IMPLÉMENTEZ UN MODÈLE DE SCORING DE LA PROBABILITÉ DE DÉFAUT DE PAIEMENT DE CLIENT

PRÉSENTATION PAR HORTENSE MONNARD

Problématique :

1. Construire un **modèle de scoring** qui donnera une prédiction de la probabilité qu'un client d'une voit sa demande de crédit acceptée ou refusée.

2. Accompagner le modèle d'un **dashboard interactif** permettant d'interpréter les prédictions faites par le modèle et d'améliorer la transparence vis-à-vis des clients quant à l'octroiement d'un crédit.

Objectifs :

- Permettre de visualiser le score et l'interprétation de ce score pour chaque client de manière simple ;
- Permettre de visualiser des informations descriptives relatives à un client ;
- Permettre de comparer les informations descriptives relatives d'un client à l'ensemble des clients.

PLAN

1. Présentation des jeux de données
2. Analyse Exploratoire
3. Création et Sélection de Features
4. Modélisation
5. Déploiement du modèle dans un Dashboard
6. Conclusions et Perspectives

1. Présentation des jeux de données

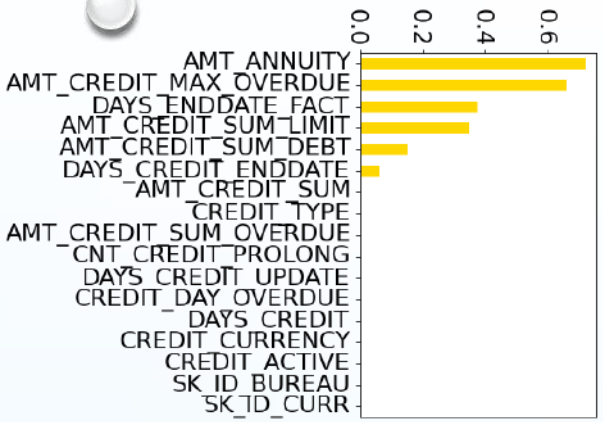
Observation des data sets

10 fichiers de données :

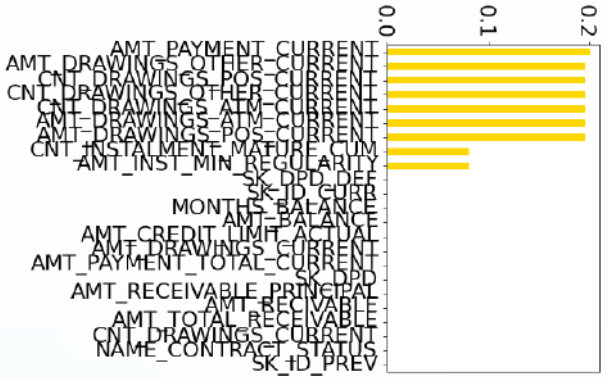
- application_train.csv (307511, 122)
- application_test.csv (48744, 121)
- bureau.csv (1716428, 17)
- bureau_balance.csv (27299925, 3)
- credit_card_balance.csv (3840312, 23)
- POS_CASH_balance.csv (10001358, 8)
- installments_payments.csv (13605401, 8)
- previous_application.csv (1670214, 37)
- sample_submission.csv
- HomeCredit_columns_description.csv

Le fichier HomeCredit_columns_description.csv donne des informations détaillées sur le sens des variables présentes dans les autres data sets.
Le fichier sample_submission.csv propose un exemple de soumission de demande de crédit.

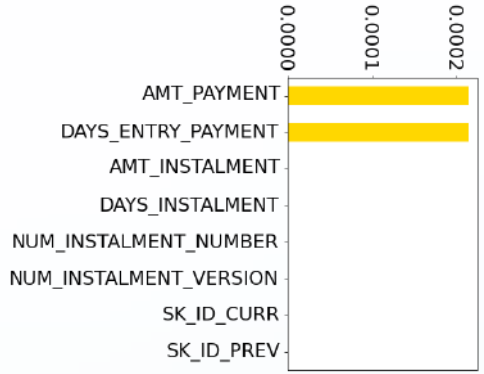
Taux de données manquantes par data set



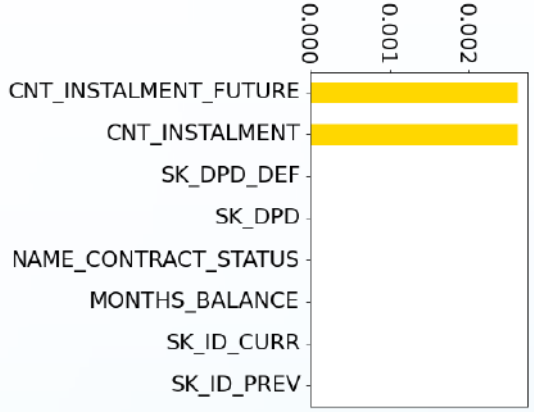
bureau



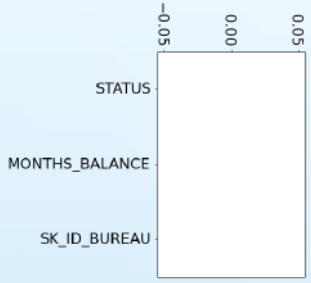
credit_card_balance



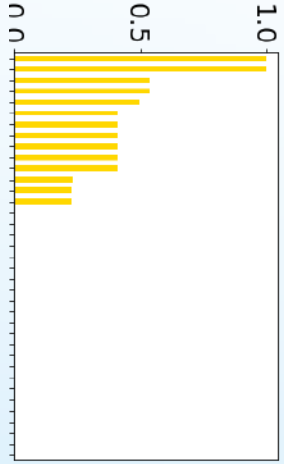
installments_payments



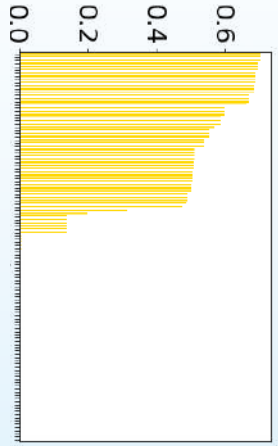
POS_CASH_balance



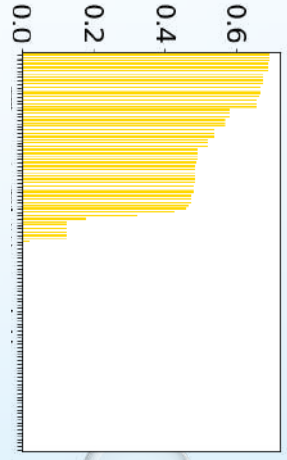
bureau_balance



previous_application



application_train



application_test

Certaines variables ont un taux de valeurs manquantes important > 50 %

Nettoyage des data sets

- Suppression des lignes pour lesquelles il y a un taux de données manquantes supérieur à 50% ;
- Suppression des lignes pour lesquelles il y a des données manquantes pour des valeurs jugées clefs ;

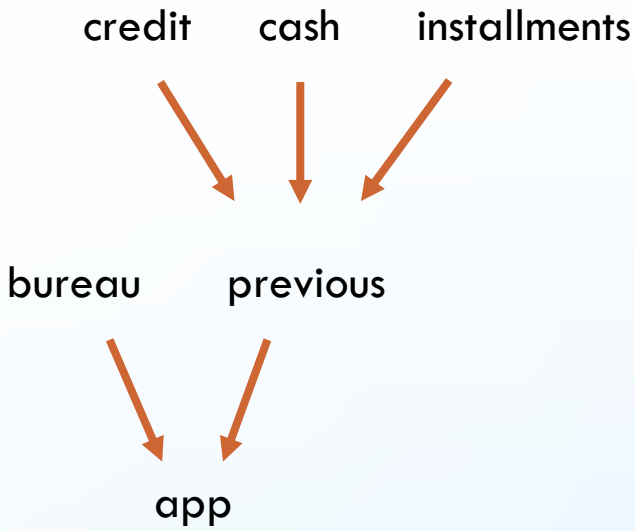
- Suppression des données aberrantes :

Valeurs qui semblent irréalistes en se servant de la méthode des percentiles (basée sur la Median Absolute Deviation et la Comparative Quantile Based Method).

- Imputation des données manquantes en se basant sur la médiane ou en remplaçant ces données manquantes par zéro.

