Projet 7 Note Méthodologique

Sofiane Mouhab - Sept 2021

Sommaire :

1 - Présentation

1.1 Les Objectifs

L’objectif est l’implémentation d'un modèle d'apprentissage supervisé pour une application de Crédit Scoring suivant plusieurs parametres:

- Le modèle doit permettre de définir la probabilité de défaut de remboursement d'un crédit à partir des informations sur le client
- Il doit également offrir un certain niveau de transparence afin de permettre aux conseillers de justifier la réponse

1.2 Les données en présence

Le jeu de données est constitué d'informations relatives aux crédits en cours et autres informations externes.

Une partie du jeu concerne des crédits terminés accompagnés d’une valeur binaire (1 ou 0) nous indiquant les difficultés à rembourser le crédit, dataset référence pour effectuer nos tests de prédictions

L’autre partie concerne les crédits en cours et nécessite donc une application de notre modèle pour prédire les difficultés de paiement

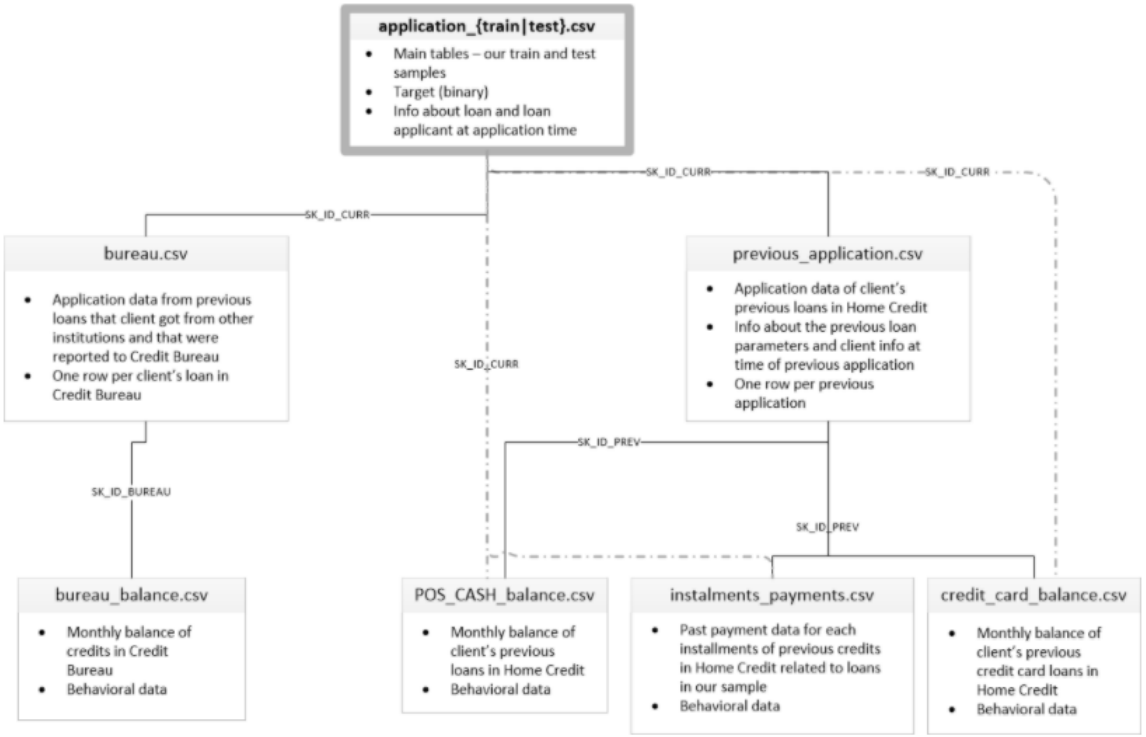
1.3 Modélisation

Cette étude se fait sous un biais, il s’agit de prédire au maximum les non-paiements d’un client. à nous donc de trouver un “seuil” à partir duquel la probabilité de difficulté du client est à envisager.

Pour cela nous allons évidemment comparer les résultats de notre travail à la base de donnée évoqué plus haut, et ainsi déterminer le seuil adéquate en accord avec notre client

2 - Analyse Exploratoire

2.1 Présentation



- Nous sommes en présence de 7 datasets différents , pour cette étude préalable nous allons nous intéresser à la base de données “Application-Test-Train”. Focus sur ce dernier :
- 307533 lignes de client
- 122 colonnes contenant des infos telle que :
    - Id du client
    - Difficulté de paiement (1 ou 0)
    - Somme empruntée
    - Âge
    - Durée depuis la prise de poste
    - Emploi
    - Nombre d’enfant
    - Nombre de personne vivant sous le même toit
    - Catégorie socioprofessionnelle
    - ...

2.2 Nettoyage

2.2.1 Par Colonne

Cette base de données contient 24% de cellules vides, notre première étape est de filtrer les colonnes qui contiennent peu d'informations.

