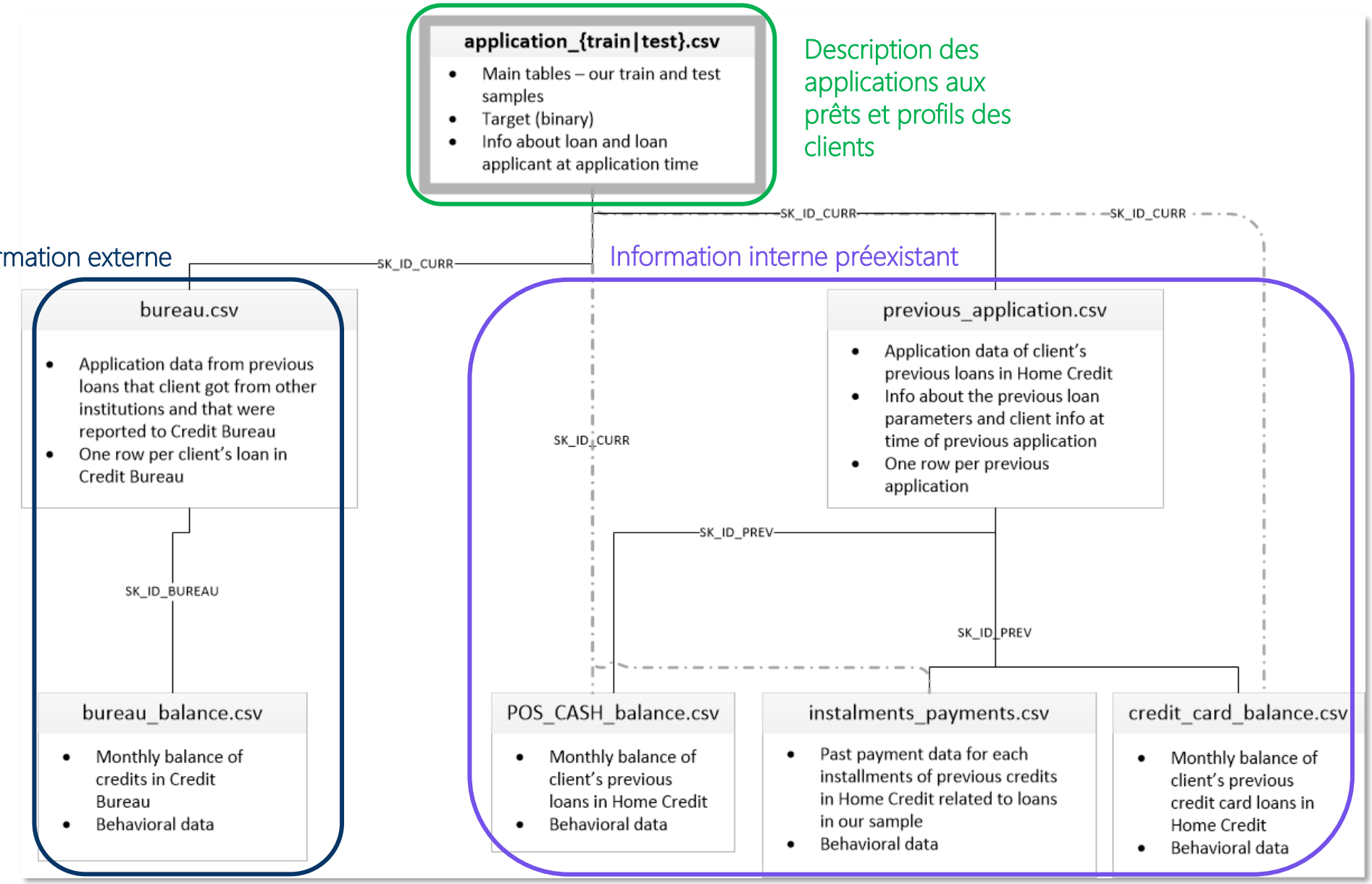


Cloud :

- Le repository Github du projet est disponible en cliquant ici
- Le dashboard est aussi disponible en cliquant ici



Analyse des donnés – Source



Analyse des donnés – Feature engineering

Processus :

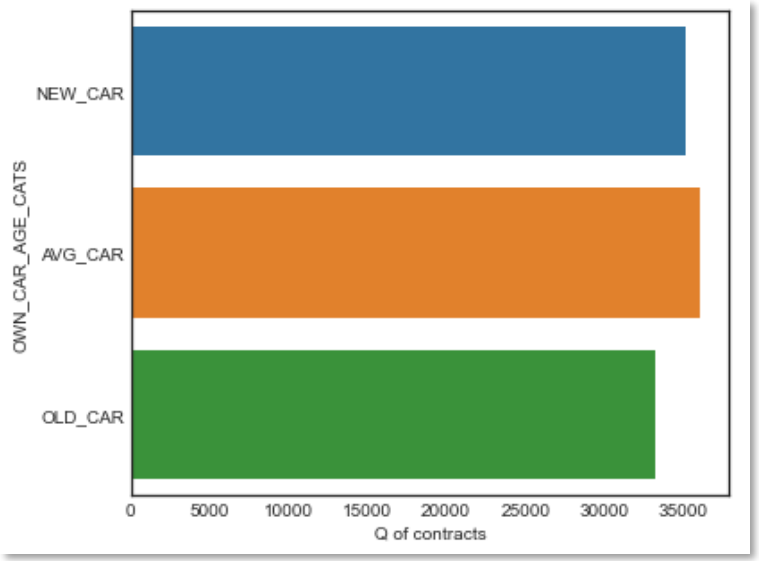
- Création des variables dummy pour les features du type catégoriel
- Détection des anomalies dans les dates et des catégories non-interprétables
- Création des nouveau features (ratios, catégories basées sur des données discrètes, durée du crédit et agréger les données)
- Suppression des features avec plus de 20% des valeurs manquants et imputation des valeurs manquants sur les colonnes restantes
- Agréger les différents dataframes sur la colonne *SK_ID_CURR*