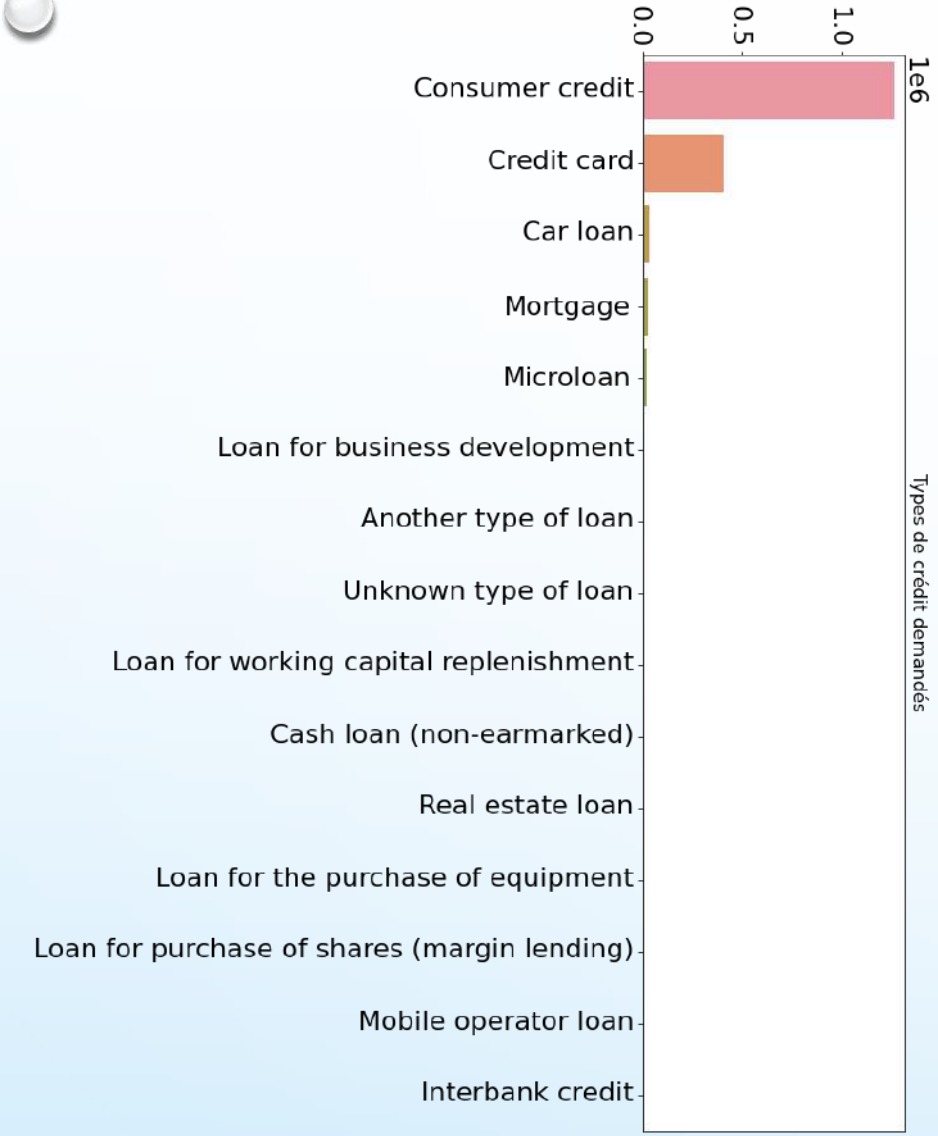
Création d'une matrice à partir d'une entité prenant en compte les relations entre les datasets

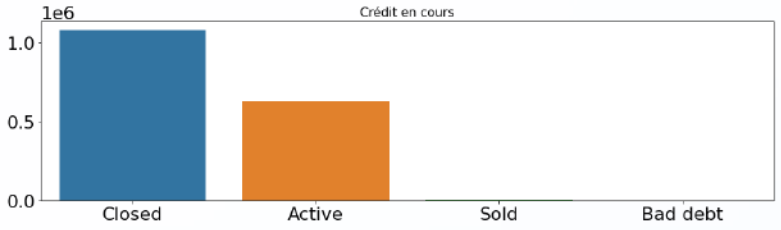
```
Entityset: clients
  Entities:
    app [Rows: 275907, Columns: 63]
    bureau [Rows: 234711, Columns: 16]
    previous [Rows: 632547, Columns: 34]
    cash [Rows: 10001358, Columns: 9]
    credit [Rows: 2122191, Columns: 24]
    installments [Rows: 12280690, Columns: 9]
  Relationships:
    bureau.SK_ID_CURR -> app.SK_ID_CURR
    previous.SK_ID_CURR -> app.SK_ID_CURR
    cash.SK_ID_PREV -> previous.SK_ID_PREV
    installments.SK_ID_PREV -> previous.SK_ID_PREV
    credit.SK_ID_PREV -> previous.SK_ID_PREV
```

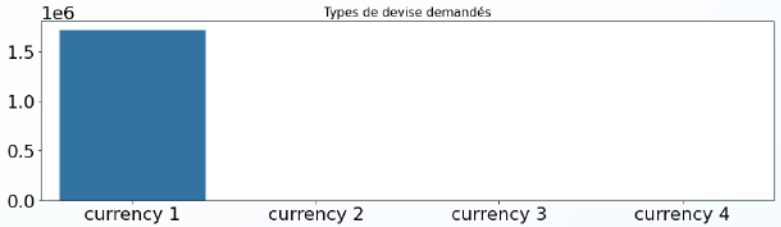


2. Analyse Exploratoire

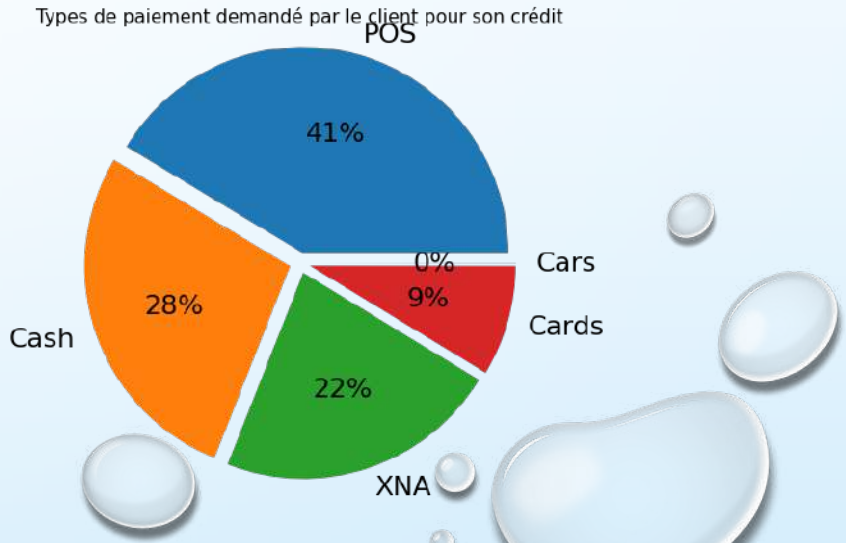
Informations sur les types de crédit demandés



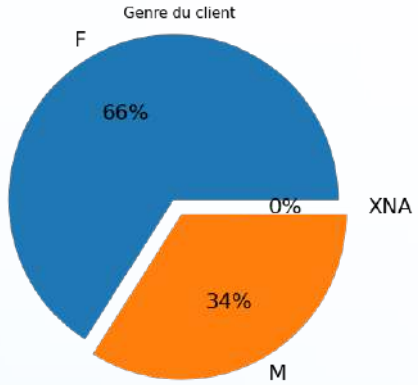


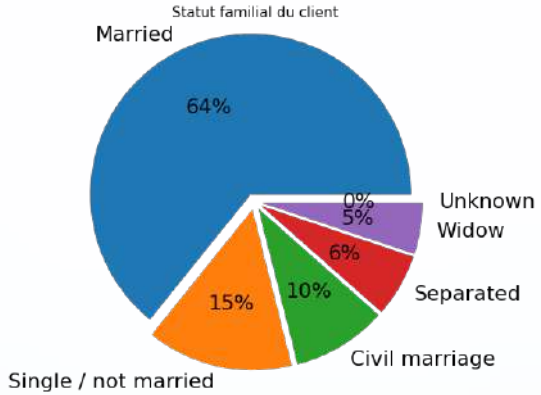


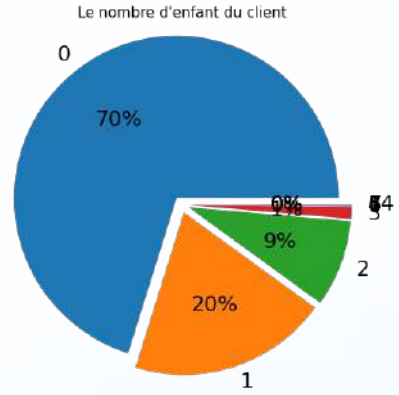
Crédit à la consommation, principalement en dollars.

