P7

Implémentation d'un modèle de scoring

Etudiante : Fatma AIDI
Mentor : Calliane YOU
Evaluateur : Ibrahima Diakite
Date :17/08/2021

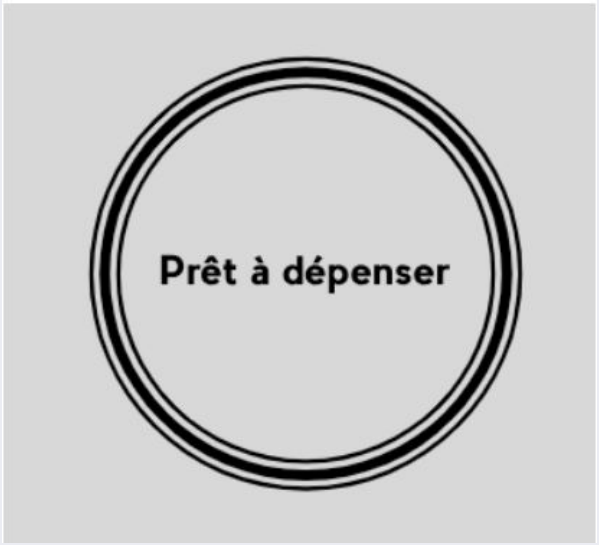
TABLE

01
MISSION

Mission

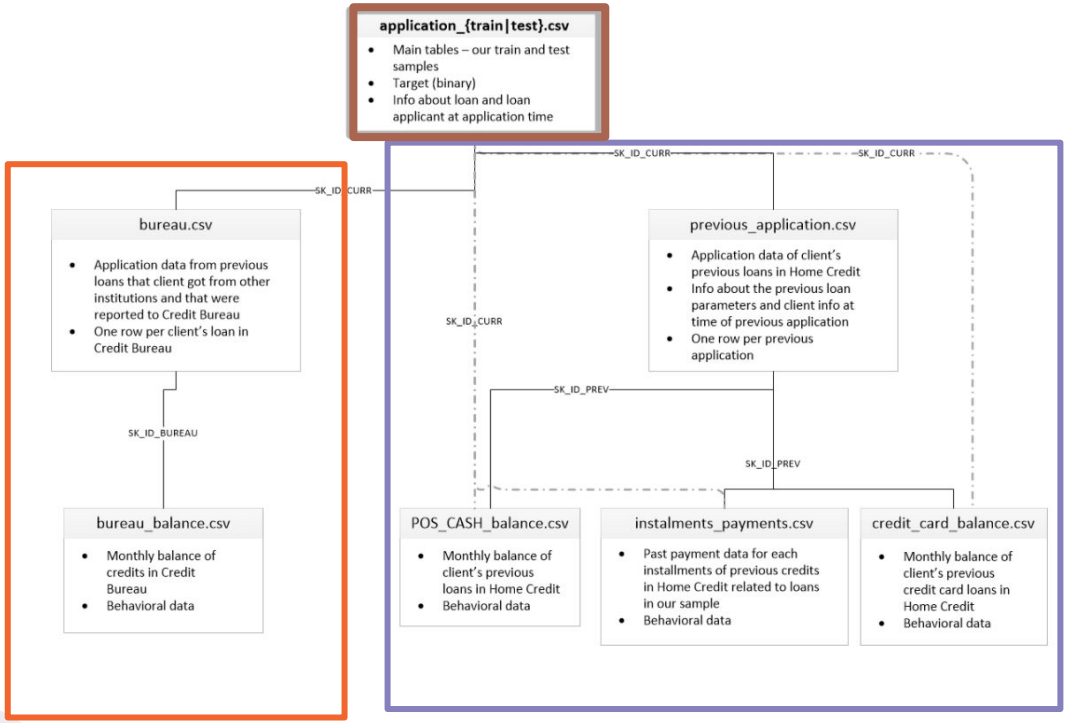
- Développer un **modèle de scoring** afin d'établir la probabilité de défaut de paiement d'un client.

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



Les informations générales sur le client (âge, revenu, nombre d'enfants..) et les données concernant le crédit(prix du bien, annuité..).

Informations sur les prêts antérieurs de chaque client auprès de toutes les institutions financières autres que Home Credit



Informations sur les anciens crédits et les données mensuelles de la carte de crédit de chaque client dans la même agence.