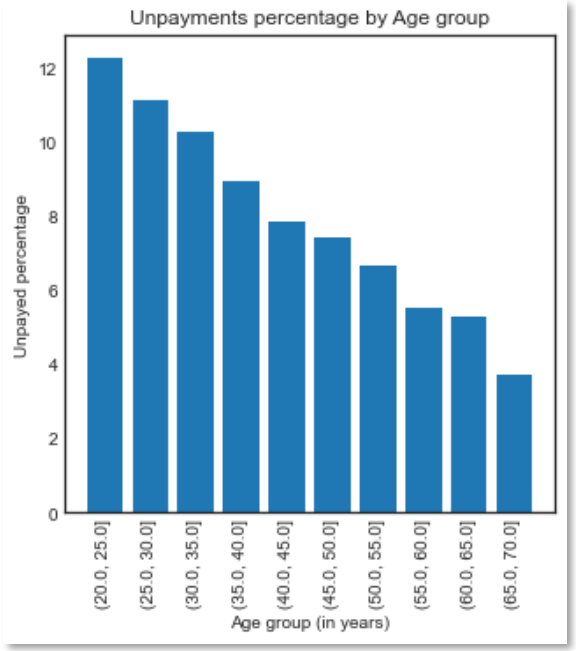
```
Clients with previous loans in general: 85.84%
Clients with previous loans from the company: 95.12%

Clients with both types of loans: 81.71%
Clients with no previous loans at all: 0.75%

Clients with any kind of loans, Test dataset: 208
Clients with any kind of loans,Train dataset: 2470

Unknown applicants, Test dataset: 0.43%
Unknown applicants, Train dataset: 0.8%
```



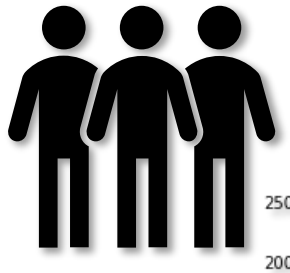


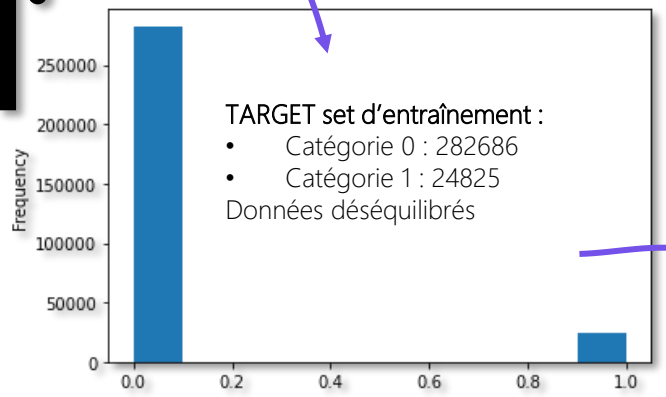
	DAYS_BIRTH	DAYS_EMPLOYED
count	307511.000000	307511.000000
mean	43.936973	-174.835742
std	11.956133	387.056895
min	20.517808	-1000.665753
25%	34.008219	0.791781
50%	43.150685	3.323288
75%	53.923288	7.561644
max	69.120548	49.073973

Analyse des donnés – Clients

Applications :

- 307511 set d'entraînement
- 48744 set de test



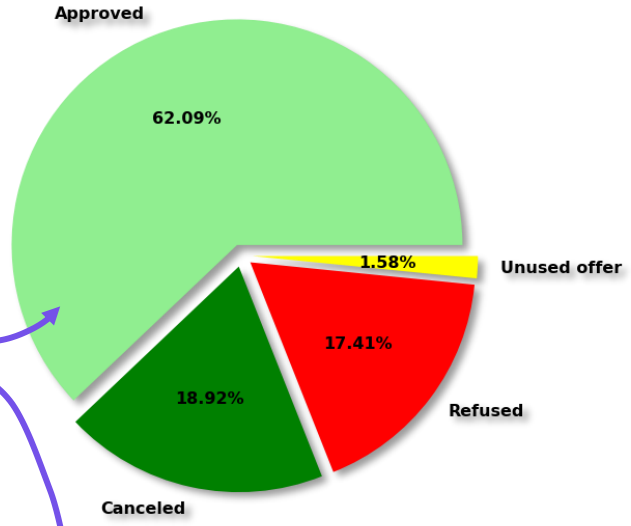


TARGET set d'entraînement :