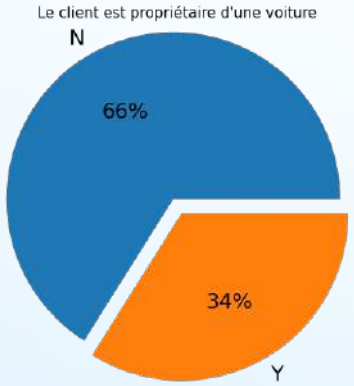


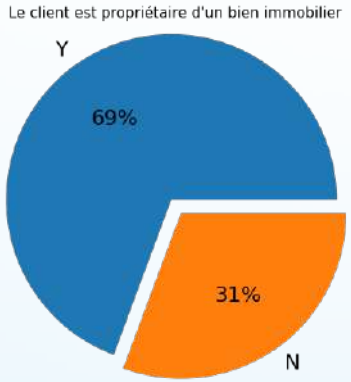
Informations sur les clients

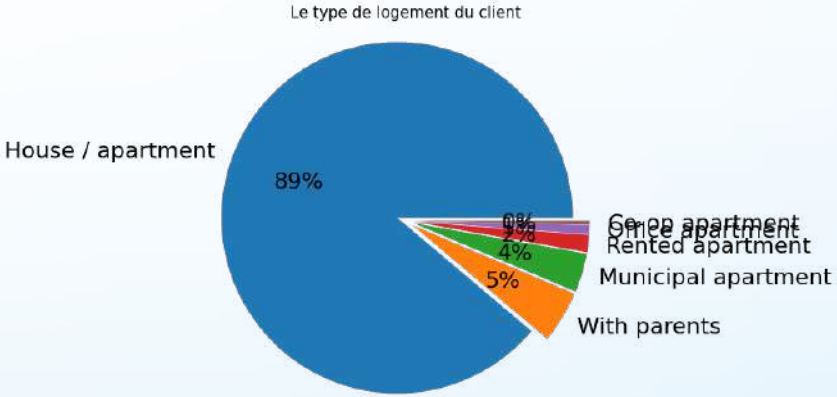






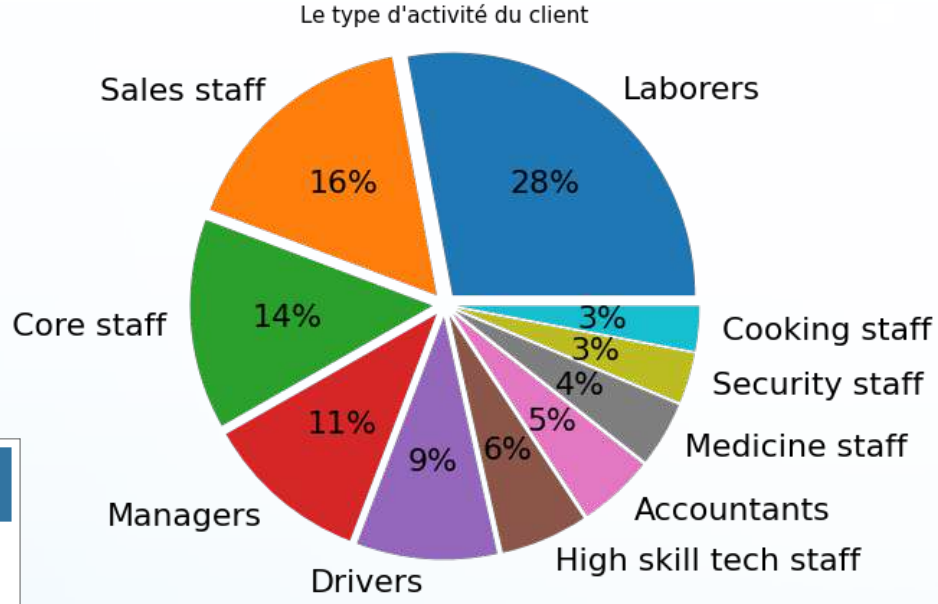


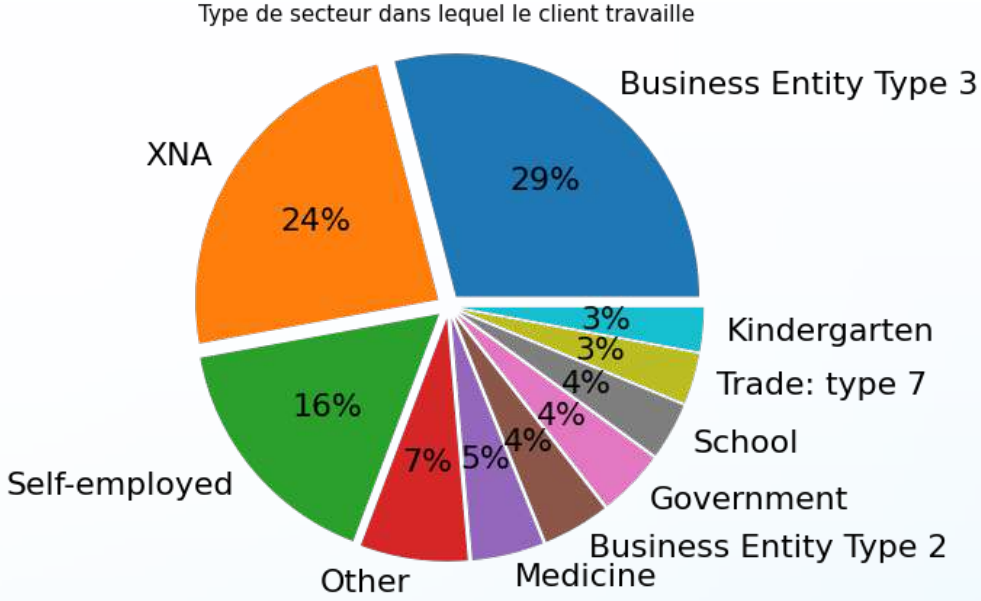


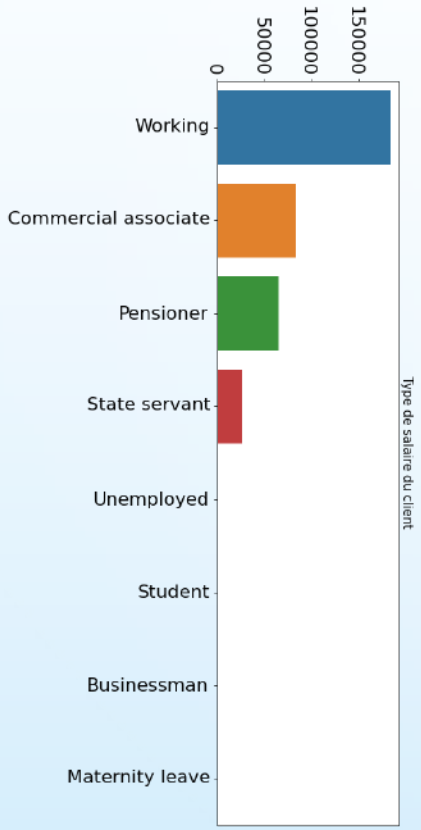


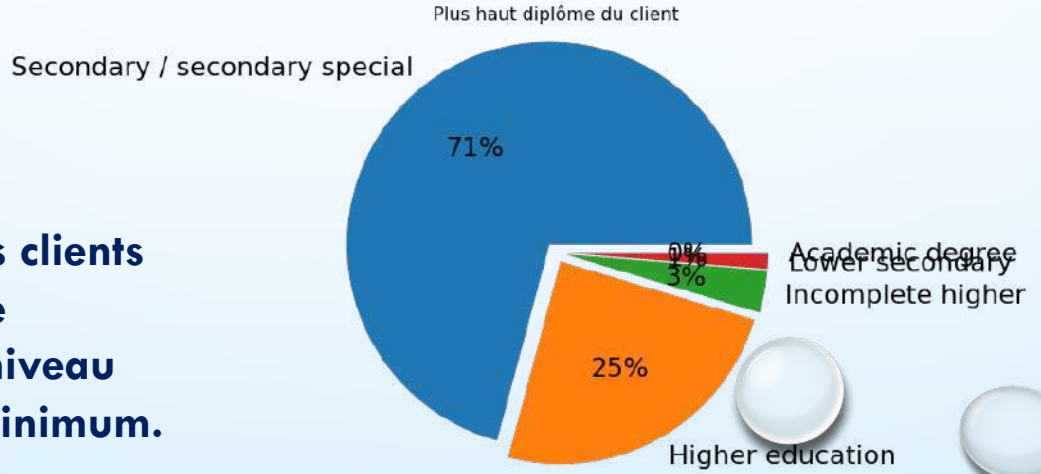
Une majorité des clients est mariée, sans enfant et vit indépendamment.

Informations sur les clients







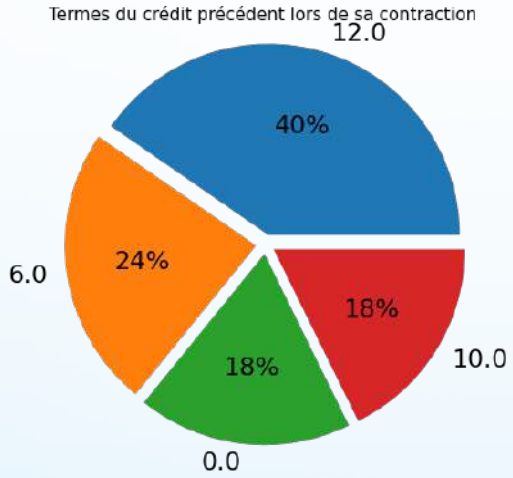


Une majorité des clients travaille et a une éducation d'un niveau secondaire au minimum.