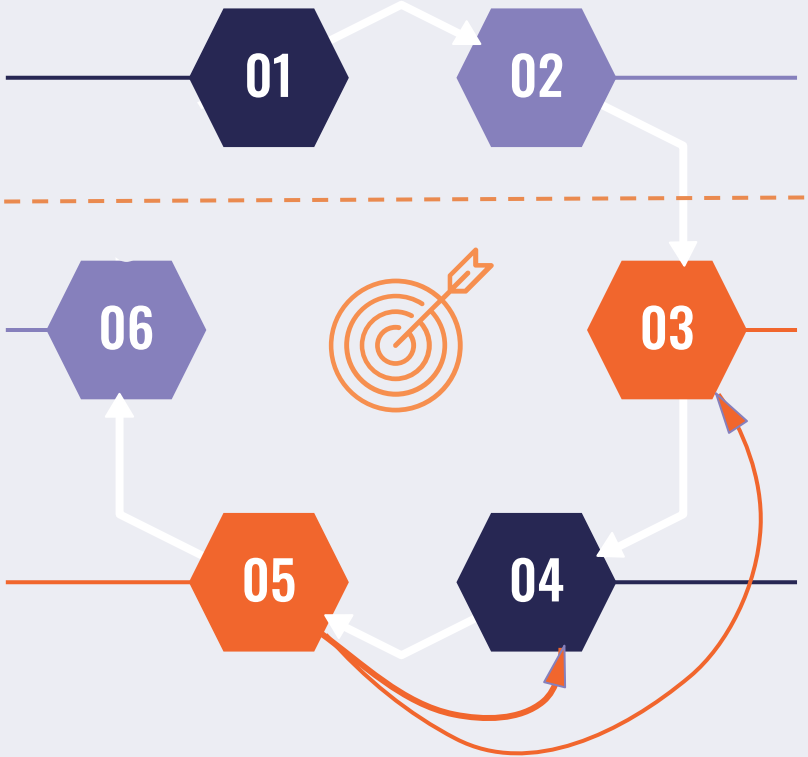
DÉMARCHE /MÉTHODOLOGIQUE

TABLE UNIQUE

Jointure et **agrégation** des tables par client unique

PREPARATION DES DONNEES

Features engineering, **nettoyage**, analyse exploratoire



INTERPRETATION

Interprétation des résultats (avec SHAP) et **sauvegarde** du modèle

PRE-PROCESSING

Encodage ,imputation, standardisation et mise à l' échelle, **features selection**

MODELISATION

Entraînement du modèle avec les paramètres optimaux, **validation**, **choix du modèle** final.

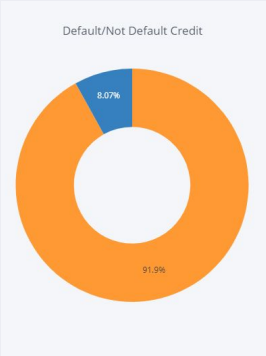
OPTIMISATION

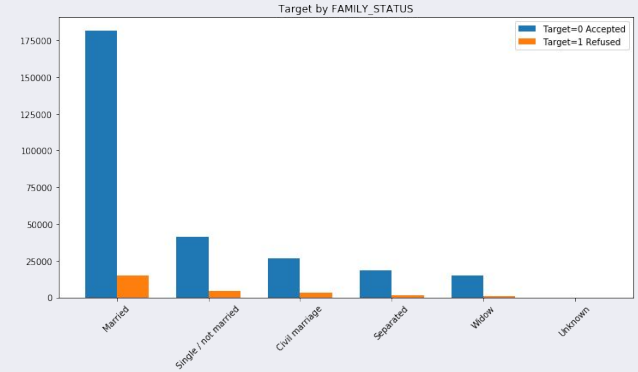
Split(3 sets), **downsampling**, choix de la **métrique d' évaluation**, Optimisation des Hyper-Paramètres