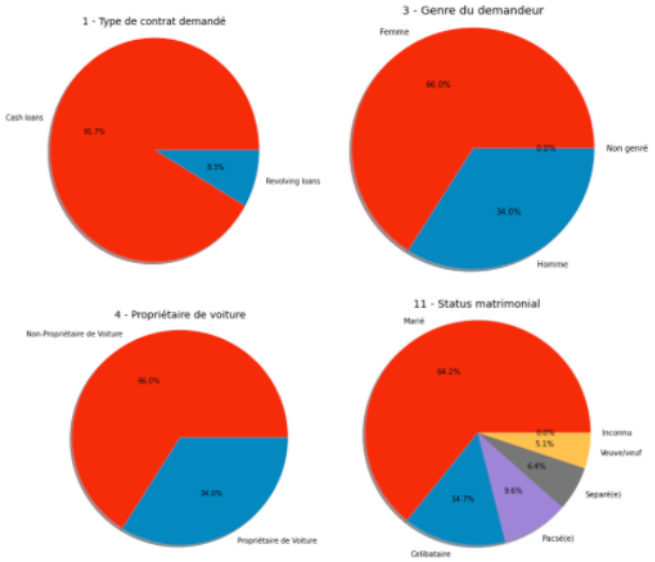
En éliminant les colonnes avec la moitié de cellules vides, nous retenons 73 Colonnes

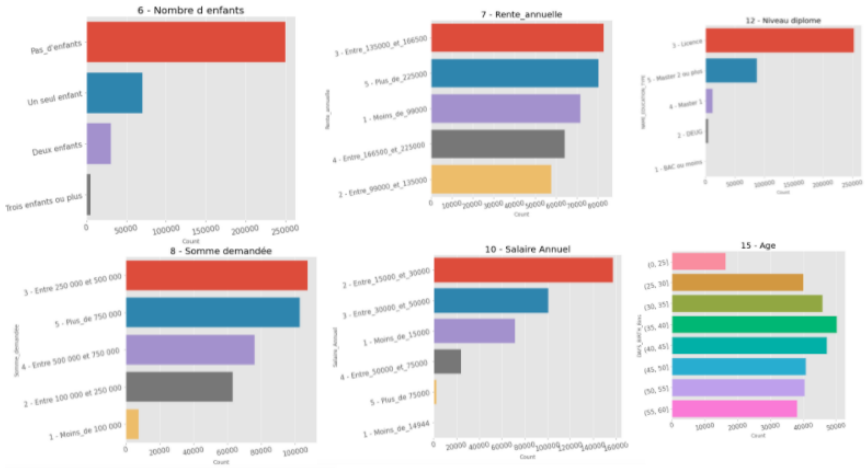
2.2.2 Par Lignes

La même manière nous effectuons un filtre par client, ceux qui disposent de peu d'information viable sont retirés du jeu de données, nous obtenons ainsi une base de 265554 lignes sur 73 colonnes