Conclusions sur les informations présentes

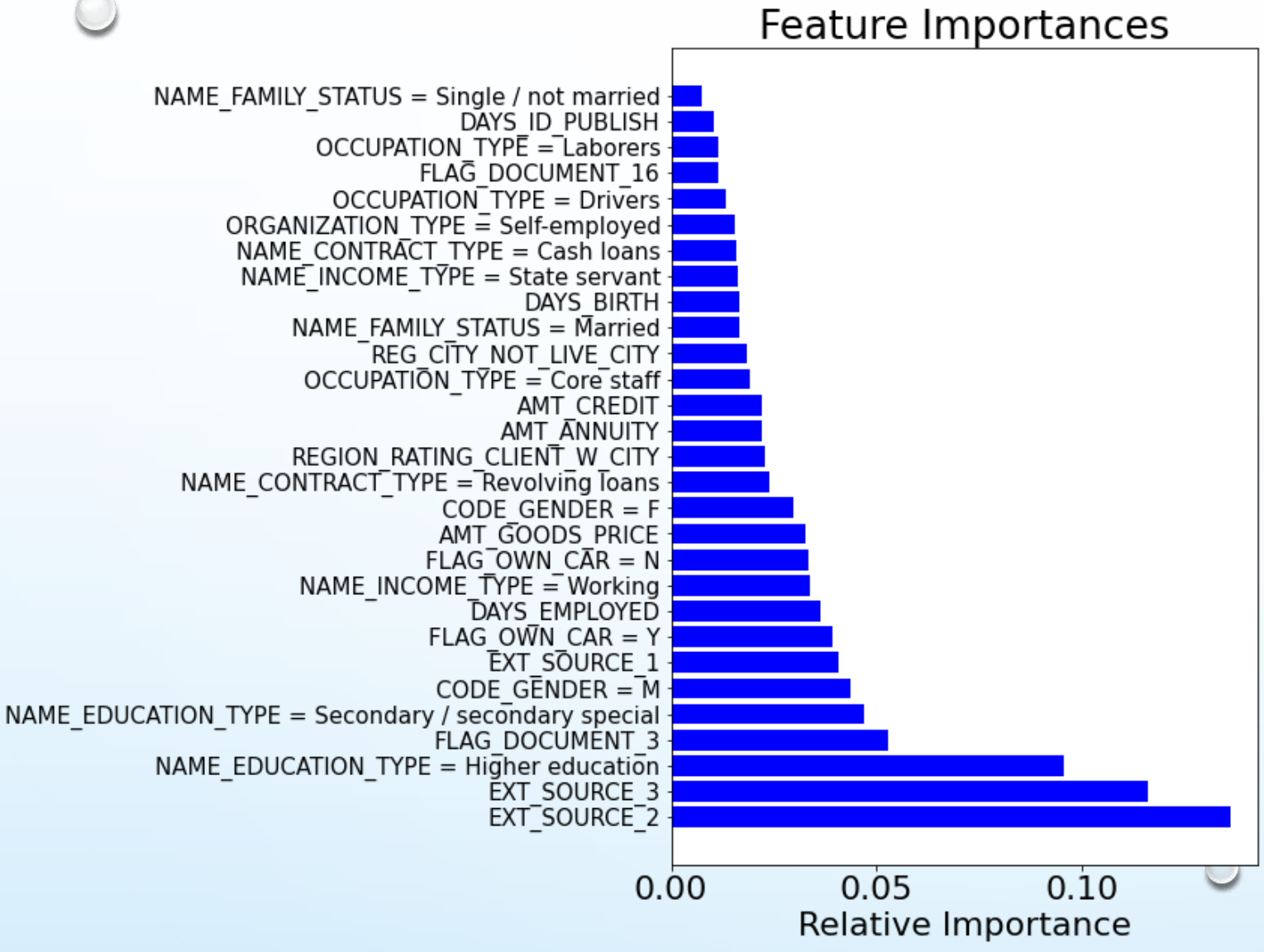
Certaines informations sont indisponibles :

- Certains documents on juste un numéro « FLAG » (1 à 19)
- Certaines sources sont juste décrites comme EXT_SOURCE_X (1, 2, 3)
- Les valeurs que peuvent prendre certaines variables sont indiquée sous la forme d'un numéro.



3. Sélection de Features

Sélection des features



Sélection des features grâce à la fonction SelectFromModel, en utilisant comme estimateur un modèle XGBoostClassifier.

Sur 123 features initiales, 29 sont sélectionnées.

4. Modélisation

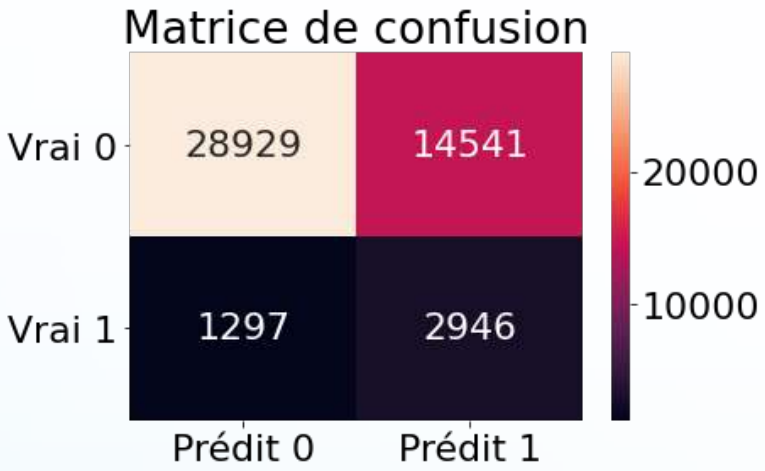
Indices clefs de la modélisation

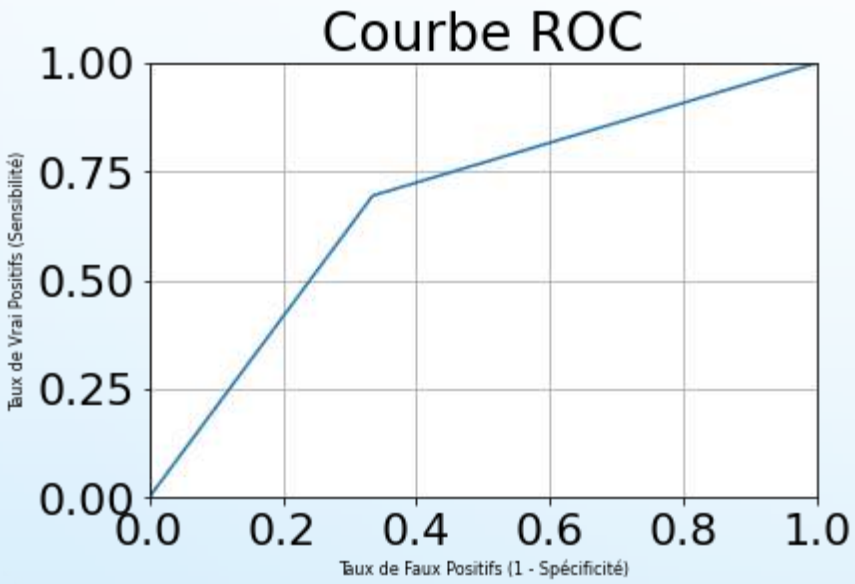
Fonction coût:
Ici, on considère que la fonction coût est l'inverse du **score auc**. Cette fonction coût doit être réduite au maximum. On cherche donc à maximiser l'espace sous la courbe roc.

Algorithme d'optimisation:
Afin de réduire la fonction coût de manière optimale, le **seuil de discrimination** établit à partir des probabilités de prédiction de classification dans une catégorie ou l'autre (TARGET= 1 ou TARGET= 0) est optimisé.

Métriques d'évaluation pour une classification binaire (TARGET= 1 ou TARGET= 0)**:**
- Le **score auc** : espace sous la courbe ROC. Vrais Positifs en fonction du taux de Faux Positifs ;
- Le **f-1 score** : combine une mesure de la précision (Vrais Positifs / (Vrais Positifs + Faux Positifs)) et du rappel (Vrais Positifs / (Vrais Positifs + Faux Négatifs)) ;
- Le **coefficient de corrélation de Matthew** : une mesure équilibrée qui prend en compte toutes les classes de la matrice de confusion (Vrais Positifs, Faux Positifs, Vrais Négatifs et Faux Négatifs).