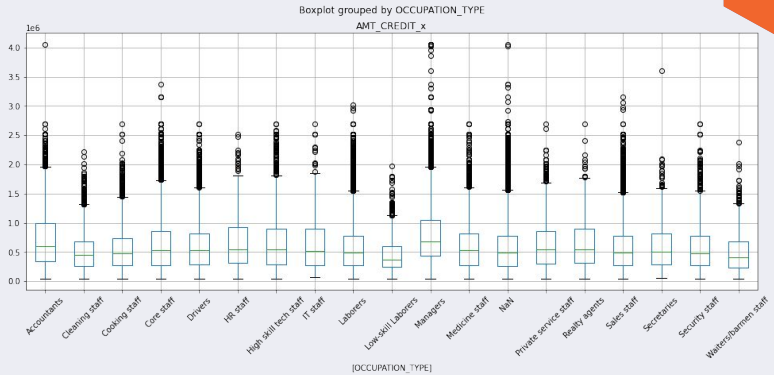
ANALYSE EXPLORATOIRE

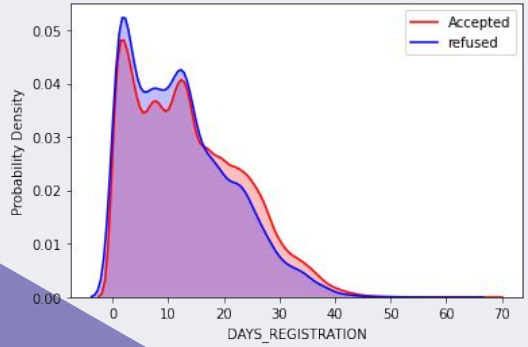
Data finale:

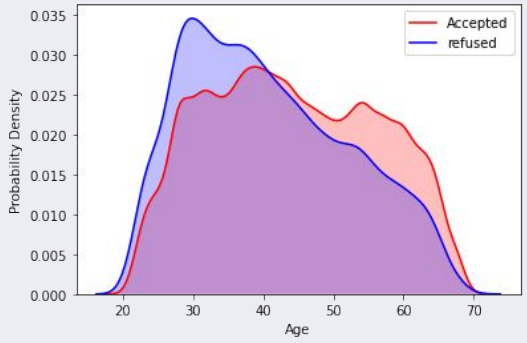
- **307511** observations (dossier client)
- **101** variables
- 92% sont des clients fiables, contre 8% sont à défaut de paiement.