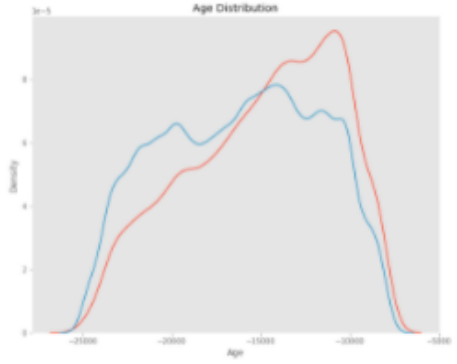
2.3 Visualisation des données

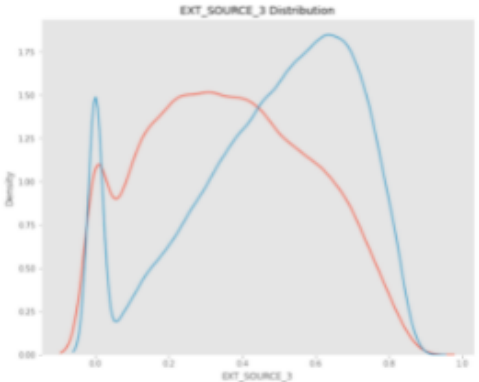
2.3.1 Univarié

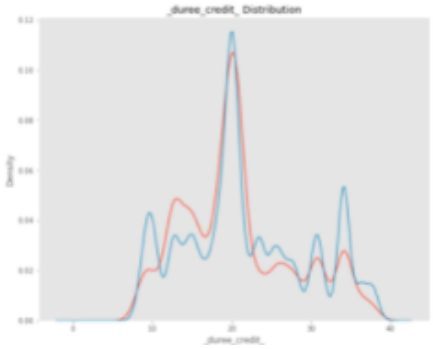


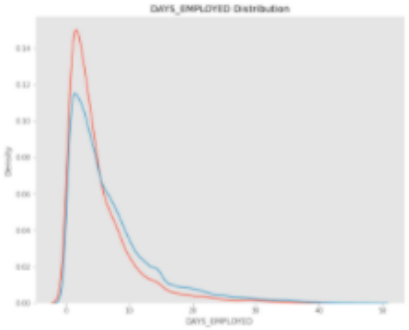


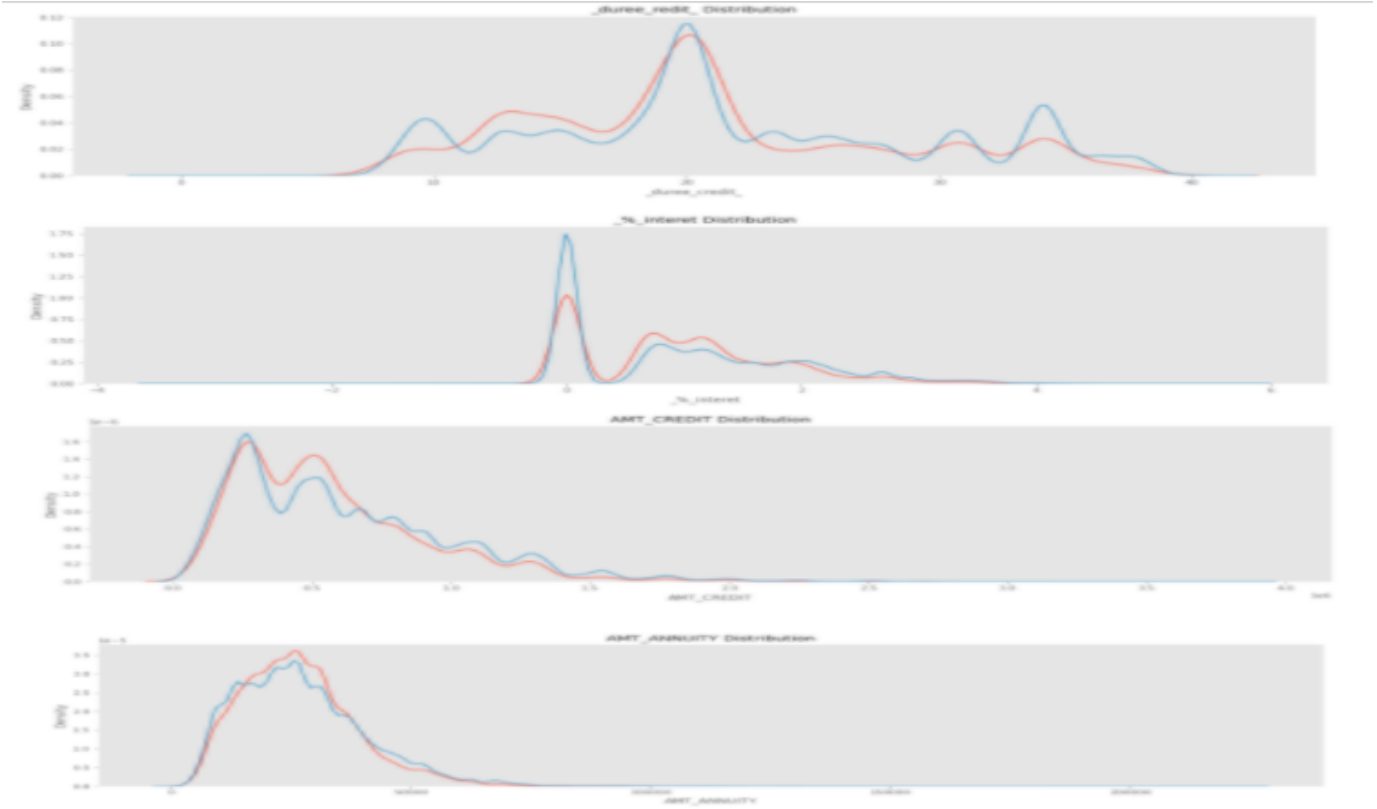
2.3.1 Bivarié


Age Distribution
Density
Age


EXT_SOURCE_3 Distribution
Density
EXT_SOURCE_3


_duree_credit_ Distribution
Density
_duree_credit_


DAYS_EMPLOYED Distribution
Density
DAYS_EMPLOYED


_duree_credit_ Distribution
_%_interet Distribution
AMT_CREDIT Distribution
AMT_ANNUITY Distribution

3 - Traitement des données

3.1 - Feature engineering

A partir des données nous pouvons créer quelques variables qui semblent pertinente :

- Rapport Salaire/Somme demandée *(AMT INCOME TOTAL/AMT CRÉDIT)*
- Durée du crédit *(AMT INCOME TOTAL/AMT ANNUITY)*
- Intérêt du prêt (AMT CRÉDIT * DURÉE CREDIT * (1/ AMT CRÉDIT- AMT GOODS PRICE))
- Part fiscale *(CNT FAM MEMBERS - COUNT CHILDRENS/2)*
- Taux endettement*(AMT CRÉDIT/DURÉE* CREDIT*/12) / (AMT INCOME TOTAL/12)*

3.2 - Refactor

Certaines colonnes se prêtent à un léger refactoring des données pour permettre un meilleur traitement de notre étude:

3.2.1 NAME EDUCATION TYPE

Regroupement en 4 catégories :

- "Higher"
- "Secondary"
- "Low"
- "Secondary"

3.2.2 OCCUPATION TYPE

Regroupement en 3 catégories :

- "CSP_1"
- "CSP_2"
- "CSP_3"

3.2.3 NAME INCOME TYPE

Regroupement en 2 catégories :

- "Unemployed"
- “Working”

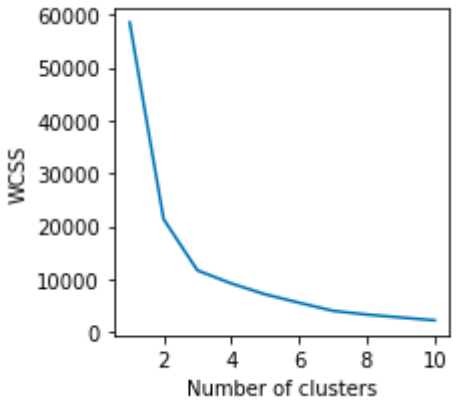
3.2.3 NAME INCOME TYPE

Regroupement en 2 catégories :