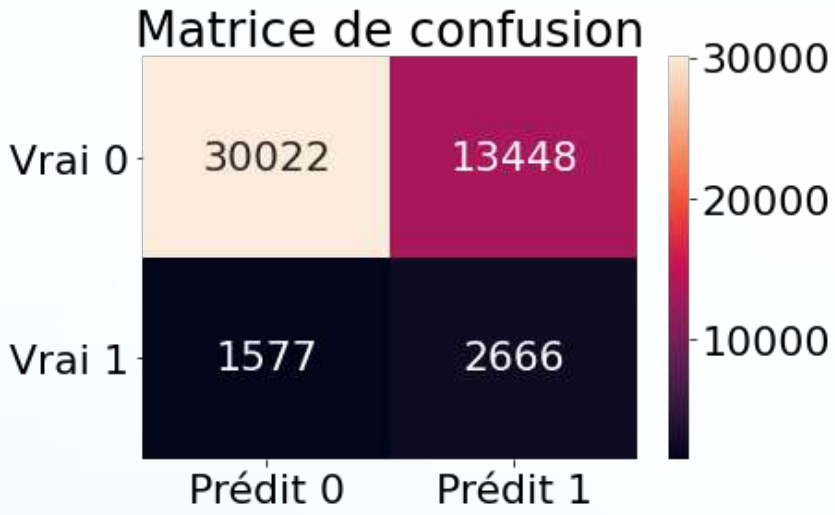
Régression Logistique

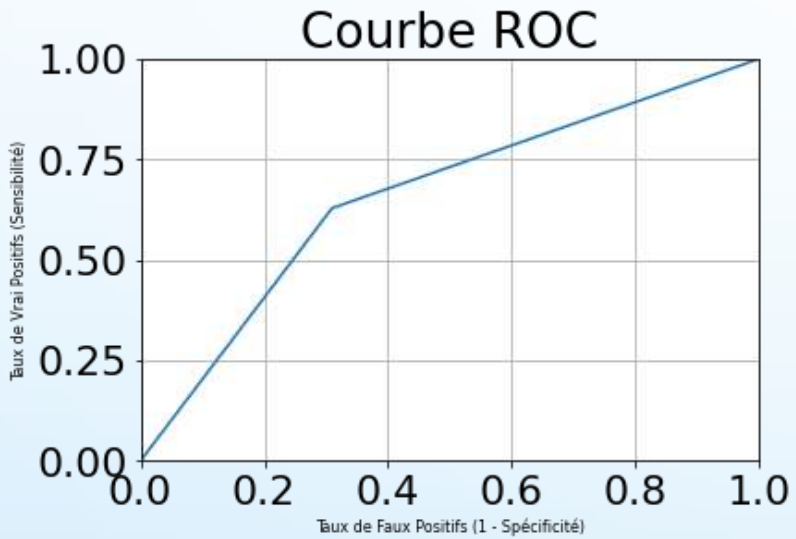





```
**********************************************************
Détermination du seuil optimal
Le threshold optimal est : 0.083
Le score AUC optimal est : 0.6799067501590327
**********************************************************
Prédictions avec le seuil optimisé
Temps Entrainement 10.31s
Temps Test 0.02s
**********************************************************
Statistiques du Test Set
Accuracy score: 0.6680569236895605
AUC score: 0.6799067501590327
F1 score: 0.27114588127013345
Matthews correlation coefficient: 0.2125498467318124
**********************************************************
Statistiques du Train Set
Accuracy score: 0.6653899356569488
AUC score: 0.6794358870380189
F1 score: 0.26550117890620506
Matthews correlation coefficient: 0.2094618669026185
**********************************************************
  Variable Stat  RegLogistique_test  RegLogistique_entrainement
0      accuracy            0.668057                    0.665390
1           AUC            0.679907                    0.679436
2            F1            0.271146                    0.265501
3           MCC            0.212550                    0.209462
4  Temps (en s)            0.023140                   10.308178
<class 'pandas.core.frame.DataFrame'>
**********************************************************
Vrais Positifs =  2946
Vrais Négatifs =  28929
Faux Positifs =  14541
Faux Négatifs =  1297
**********************************************************
Spécificité =  0.6654934437543133
Taux de Faux Positifs =  0.3345065562456867
**********************************************************
```

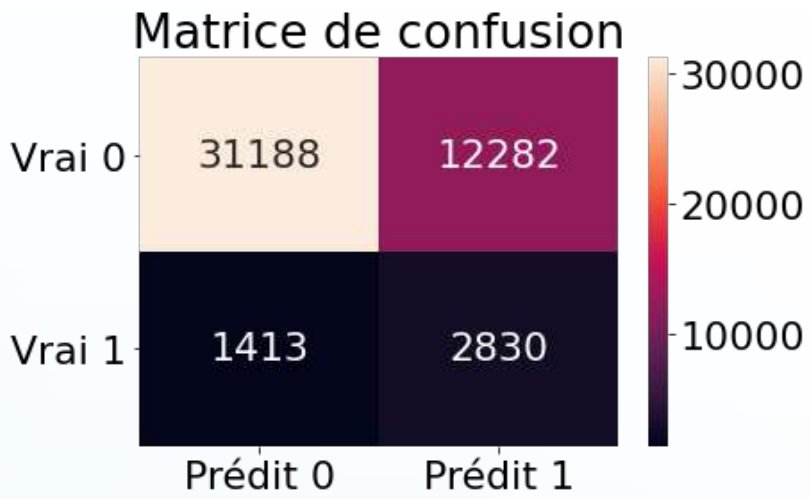
Random Forest Classifier

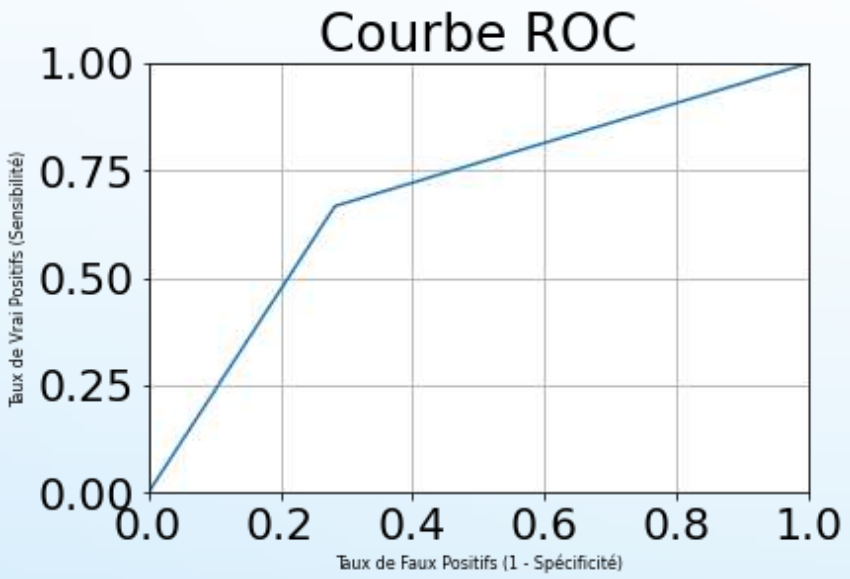





```
**********************************************************
Détermination du seuil optimal
Le threshold optimal est : 0.099
Le score auc optimal est : 0.6594831167815829
**********************************************************
Prédictions avec le seuil optimisé
Temps Entrainement 96.33s
Temps Test 0.85s
**********************************************************
Statistiques du Test Set
Accuracy score: 0.6850963049902542
AUC score: 0.6594831167815829
F1 score: 0.26192464508522867
Matthews correlation coefficient: 0.1919719149362709
**********************************************************
Statistiques du Train Set
Accuracy score: 0.8552962504977679
AUC score: 0.9200580339157418
F1 score: 0.5451071469750127
Matthews correlation coefficient: 0.5610741778332595
**********************************************************
  Variable Stat  Random Forest Test  Random Forest Entrainement
0      accuracy            0.685096                    0.855296
1           AUC            0.659483                    0.920058
2            F1            0.261925                    0.545107
3           MCC            0.191972                    0.561074
4  Temps (en s)            0.845892                   96.326613
<class 'pandas.core.frame.DataFrame'>
**********************************************************
Vrais Positifs =  2666
Vrais Négatifs =  30022
Faux Positifs =  13448
Faux Négatifs =  1577
**********************************************************
Spécificité =  0.6906372210720036
Taux de Faux Positifs =  0.3093627789279963
**********************************************************
```