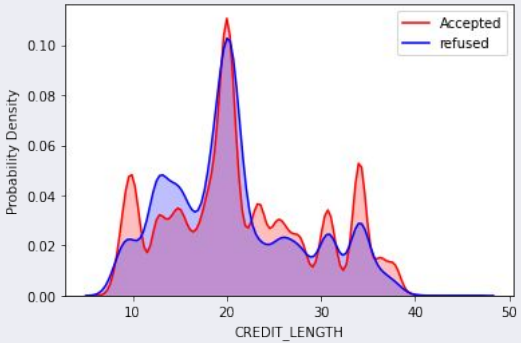












02

Modèle de scoring

PRE-PROCESSING

Encodage des données catégoriques

méthode utilisée:
TargetEncoder

Imputation de données manquantes

méthode utilisée:
IterativeImputer

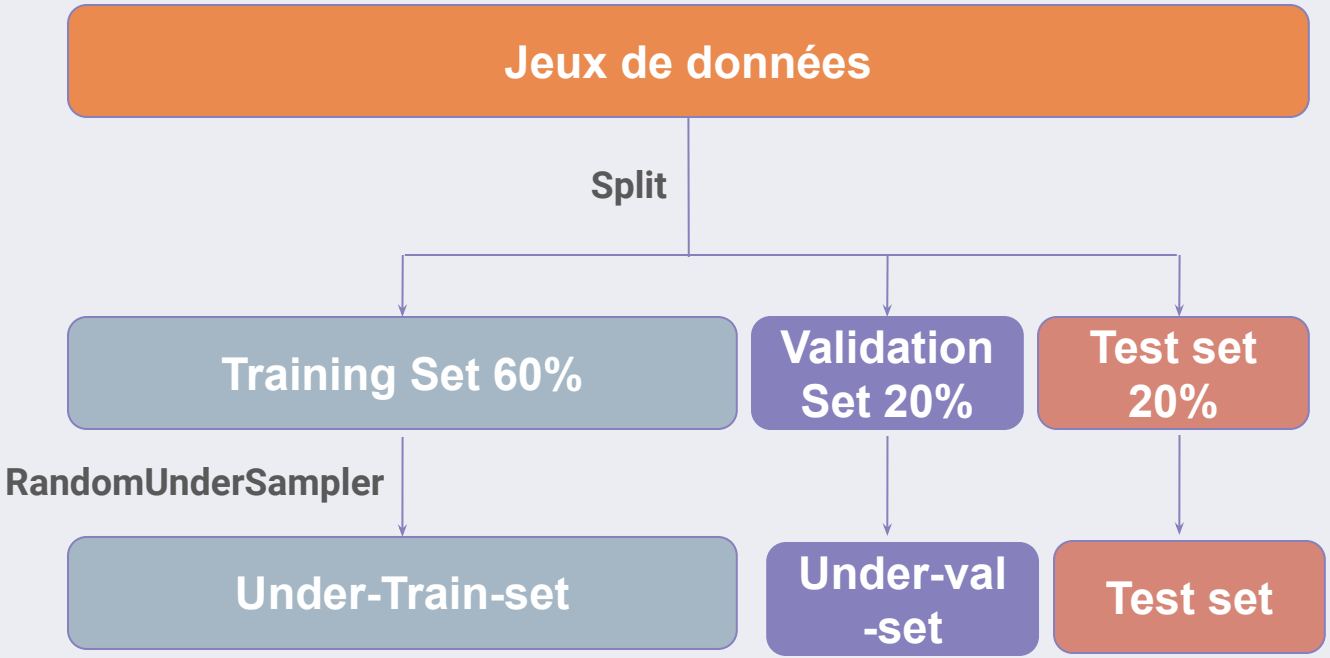
Mise à l' échelle

méthode utilisée:
StandardScaler

Features selection (40 variables)

méthode utilisée:
SelectKBest

Equilibrage des classes: Downsampling



Modélisation

***Modèle de classification supervisé:**

Le choix des modèles est:

- La régression logistique: **Logistic Regression**
- Forêts aléatoires: **Random Forest**
- Le Gradient Boosting: **XGBClassifier et Light GBM classifier**

*** Optimisation des Hyper-Paramètres:**

Pour le réglage des hyperparamètres afin de déterminer les valeurs optimales pour un modèle donné on a choisi une recherche par grille **GridSearch** par validation croisée(cross-validation) **k-fold.**

Métrique d'évaluation

Matrice de confusion	Classe 0 Prédiction:Non Défaillant	Classe 1 Prédiction: Défaillant
Classe 0 Réel:Non Défaillant	**TN** True Negative	**FN** False Negative
Classe 1 Réel: Défaillant	**FP** False Positive	**TP** True Positive

- **Recall** ou Sensibilité :pourcentage des vrais positifs (Défaillants)

$$Recall = TP/(TP+FN)$$

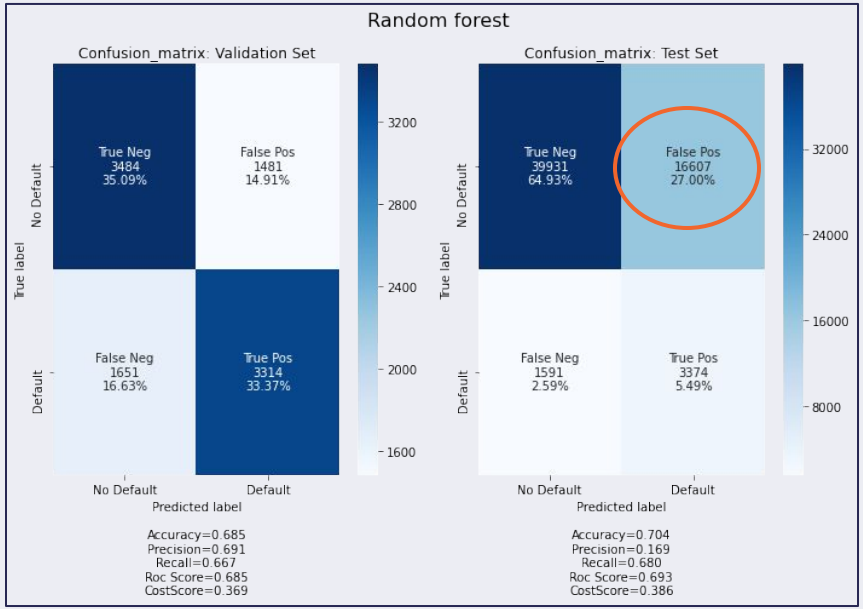
- **Spécificité** = pourcentage des vrais négatifs (Non-Défaillants)