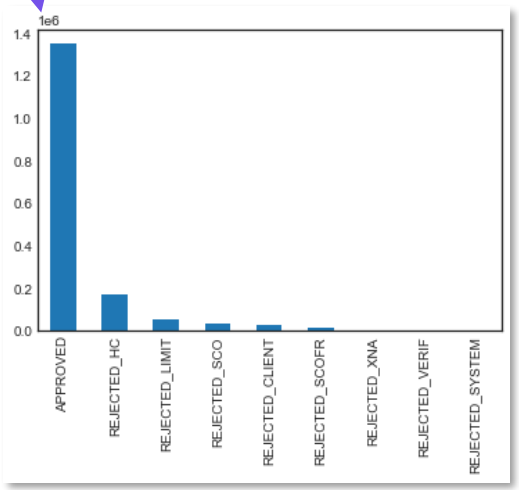
- Catégorie 0 : 282686
- Catégorie 1 : 24825

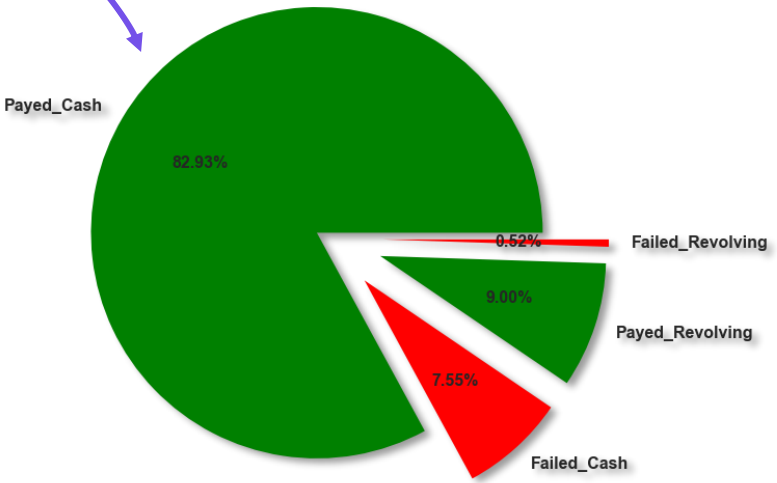
Données déséquilibrés

95.1% des clients ont des demandes de prêts antérieures avec Prêt à dépenser



85.8% des clients ont des demandes de prêts antérieures avec des autres établissements





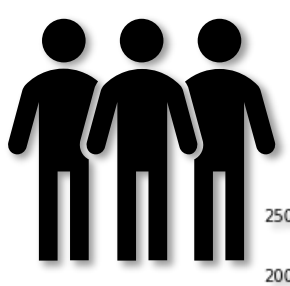
Target 0 : client qui n'a eu aucune difficulté à rembourser son prêt

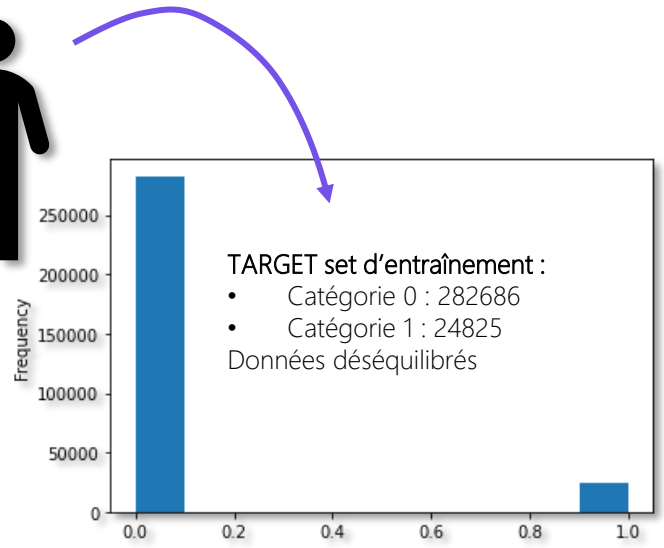
Target 1 : client qui n'a pas remboursé son prêt

Analyse des donnés – Déséquilibre

Applications :

- 307511 set d'entraînement
- 48744 set de test





TARGET set d'entraînement :

- Catégorie 0 : 282686
- Catégorie 1 : 24825

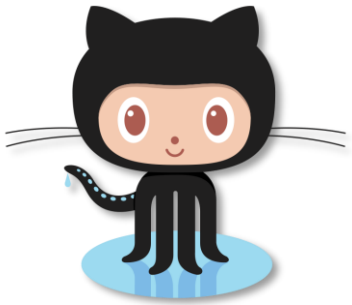
Données déséquilibrés

Comme est-ce qu'on peut réduire les conséquences de se déséquilibre ?

- SMOTE (Synthetic Minority Oversampling Technique)
- RandomUnderSampler (undersampling)

SMOTE consiste en la création des individus « synthétiques » sur la base de ceux déjà existants en choisissant au hasard un point dans la classe minoritaire et en calculant les k plus proches voisins pour ce point. Les points synthétiques sont ajoutés entre le point choisi et ses voisins.