- "Apartment/house"
- “Other

3.2.4 FLAG DOCUMENT

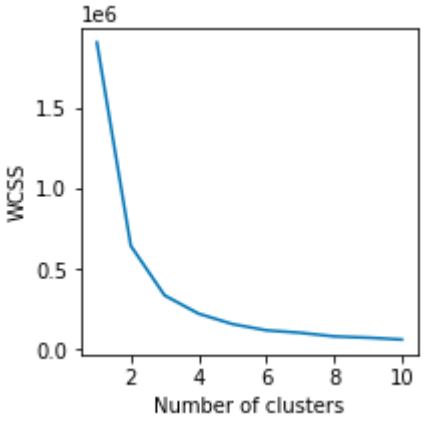
Ces données concernent les documents transmis à l'organisme pour l'étude du dossier. Après une rapide étude, on observe qu'il existe 3 grandes catégories de clients :



L'ensemble des colonnes est donc regroupé au sein d'une nouvelle entrée nommée "Cluster_doc" et reprend pour chaque client un code allant de 1 à 3.

3.2.5 CNT_SOCIAL_CIRCLE

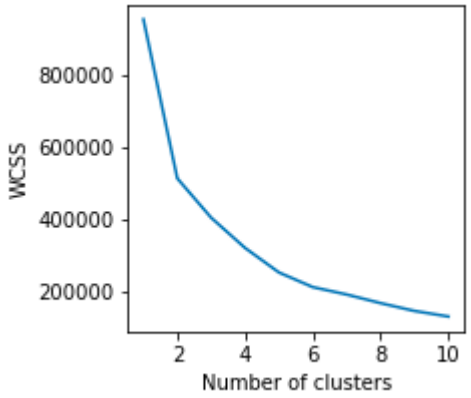
Nous avons procédé de même pour le défaut de paiement dans le cercle social du demandeur



L'ensemble des colonnes est donc regroupé au sein d'une nouvelle entrée nommée "Cluster_Social_C" et reprend pour chaque client un code allant de 1 à 3.

3.2.6 CREDIT BUREAU

Une dernière application concerne l'heure et le jour de la demande du client

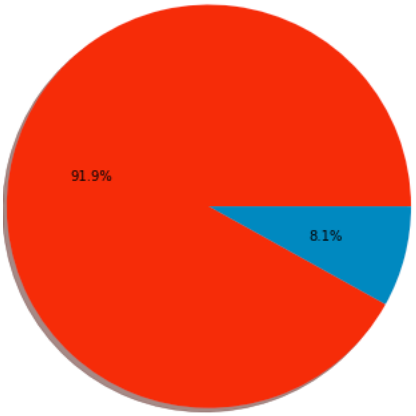


L'ensemble des colonnes est donc regroupé au sein d'une nouvelle entrée nommée "Cluster Time Bureau" et reprend pour chaque client un code allant de 1 à 3.

4 - Modélisation

4.1 - Resampler

Nous sommes ici en présence d'un souci, en effet le nombre de target est largement déséquilibré.



Nous allons donc appliquer une méthode de “ReSampling” à savoir un rééquilibrage des données via la librairie ImbLearn et la méthode Random Over Sampling. Divers outils sont disponibles et après quelques essais avec Smote() et Random Under Sampling. Cette méthode nous paraît la plus adaptée à notre travail.

La conséquence directe est malheureusement une augmentation du temps de calcul, car la base double quasiment de volume, mais les résultats s’en trouvent nettement améliorés.

4.2 - Preprocessing

Nous créons désormais un Pipeline afin de traiter une dernières fois nos données avant la création d’un modèle.

4.2.1 - Variables numériques

- SimpleImputer

  Tout d'abord, il s’agit de remplacer les valeurs manquantes, On applique ici un Simple Imputer avec la stratégie ‘mean’, les valeurs sont donc remplacées par la moyenne constatée dans la colonne.

- Quantile Transformer

  Pour faciliter l’apprentissage les données sont transformé par Quartile qui nous permet d’établir des intervalles des données

4.2.2 - Variables Catégorielles

- SimpleImputer

  De même on remplace les valeurs manquantes, On applique donc un Simple Imputer avec la stratégie ‘most_frequent’, les valeurs sont donc remplacées par la valeurs la plus constatée dans le dataset

- OneHotEncoder

Enfin les données catégorielles sont transformé en donnée numériques grâce au One Hot Encorder qui crée une matrice pour chaques colonnes présentes

4.2.3 - Reunion

- SelectKBest

  Avant d'envoyer l'ensemble des données au modèle, on applique un SelectKBest qui filtre nos données avec le nombre de variable souhaitée classement en fonction de leur importance pour le modèle

4.3 - Choix des métriques

4.3.1 Roc Score

On représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs (fraction des positifs qui sont effectivement détectés) en fonction du taux de faux positifs (fraction des négatifs qui sont incorrectement détectés).

4.3.2 Matrice confusion

4.3.3 F1

Le F1-Score combine la précision et le rappel.