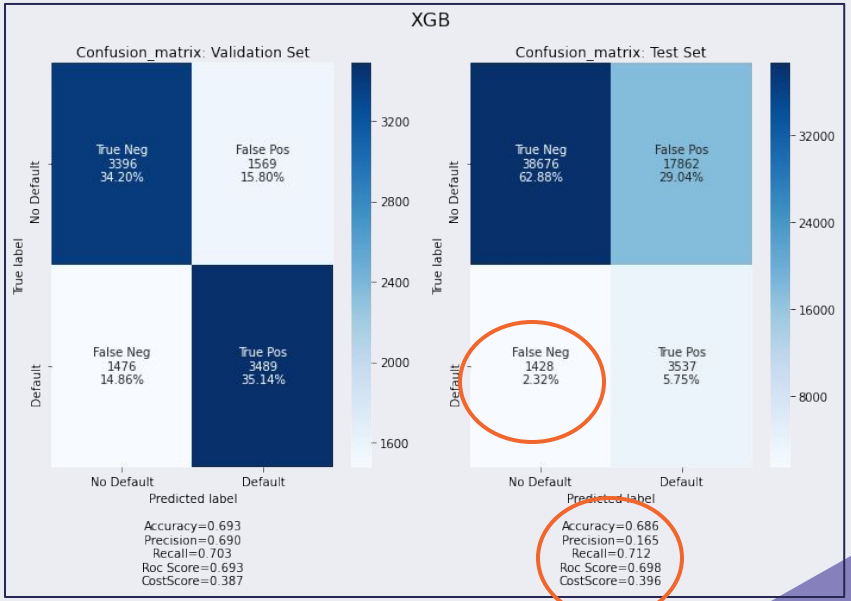
$$Spécificité = TN/(TN+FP)$$

1ère stratégie: Youden's J statistic

Maximiser le nombre des TP et TN → **Maximiser** Recall et Spécificité

J_index=Recall+Spécificité-1





2ème stratégie: Cost-Sensitive Learning

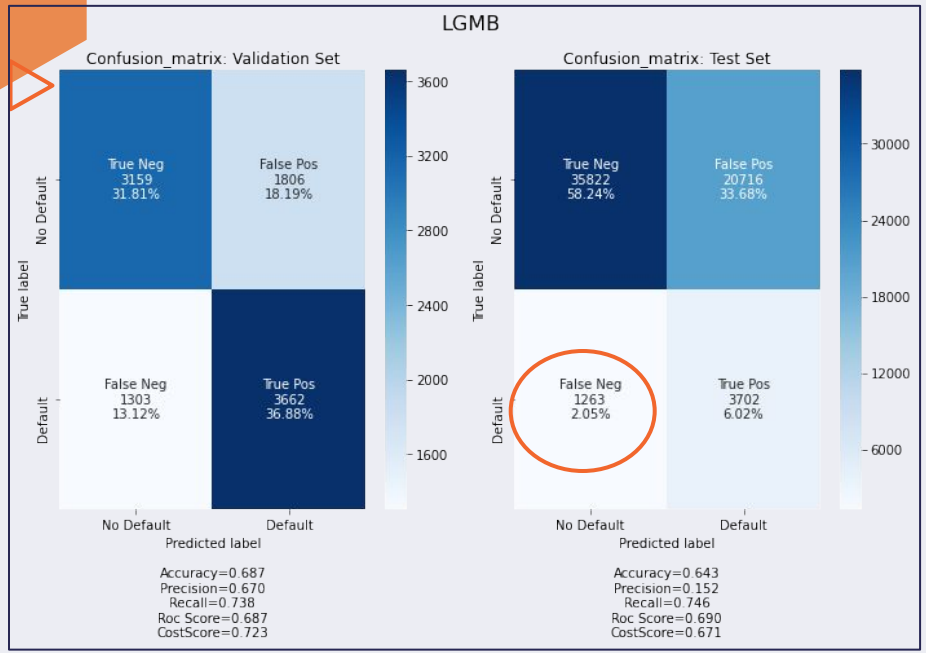
L'ensemble de données est déséquilibré, nous utiliserons donc un apprentissage sensible aux coûts et pondéré par classe.