XG Boost Classifier

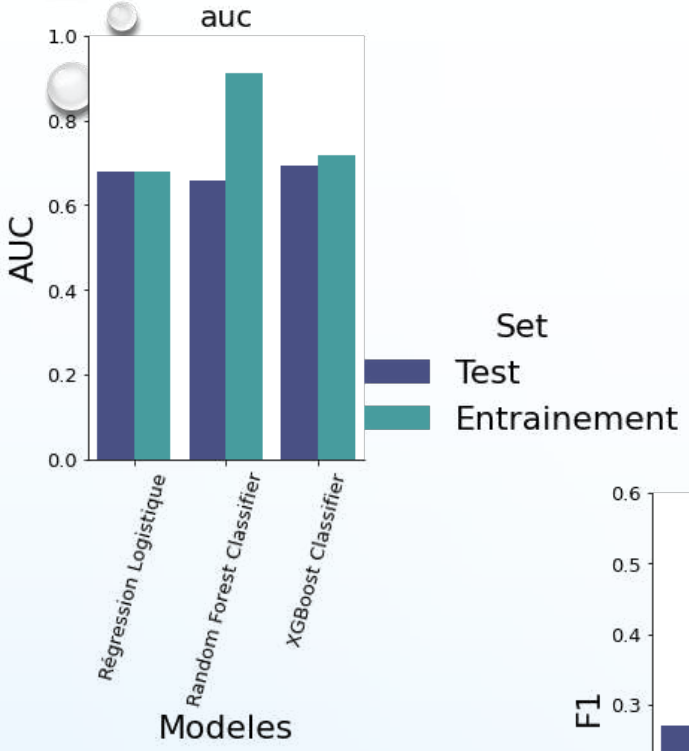


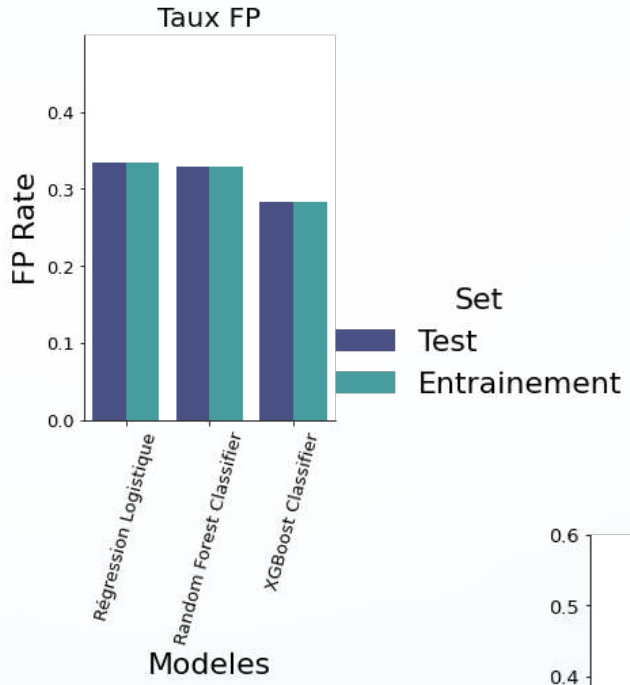



```
**********************************************************
Détermination du seuil optimal
Le threshold optimal est : 0.092
L'auc score optimal est : 0.6922206135969983
**********************************************************
Prédictions avec le seuil optimisé
Temps Entrainement 256.34s
Temps Test 2.85s
**********************************************************
Statistiques du Test Set
Accuracy score: 0.7129713076100853
AUC score: 0.6922206135969983
F1 score: 0.2924308964091966
Matthews correlation coefficient: 0.2352257476294144
**********************************************************
Statistiques du Train Set
Accuracy score: 0.7212185358288098
AUC score: 0.7193414893860381
F1 score: 0.30877958791280175
Matthews correlation coefficient: 0.26560022736506
**********************************************************
  Variable Stat  XGBoost Test  XGBoost Entrainement
0      accuracy      0.712971              0.721219
1           AUC      0.692221              0.719341
2            F1      0.292431              0.308780
3           MCC      0.235226              0.265600
4  Temps (en s)      2.850174            256.340018
<class 'pandas.core.frame.DataFrame'>
**********************************************************
Vrais Positifs =  2830
Vrais Négatifs =  31188
Faux Positifs =  12282
Faux Négatifs =  1413
**********************************************************
Spécificité =  0.7174603174603175
Taux de Faux Positifs =  0.28253968253968254
**********************************************************
```

Le seuil de discrimination par le modèle choisi (XGBoostClassifier) est de 0.092. Si la probabilité d'avoir TARGET=1 est au dessus de ce seuil, le client devrait être éligible à un crédit.

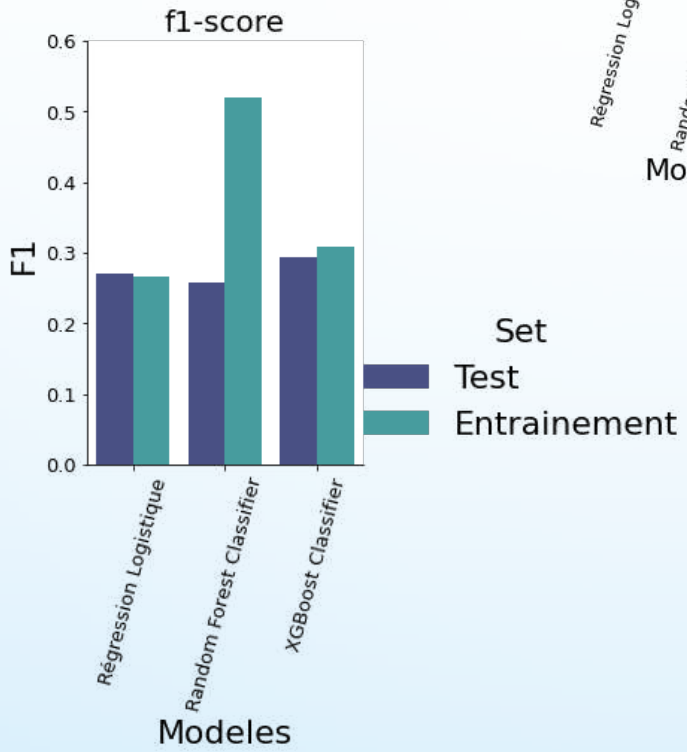
Comparaison des performances des modèles

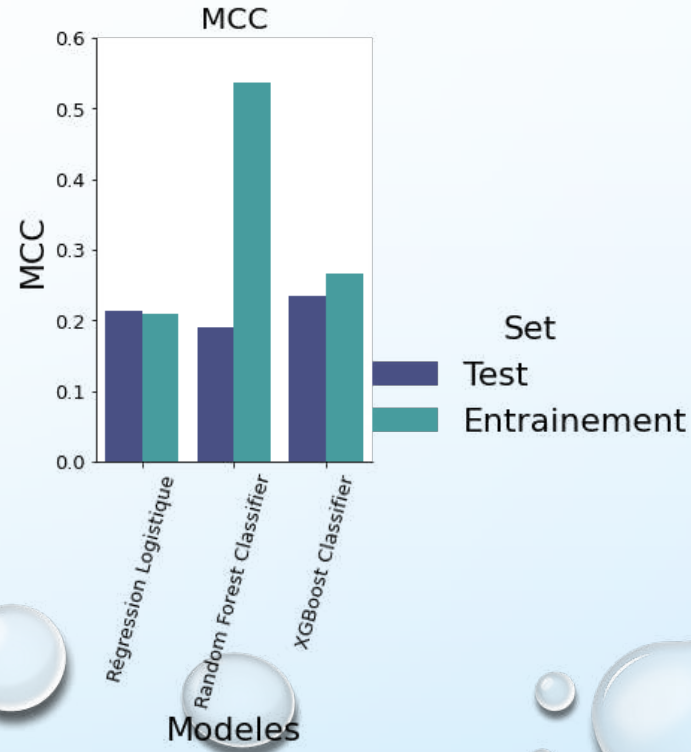




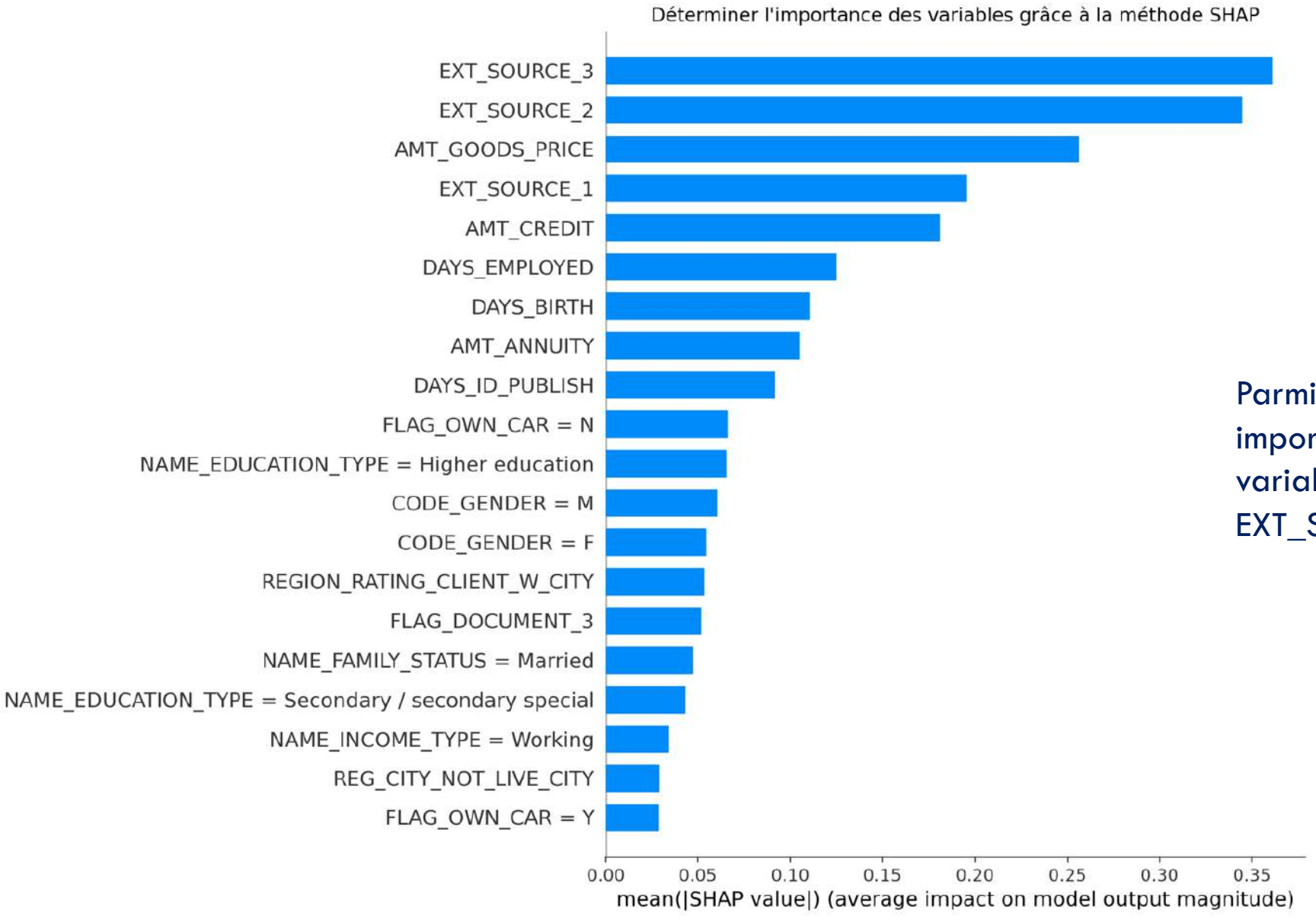
Avec XG Boost Classifier :