4.3.4 Recall

Le rappel correspond au nombre de documents correctement attribués par rapport au nombre total appartenant réellement à la classe

4.3.5 Précision

La précision correspond au nombre de documents correctement attribués par rapport au nombre total de documents prédits

4.3.5 Accuracy

L'accuracy permet de connaître la proportion de bonnes prédictions par rapport à toutes les prédictions

4.3 - Entraînement

4.3.1 Logistic

Roc Score : 0.7357

4.3.2 Random Forest :

Roc Score : 0.7156

4.3.3 XGBoost

Roc Score : 0.7461

4.4 - Optimisation

4.4.1 Logistic avec hyperparamètre

```
{'model__C': 0.615848211066026, 'model__penalty': 'l1', 'model__solver': 'liblinear', 'selector__k': 69}
```

Roc Score : 0.7326

4.3..2 Random Forest avec hyperparamètre

```
{'model__bootstrap': False, 'model__max_depth': 100, 'model__max_features': 'sqrt', 'model__min_samples_leaf': 4, 'model__min_samples_split': 2,
'model__n_estimators': 230, 'selector__k': 54}
```

Roc Score : 0.7334

4.3.3 XGBoost avec hyperparamètre

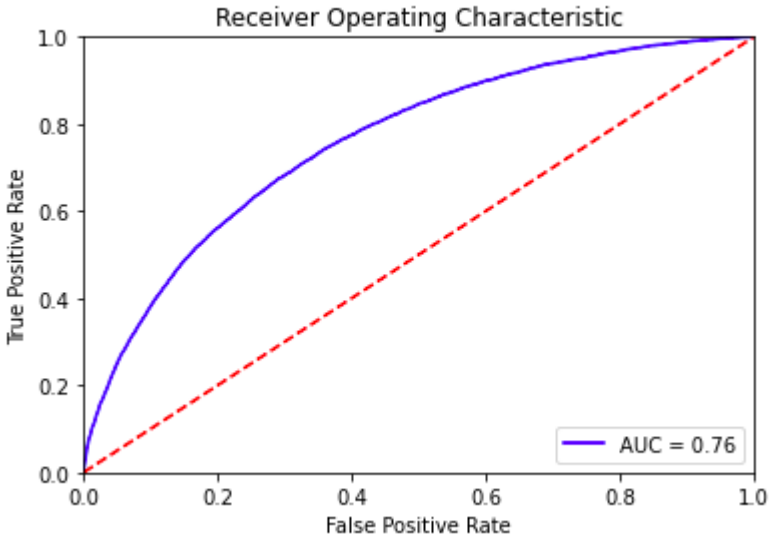
```
{'model__alpha': 0.43941871338822636, 'model__gamma': 0.4275061568251538, 'model__learning_rate': 0.14929882428726363, 'model__max_depth': 4, 'm
odel__min_child_weight': 4, 'model__subsample': 0.75, 'selector__k': 73}
```