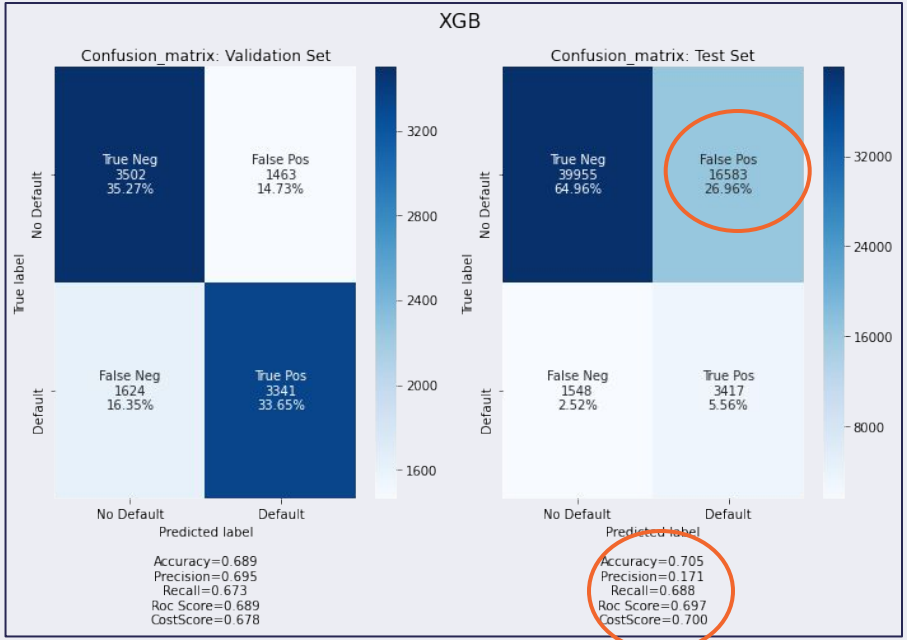
$$Fonction_{coût} = TN \times C_{tn} + FN \times C_{fn} + FP \times C_{fp} + TP \times C_{tp}$$

Matrice de coût	Classe 0 Prédiction:Non Défaillant	Classe 1 Prédiction: Défaillant
Classe 0 Réel:Non Défaillant	**Ctn** Valeurs des intérêts annuels	**Cfn** - Valeurs des intérêts annuels
Classe 1 Réel: Défaillant	**Cfp** - 30% valeur des crédits	**Ctp** -frais du dossier

- **Ctn** = valeur moyenne annuelle d'intérêts payée par les comptes sans défaut(Target=0).
- **Cfn**: valeur moyenne du montant du crédit demandé par les comptes à défaut de paiement et que l'assurance ne couvre pas (hypothèse : 30% du montant et à vérifier avec la banque).
- **Ctp** : la banque ne gagne pas de l'argent mais elle perd les frais administratifs pour l'étude du dossier du client et qui est un service gratuit pour le client: valeur forfaitaire à définir avec la banque(300).