- **Meilleur AUC**
- **Moins de Faux Positifs**
- **Modèle stable**



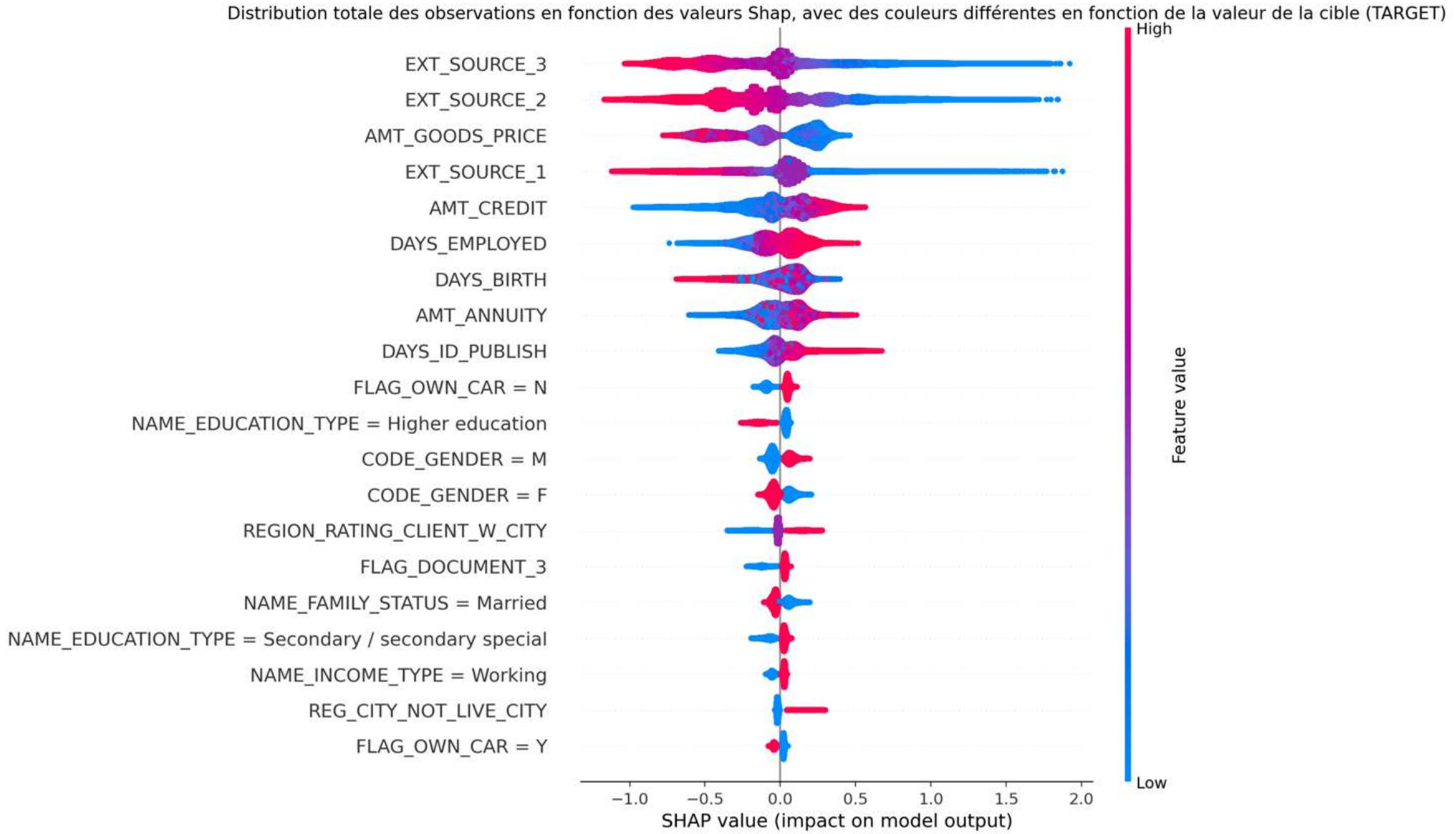


Visualiser l'impact des features sur la prédiction

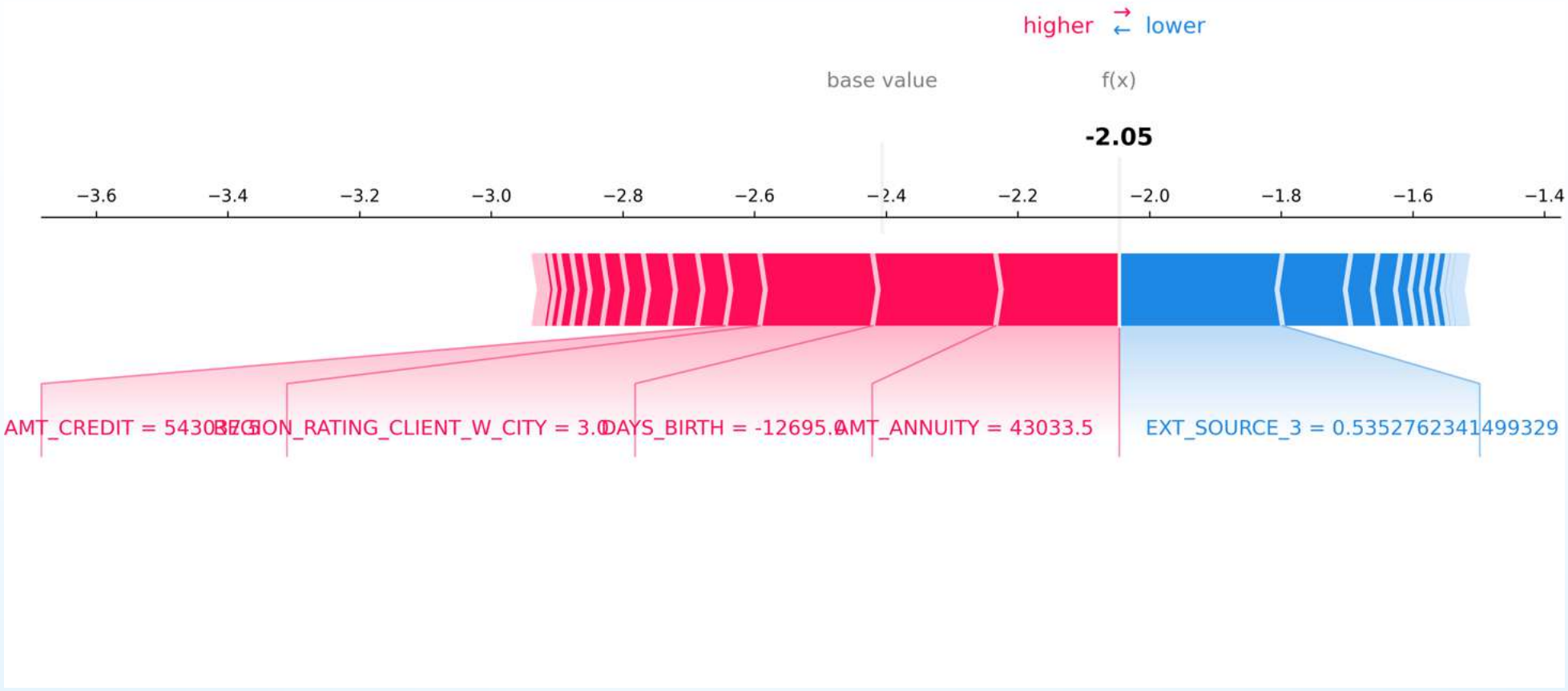


Parmi les variables les plus importantes on retrouve des variables dont on ne connait rien : EXT_SOURCE 1, 2, 3.

Déterminer le type d'impact de chaque feature sur la prédiction



Visualisation détaillée par client du type d'impact de chaque feature sur la prédiction



Ceci permet d'expliquer aux clients la motivation de la décision et de les aider à améliorer la probabilité que leur prochaine demande de crédit soit acceptée.