RandomUnderSampler supprime des individus appartenant à la classe la plus lourde en choisissant des points aléatoires sans ou avec replacement (sans dans notre cas)



Libraire **imblearn**, méthode **under_sampling**

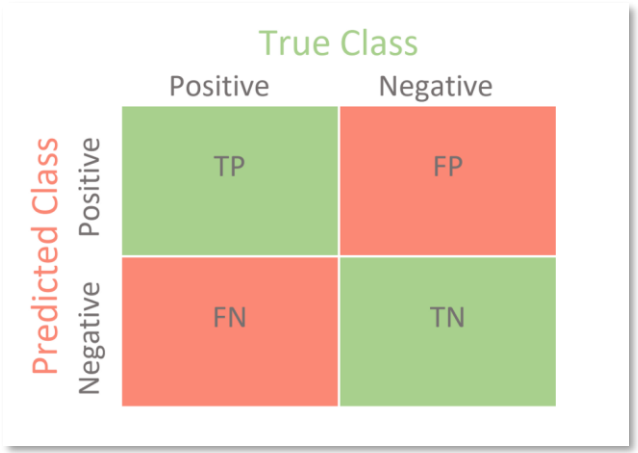
Modélisation – Problématique

Problématique :

- L'analyse du risque de crédit est une forme d'analyse effectuée par un analyste de crédit pour déterminer la capacité d'un emprunteur à honorer ses dettes
- Eviter cataloguer comme applications à risque des potentiels clients qui ne présentent pas de risquer

Une façon de résoudre cette situation c'est utilisant une matrice de confusion pour mesurer la qualité du système de classification.

- **Vrai négatif (TN) et Vrai positif (TP)** : prédictions correctes
- **Faux négatif (FN)** : TARGET 1 prédit comme TARGET 0. **Haut exposition au risque**
- **Faux positif (FP)** : TARGET 0 prédit comme TARGET 1. **Potentielle client perdu**

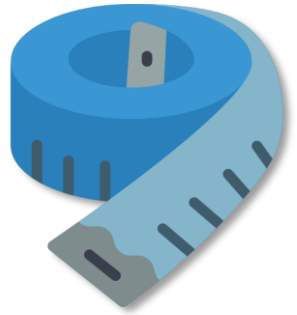


On cherche à **diminuer** les prédictions fausses (FP+FN) avec un intérêt principale sur les **Faux négatifs**.

Modélisation – Scoring

Problème :

- Limiter les Faux négatifs
- Limiter les Faux positifs



Mesures

- Accuracy : Proportion de prédictions que le modèle a classées correctement.
- Precision: Quelle proportion d'identifications positives était réellement correcte ?
- Recall: Quelle proportion de positifs réels a été identifiée correctement ?
- F1-Score : Mesure de la précision d'un test, c'est la moyenne harmonique de Precision et Recall fournissant un score unique qui équilibre à la fois les préoccupations de Precision et de Recall en un seul nombre avec un score maximum de 1 (précision et rappel parfaits) et 0. Globalement, c'est une mesure de la précision et de la robustesse du modèle.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + TP + FP + FN}$$

Correspondances :

- TP : Vrai positif
- TN : Vrai négatif
- FP : Faux positif
- FN : Faux négatif

Modélisation – Méthodologie

Baseline :

- Logistic Regression avec imputation des valeurs manquants et normalisation des features (feature scaling)

Rééquilibrage :

- RandomUnderSampler

Algorithmes :

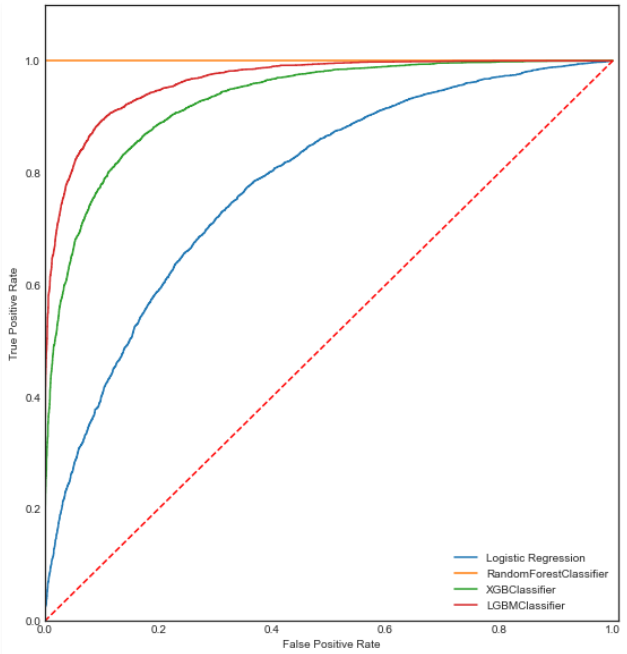
- Logistic Regression
- Random Forest Classifier
- XGBClassifier
- LGBMClassifier