2ème stratégie: Cost-Sensitive Learning

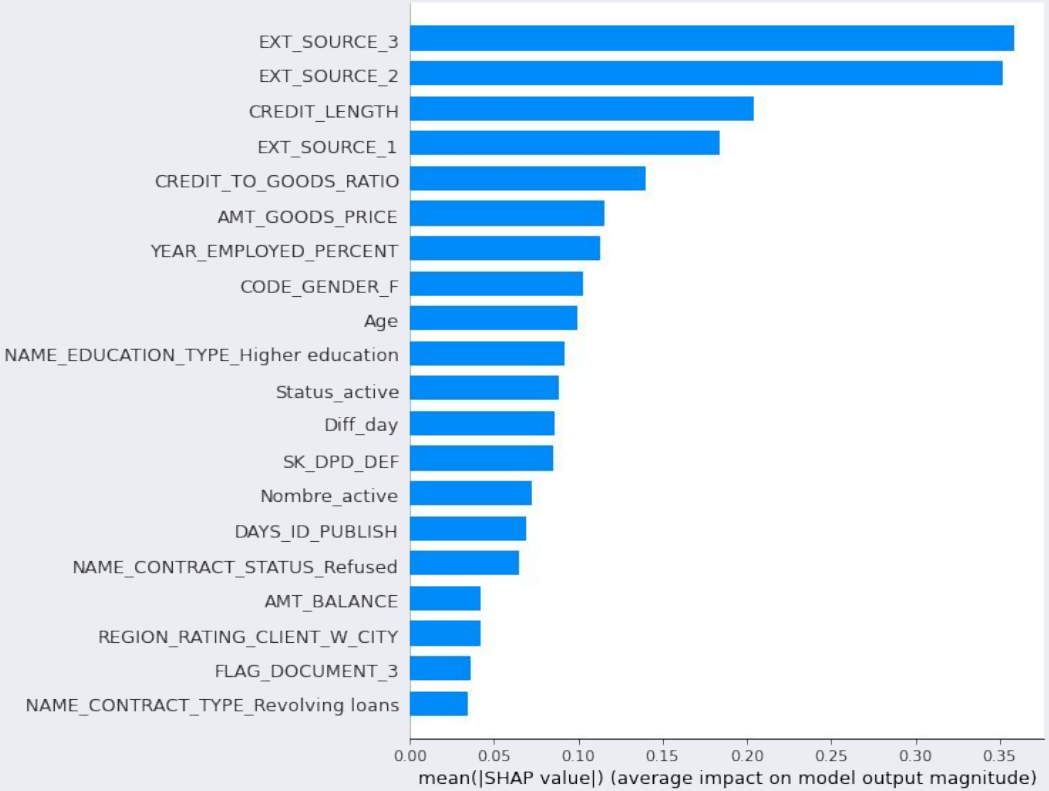




03

Interprétabilité du modèle

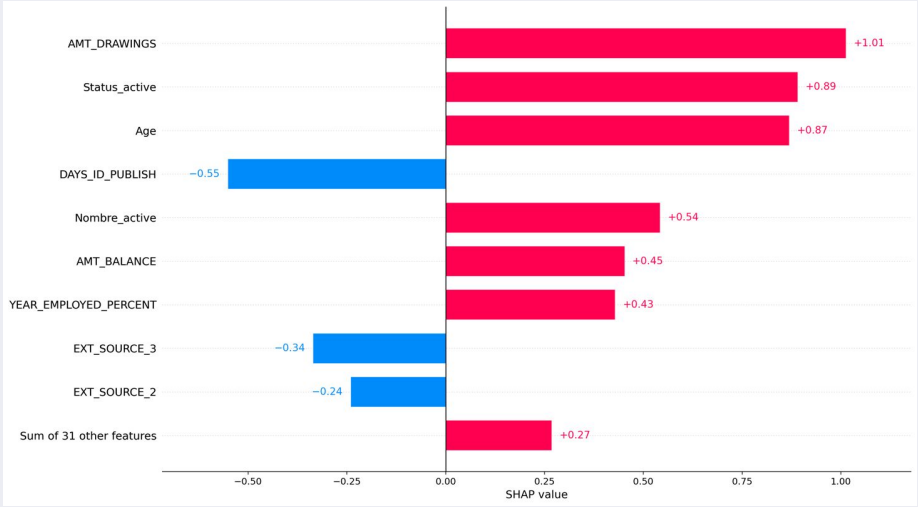
L’importance des variables



L'impacte des variables sur une prédiction

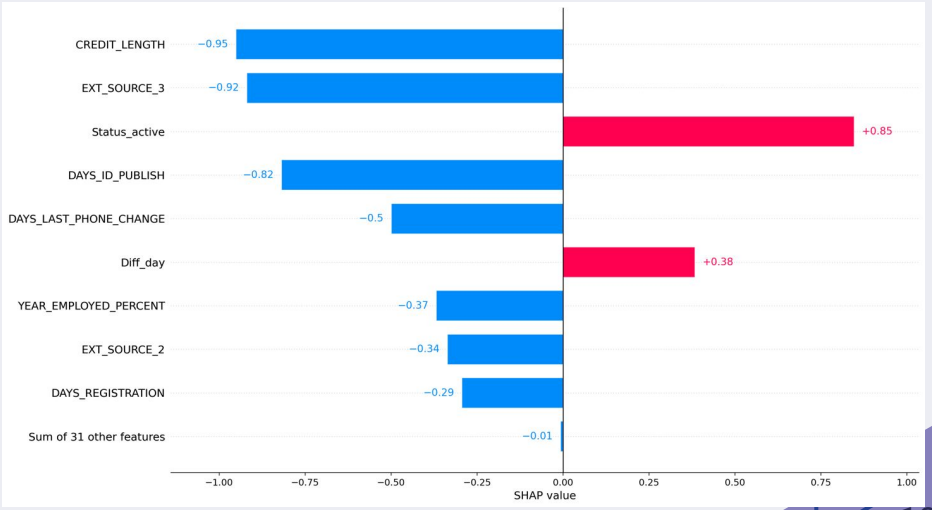
Refusé

Dossier: 227742
Risque: 96%
Montant de prélèvement/mois :4474 (elevé)