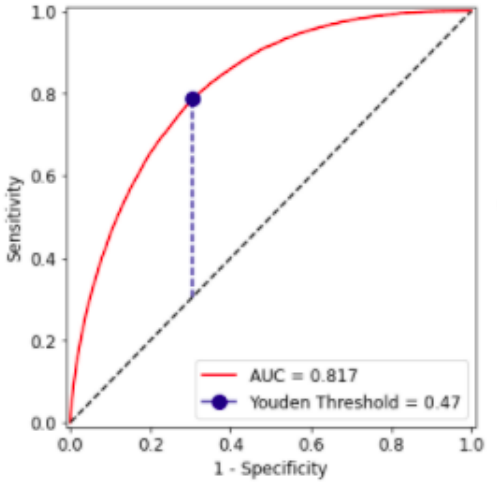
Roc Score : 0.7534

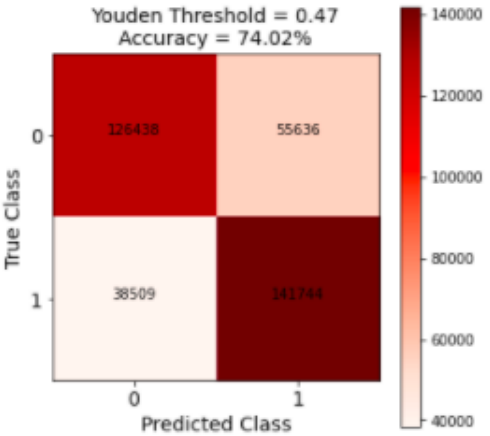
Meilleur Choix XGBOOST

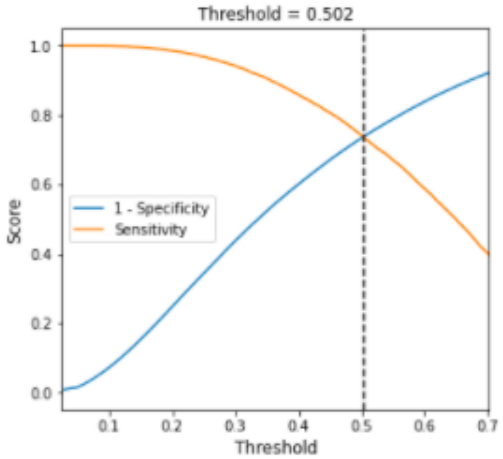
5 . Evaluation et seuil

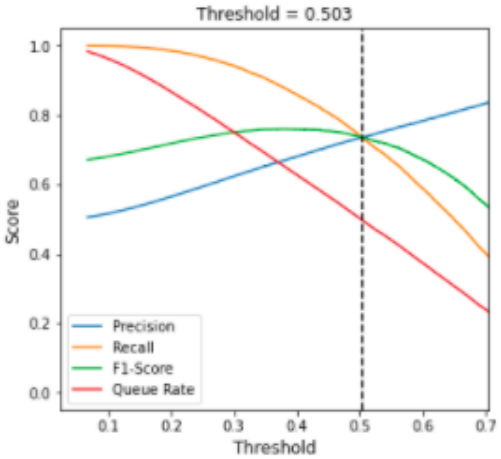
```
XGB classifier roc score  : 0.7601673340214006
```











Au vu du travail demandé, aucune métrique ne correspond parfaitement à nos attentes. Nous allons donc "créer" un scoring qui correspond mieux à nos attentes, le but étant donc de maximiser les taux de TN, mais surtout de pénaliser les FP , car ce sont les clients les plus risqués pour la banque. Nous ne pouvons nous permettre d'accepter un crédit à un client avec un risque élevé.

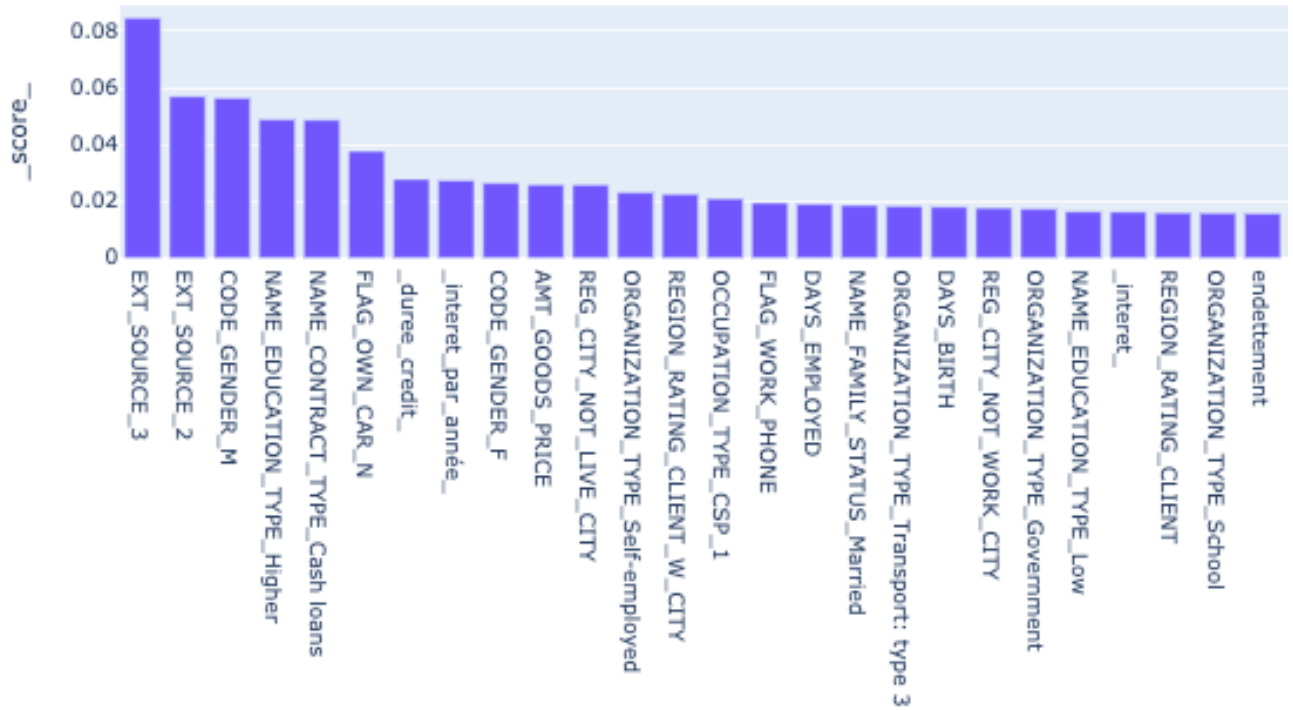
Voici les données et seuil choisi, de manière arbitraire, et pouvant évidemment donner lieux à un échange.