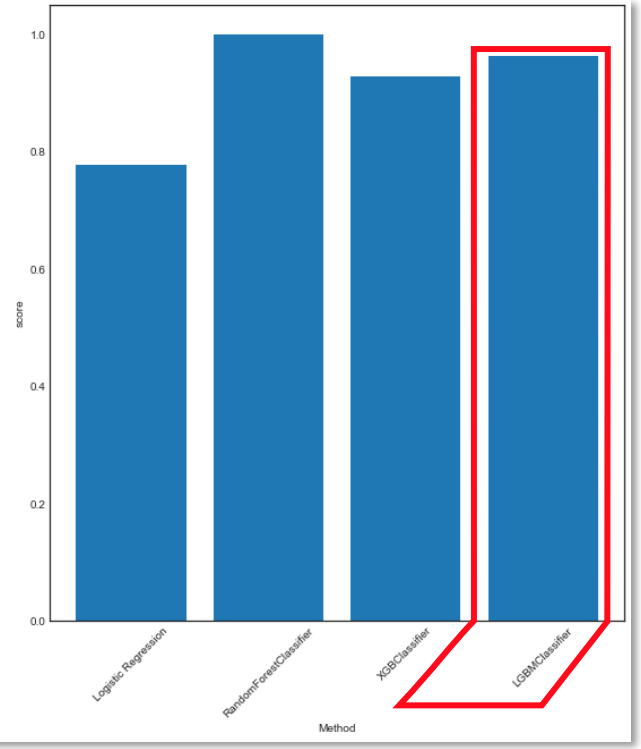
Modélisation – Comparaison des modèles

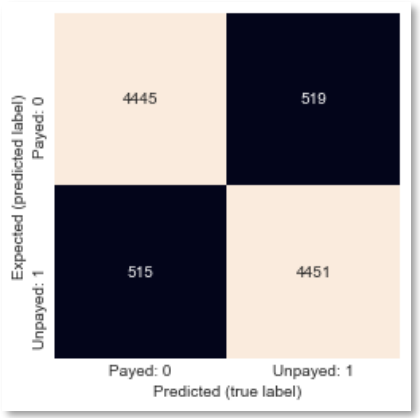
{0:	Accuracy	Precision	Recall	F1	Time
Logistic Regression	0.68	0.69	0.68	0.68	00:00:14
RandomForestClassifier	1.0	1.0	1.0	1.0	00:00:24
XGBClassifier	0.82	0.82	0.81	0.82	00:00:10
LGBMClassifier	0.87	0.87	0.87	0.87	00:00:13,
1:	Accuracy	Precision	Recall	F1	Time
Logistic Regression	0.71	0.71	0.71	0.71	00:00:19
RandomForestClassifier	1.0	1.0	1.0	1.0	00:00:40
XGBClassifier	0.85	0.85	0.84	0.85	00:00:27
LGBMClassifier	0.9	0.9	0.9	0.9	00:00:30,
2:	Accuracy	Precision	Recall	F1	Time
Logistic Regression	0.7	0.7	0.69	0.69	00:00:31
RandomForestClassifier	1.0	1.0	1.0	1.0	00:00:32
XGBClassifier	0.83	0.83	0.83	0.83	00:00:20
LGBMClassifier	0.88	0.88	0.88	0.88	00:00:21}

```
Scores for second dataset (complete traeatment)
Logistic Regression AUC Score: 0.7773565726391518
RandomForestClassifier AUC Score: 1.0
XGBClassifier AUC Score: 0.9280718880328213
LGBMClassifier AUC Score: 0.9631260094833426
```

```
LGBMClassifier
TARGET 1 predicted: 4970
Confusion Matrix:
 [[4445  519]
 [ 515 4451]]
```







RandomForestClassifier a des problèmes de surapprentissage

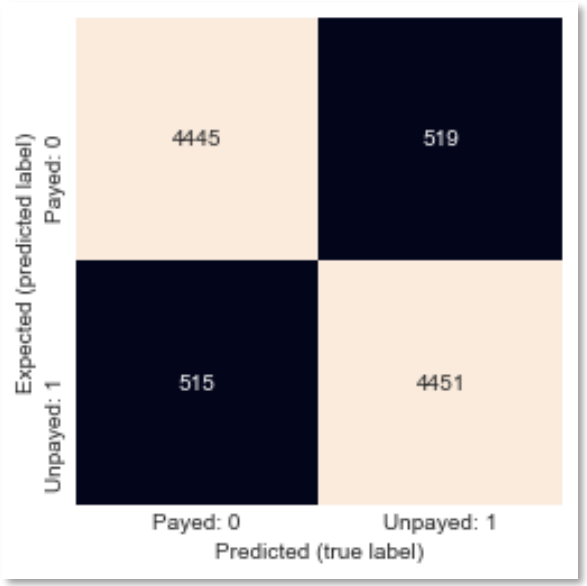
Modélisation – Sélection des features

Taille originale :

- 394 features

1:	Accuracy	Precision	Recall	F1	Time
Logistic Regression	0.71	0.71	0.71	0.71	00:00:19
RandomForestClassifier	1.0	1.0	1.0	1.0	00:00:40
XGBClassifier	0.85	0.85	0.84	0.85	00:00:27
LGBMClassifier	0.9	0.9	0.9	0.9	00:00:30,

```
Logistic Regression AUC Score: 0.7773565726391518
RandomForestClassifier AUC Score: 1.0
XGBClassifier AUC Score: 0.9280718880328213
LGBMClassifier AUC Score: 0.9631260094833426
```