5. Déploiement du modèle dans un Dashboard

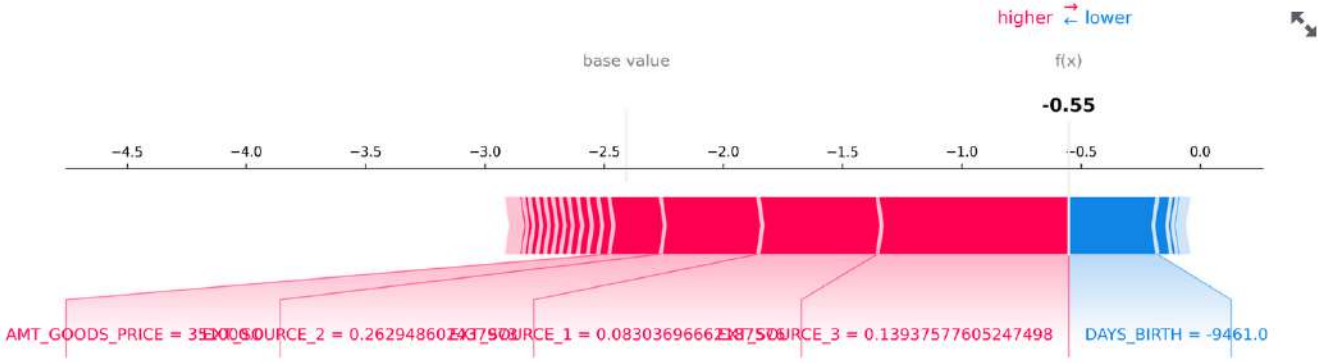
Dashboard pour accéder à la modélisation et l'analyse de l'impact de chaque feature sur la prédiction

En rentrant l'identifiant du client on peut voir s'afficher la réponse à la demande de crédit du client et les features qui ont principalement jouées dans la décision, ainsi que le type d'impact d'elles ont eu.

Observer la corrélation des différentes variables à la prédiction pour chaque client

Entrer l'identifiant client, IDD_CURR:

100002 − +



Etat de la demande de crédit pour le client :

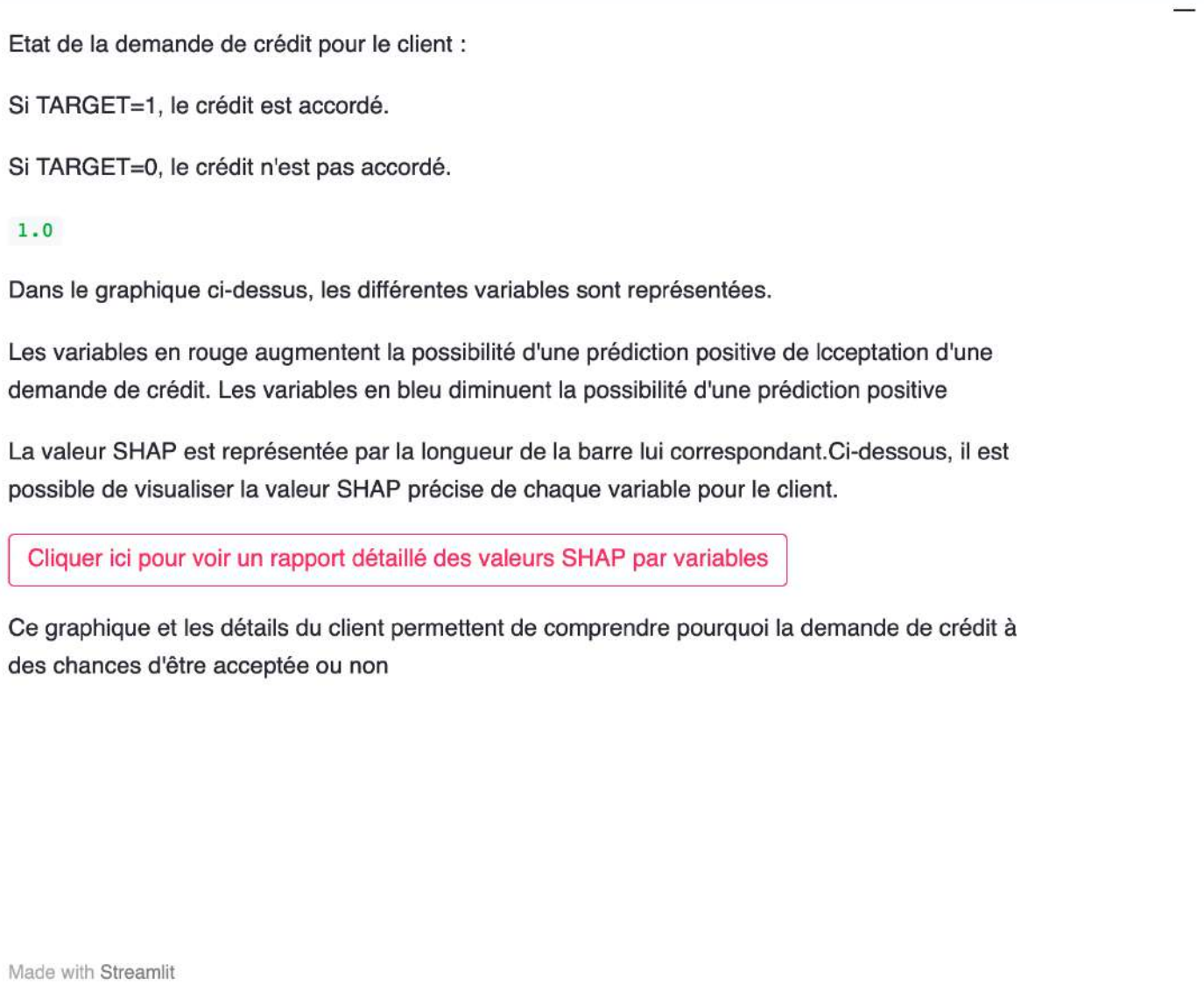
Si TARGET=1, le crédit est accordé.

Si TARGET=0, le crédit n'est pas accordé.

1.0

Dashboard pour accéder à la modélisation et l'analyse de l'impact de chaque feature sur la prédiction

Possibilité d'avoir un rapport détaillé par client où toutes les valeurs SHAP sont présentées pour chaque variable.
Ceci permet au conseiller d'aider le client à améliorer ses chances d'acceptation pour une prochaine demande de crédit.



Dashboard pour accéder comparer les features du client à celles des autres clients

	100002
NAME_CONTRACT_TYPE = Cash loans	0.0098
NAME_CONTRACT_TYPE = Revolving loans	0.0047
CODE_GENDER = F	0.0254
CODE_GENDER = M	0.0420
FLAG_OWN_CAR = N	0.0325
FLAG_OWN_CAR = Y	0.0205
AMT_CREDIT	-0.0243
AMT_ANNUITY	0.0081
AMT_GOODS_PRICE	0.2158
NAME_INCOME_TYPE = Working	0.0111
NAME_INCOME_TYPE = State servant	0.0052
NAME_EDUCATION_TYPE = Secondary / secondary special	0.0329
NAME_EDUCATION_TYPE = Higher education	0.0384
NAME_FAMILY_STATUS = Married	0.0329

- Comparaison aux moyennes des clients dont la prédiction est positive ou négative.

- Comparaison à tous les clients, de manière générale.

	NAME_CONTRACT_TYPE = Cash loans	NAME_CONTRACT_TYPE = Revolving loans	CODE_GENDER = F	CODE_GENDER = M	FLAG_OWN_CAR = N	FLAG_OWN_CAR = Y	AMT_CREDIT	AMT_A
0	-0.0035	-0.0012	-0.0073	-0.0093	-0.0029	-0.0016	-0.0112	
1	0.0034	0.0012	0.0076	0.0060	0.0066	0.0026	0.0075	

	NAME_CONTRACT_TYPE = Cash loans	NAME_CONTRACT_TYPE = Revolving loans	CODE_GENDER = F	CODE_GENDER = M	FLAG_OWN_CAR = N	FLAG_OWN_CAR = Y	AMT_CREDIT	A
count	238565	238565	238565	238565	238565	238565	238565	
mean	-0.0029	-0.0010	-0.0060	-0.0080	-0.0021	-0.0012	-0.0096	
std	0.0474	0.0163	0.0605	0.0645	0.0708	0.0312	0.2292	
min	-0.3031	-0.1086	-0.1487	-0.1362	-0.1795	-0.0794	-0.9754	
25%	0.0094	0.0030	-0.0525	-0.0587	-0.0841	-0.0367	-0.1441	
50%	0.0116	0.0040	-0.0290	-0.0378	0.0411	0.0175	-0.0021	
75%	0.0137	0.0049	0.0498	0.0560	0.0519	0.0228	0.1603	
max	0.0259	0.0096	0.2074	0.1962	0.1114	0.0510	0.5649	

6. Conclusion et Perspectives

Conclusions :

- Les métriques a réduire au maximum lors de la modélisation des prédictions d'acceptation de crédit sont : le score auc et le taux de faux positifs.
- Le modèle XGBoost Classifier, avec optimisation bayésienne des hyper-paramètres, est le modèle qui s'est montré le plus performant.
- Les features les plus importantes pour la prédiction de l'acceptation d'un crédit sont des variables pour lesquelles on a très peu d'informations (ex: EXT_SOURCE_1 , EXT_SOURCE_2, EXT_SOURCE_3, FLAG_DOCUMENT_3).

Perspectives :

- Les performances de la modélisation pourraient encore être améliorées et il conviendrait d'essayer d'autres modèles de utilisant des gradients de boosting, comme LightGBM ou CatBoost.
- Le pre-processing des données pourrait être amélioré grâce à l'apport des connaissances métier par d'autres membres de l'équipe.
- Identifier les variables extérieurs qui ont un impact sur la modélisation serait aussi important pour la transparence vis-à-vis des clients.

MERCI !