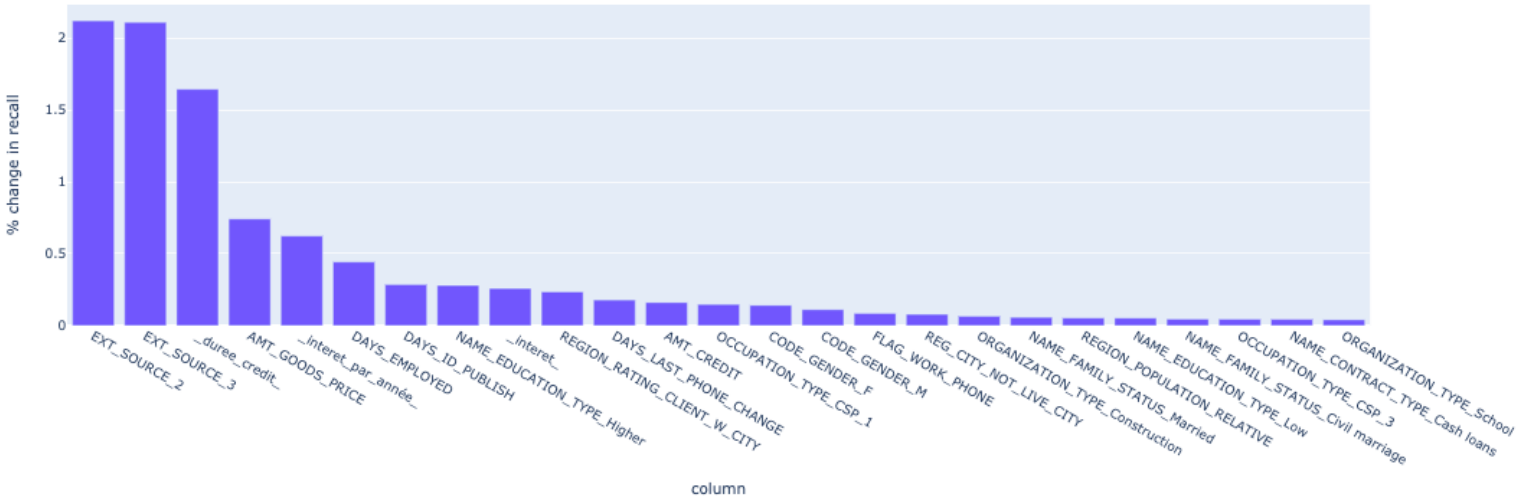
Pénalisation des FN = -25

Pénalisation des FP = -100

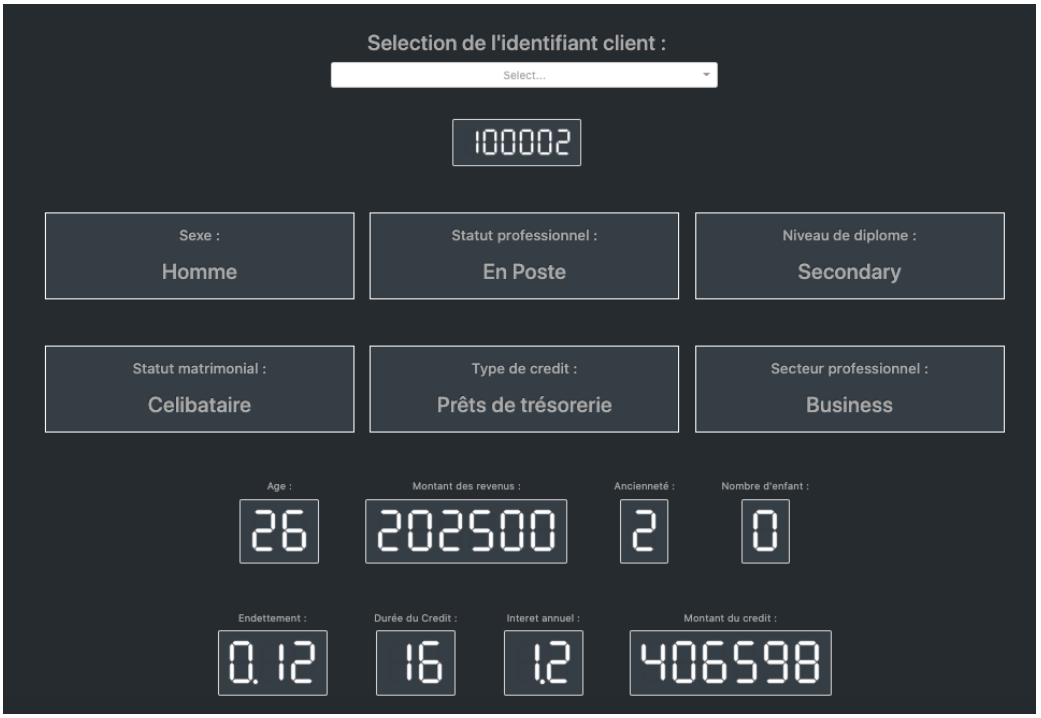
	seuil	base_0	all_0	true_0	false_0	base_1	all_1	true_1	false_1	my_metrics
699	0.799	36389	64332	35369	28963	36077	8134	7114	1020	-2038621
698	0.798	36389	64236	35357	28879	36077	8230	7198	1032	-2030833
697	0.797	36389	64155	35339	28816	36077	8311	7261	1050	-2025514
696	0.796	36389	64083	35322	28761	36077	8383	7316	1067	-2020945
695	0.795	36389	64000	35313	28687	36077	8466	7390	1076	-2013993
...	...	...	...	...	...	...	...	...	...	...
277	0.377	36389	24691	20482	4209	36077	47775	31868	15907	-401913
256	0.356	36389	22814	19290	3524	36077	49652	32553	17099	-401864
250	0.350	36389	22244	18929	3315	36077	50222	32762	17460	-401693
253	0.353	36389	22515	19109	3406	36077	49951	32671	17280	-400444
252	0.352	36389	22420	19051	3369	36077	50046	32708	17338	-400071