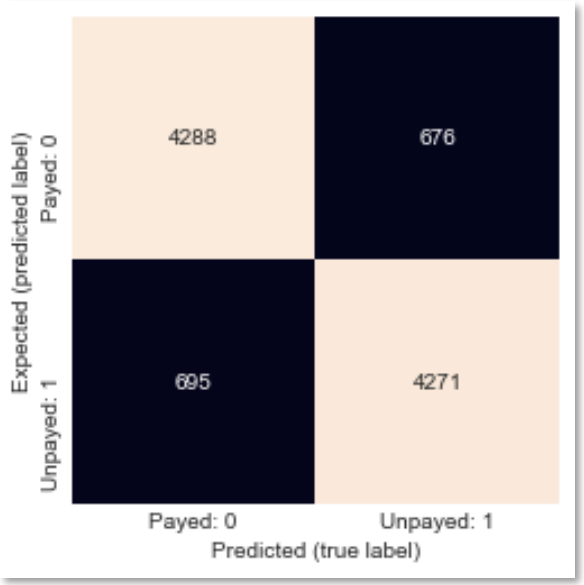


Taille après réduction :

- 40 features

{0:	Accuracy	Precision	Recall	F1	Time
Logistic Regression	0.68	0.68	0.67	0.68	00:00:05
RandomForestClassifier	1.0	1.0	1.0	1.0	00:00:28
XGBClassifier	0.82	0.82	0.82	0.82	00:00:05
LGBMClassifier	0.86	0.86	0.86	0.86	00:00:06}

```
Logistic Regression AUC Score: 0.7459157808959102
RandomForestClassifier AUC Score: 1.0
XGBClassifier AUC Score: 0.9017131765952068
LGBMClassifier AUC Score: 0.9423653365041833
```



Dashboard – Technologies utilisées

Gestion de versions	Github	
Dashboard frontend	Streamlit	
Décryptage de la prédiction	LIME	

Dashboard – Présentation

Filtrage par client

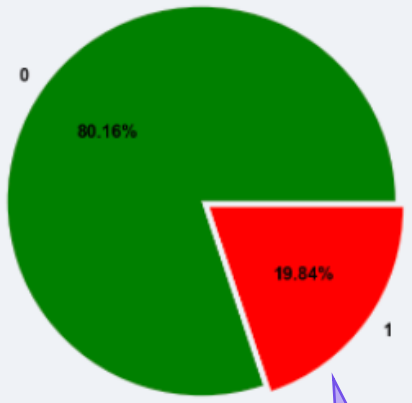
Information basique du prêt demandé

Information personnelle basique du client

Client Selection

SK_ID_CURR

100001



Target distribution

Credit Scoring

Loan info

Default risk: 0.4326

Category (0/1): 0

Contract: Cash loans

Amount: 568800.0

Annuity: 20560.5

Credit over income: 4.215

Previous known loans: 8.0

Rejected loans ratio: 0.0

Client info

Gender: F

Age: 53

Status: Married

Education: Higher education

Occupation: nan

Employer: Kindergarten , for 6 years

Housing: House / apartment

Car: N

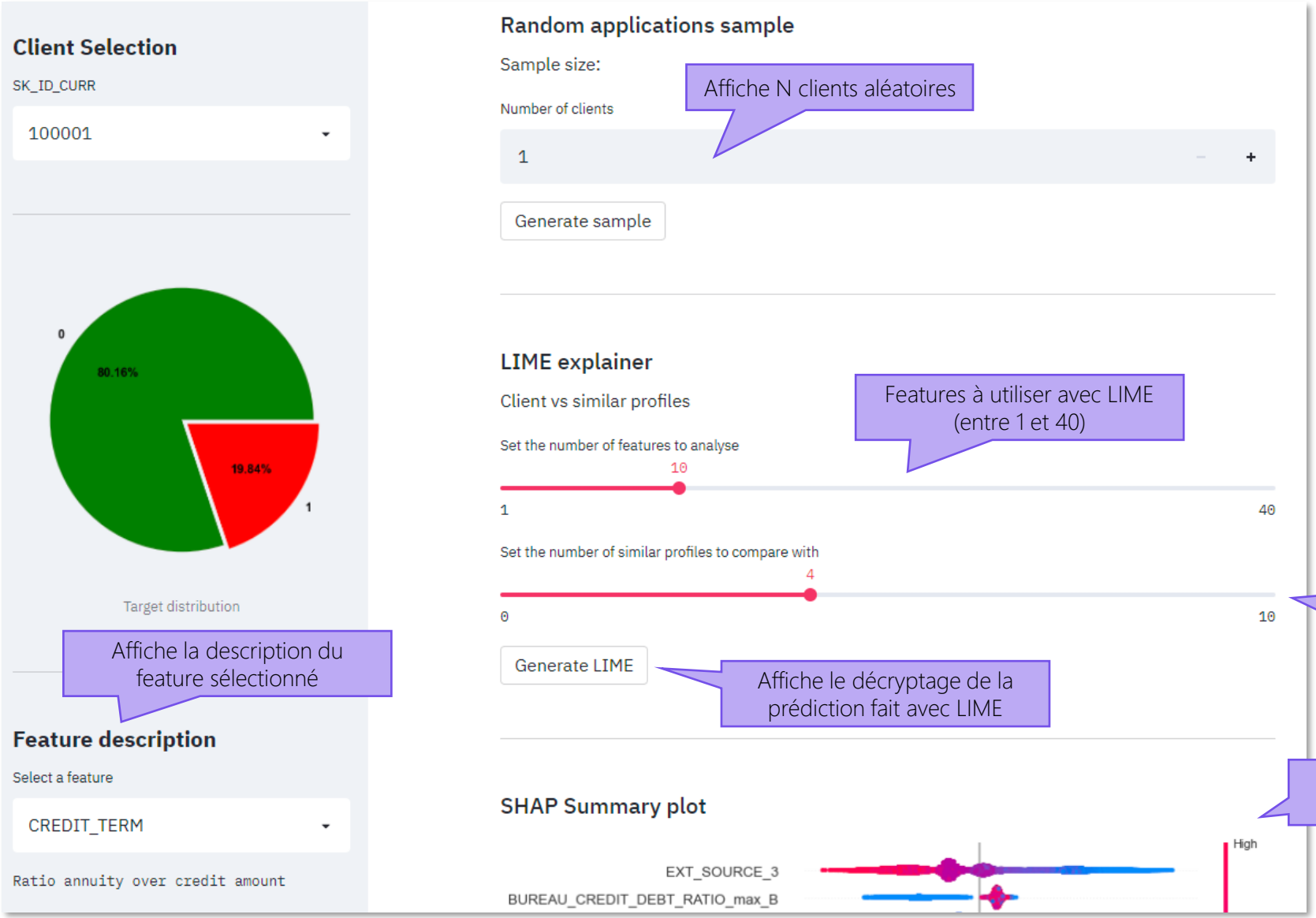
Application data:

	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	CNT_CHILDREN	AMT_INCOME
100001	0	1	0	0	

Prédiction de TARGET

Détails du client

Dashboard – Présentation



Affiche N clients aléatoires

Features à utiliser avec LIME (entre 1 et 40)

Clients similaires à trouver pour la comparaison avec LIME (entre 0 et 10)

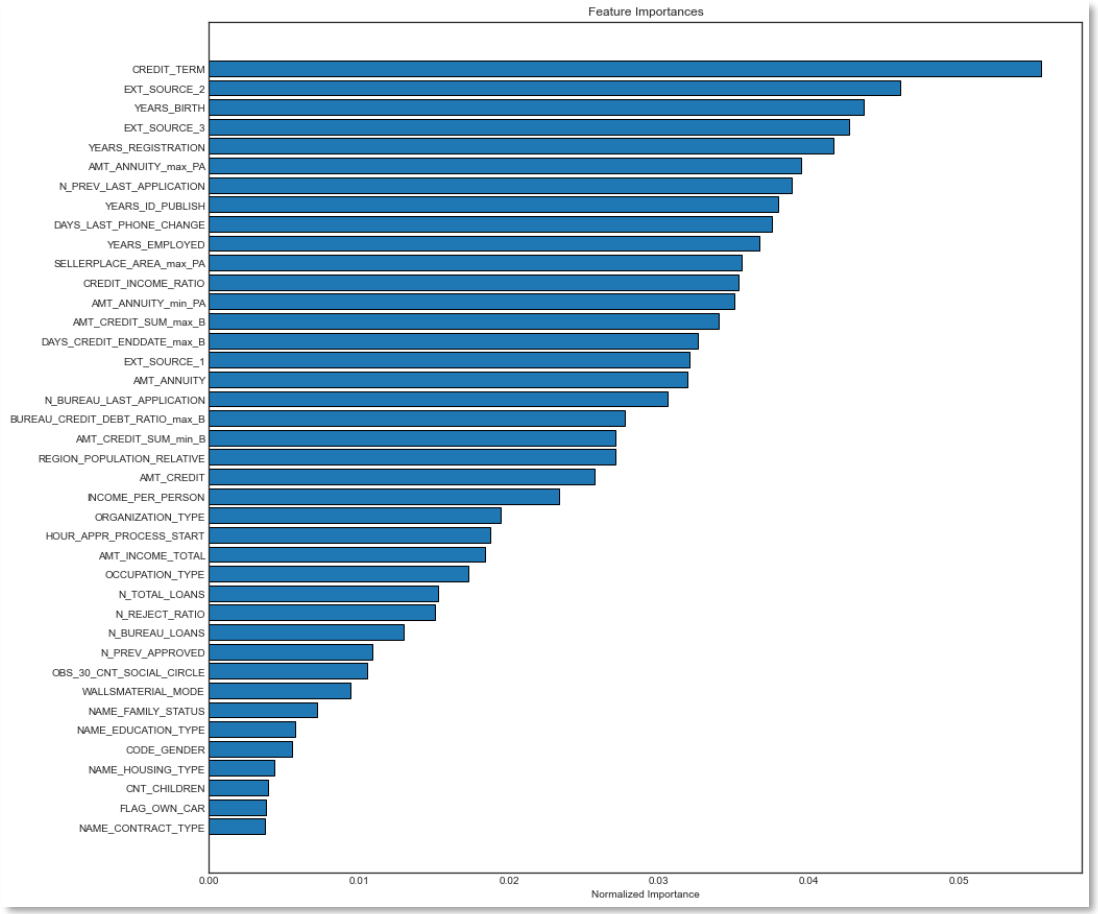
Affiche la description du feature sélectionné

Affiche le décryptage de la prédiction fait avec LIME

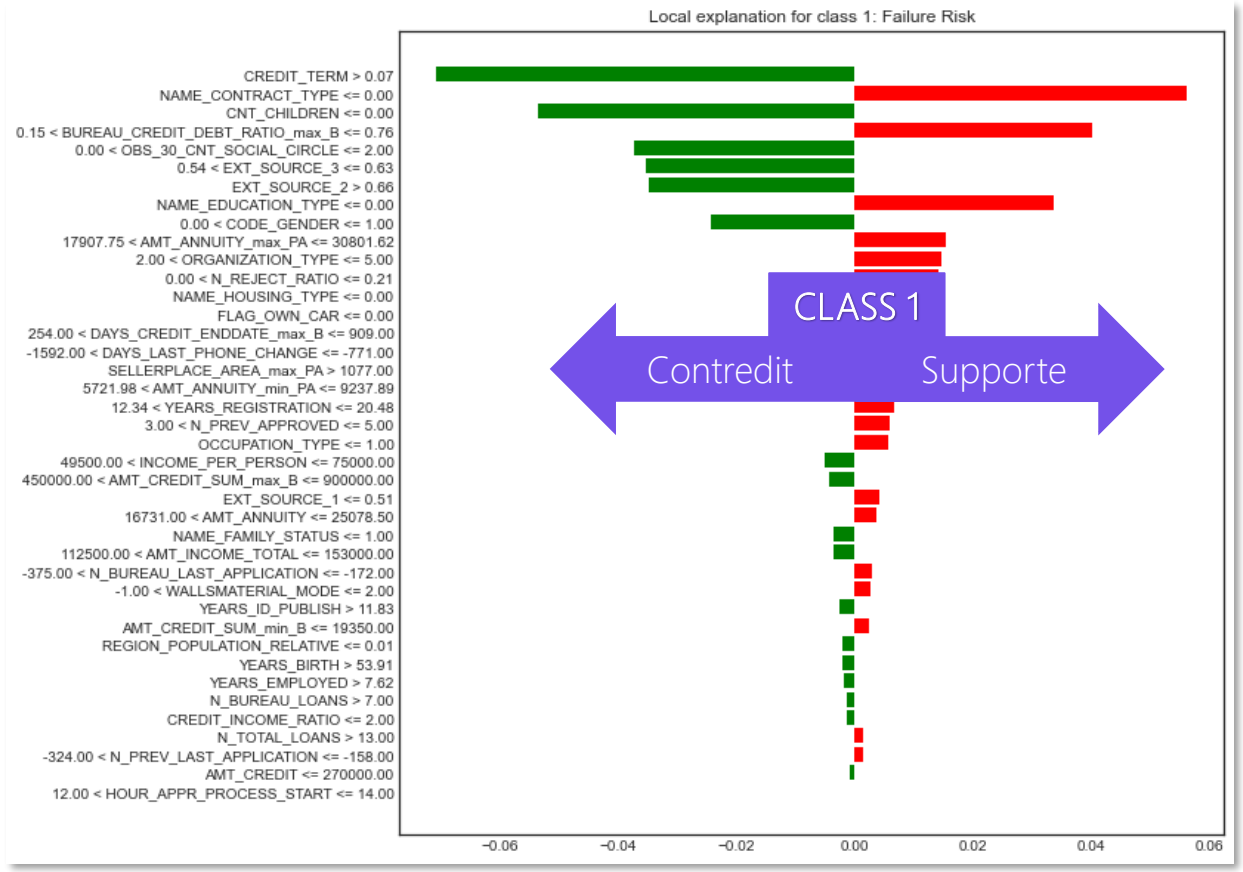
Interprétation SHAP statique

Dashboard – Interprétabilité

LightGBM : L'interprétabilité commence avec la vision globale du modèle que l'importance des features proportionne



LIME : L'importance globale d'un feature peut ne pas être la même dans le contexte local (et vice versa) également si le modèle a des centaines de variables (local fidelity)



Conclusions – Aller plus loin

Un modèle plus performant :

Même si SHAP est disponible sur le Notebook d'interprétation, son calcul est particulièrement lourd car il doit faire toutes les permutations possibles donnant comme résultat un fichier au-dessus de 100MB qui manque de réactivité lors de l'utilisation du Dashboard. Il faudrait améliorer la sélection des features ou le modèle dans son ensemble.

Améliorer le Dashboard :

- Explicabilité plus précise (notamment liée au point précèdent)
- Ajouter des graphiques interactifs (autre fois, liée à SHAP)
- Faire évoluer le scoring du client en même temps que les features sont modifiées (interactivité)

Maintenance :

Itération mensuelle de l'algorithme et contrôle de l'évolution des prédictions. Selon la situation, des nouveaux features peuvent être ajoutés ou peuvent remplacer des features existants.