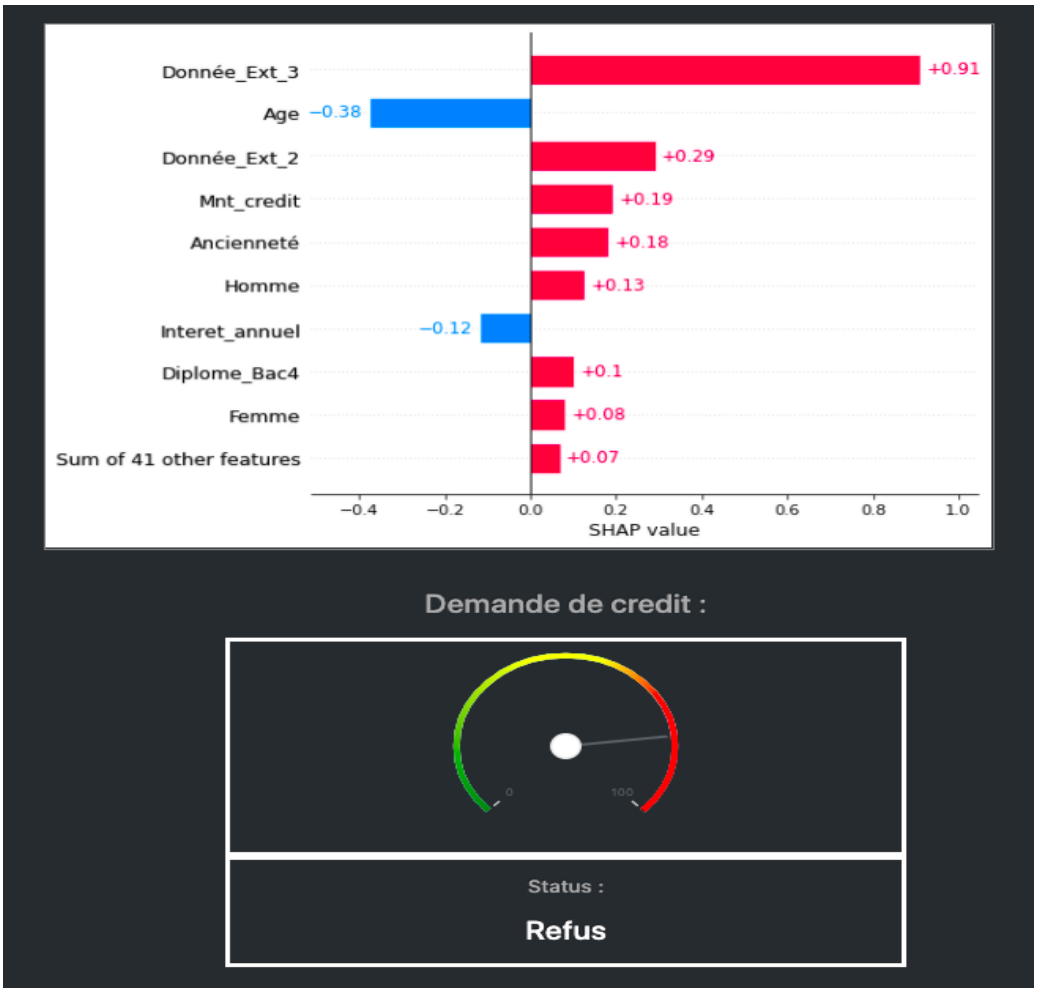
6 . Features Importances





7 . DashBoard


Selection de l'identifiant client :
Select...
100002
Sexe :
Homme
Statut professionnel :
En Poste
Niveau de diplome :
Secondary
Statut matrimonial :
Celibataire
Type de credit :
Prêts de trésorerie
Secteur professionnel :
Business
Age :
26
Montant des revenus :
202500
Ancienneté :
2
Nombre d'enfant :
0
Endettement :
0.12
Durée du Credit :
16
Interet annuel :
12
Montant du credit :
406598


Donnée_Ext_3
+0.91
Age
−0.38
Donnée_Ext_2
+0.29
Mnt_credit
+0.19
Ancienneté
+0.18
Homme
+0.13
Interet_annuel
−0.12
Diplome_Bac4
+0.1
Femme
+0.08
Sum of 41 other features
+0.07
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
SHAP value
Demande de credit :
0
100
Status :
Refus

6.1 - Choix de l’ID client

- liste déroulante des Id utilisateur
- Affichage LED de l’id sélectionné

6.2 - Information client général

- Sexe
- Statut professionnel
- Niveau de diplôme
- Statut matrimonial
- Type de crédit
- Secteur professionnel

6.3 - Information client sur la situation financière

- Âge
- Montant des revenus
- Ancienneté
- Nombre d'enfant
- Endettement
- Durée du Crédit
- Intérêt annuel
- Montant du crédit

6.4 - Explication du modèle

Mise en avant des 10 variables les plus importantes ayant contribué au score.

En bleu les variables en faveur de l’accord

En rouge les variables ayant contribuées à une réponse défavorable

6.5 - Résultat du modèle

- Jauge du score client
- Avis rendu

7 . Conclusion

Voici donc un résumé des différentes méthodes qui nous ont permis de mener à bien ce travail très intéressant :

- Random Over Sampling
- SimpleImputer
- Quantile Transformer
- OneHotEncoder
- SelectKBest
- Logistic Regression
- Random Forest
- XGBoost
- Shap