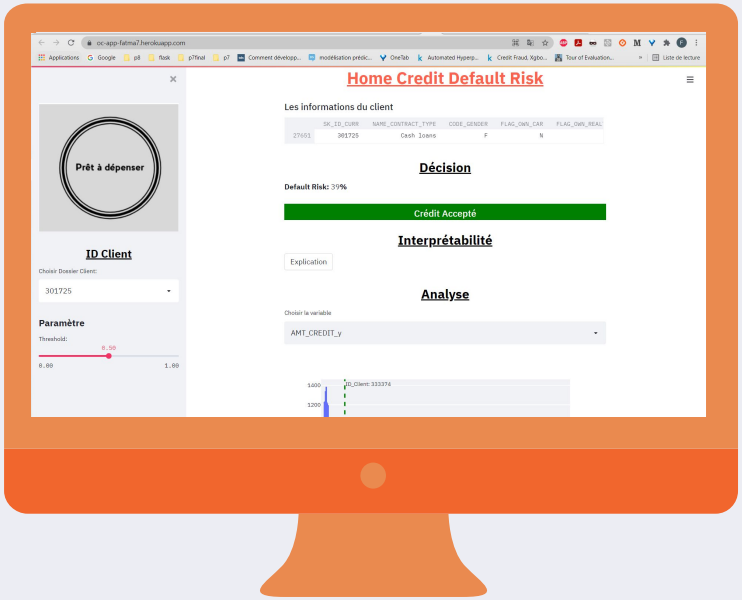
Accepté

Dossier: 262246
Risque: 4%
Durée du crédit:10 ans (court)

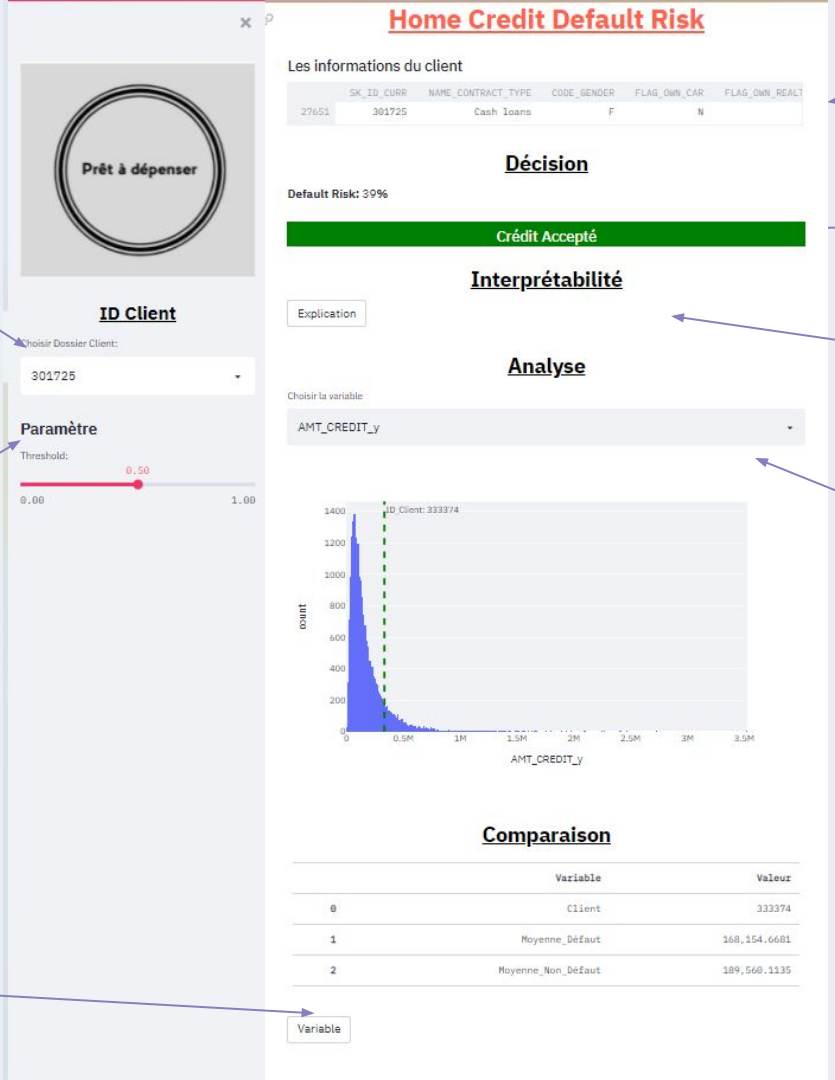


04

Dashboard API



Pour avoir une interface interactive que les chargés client peuvent utiliser facilement , on a développé une application avec **Streamlit** et on l'a déployé sur **Heroku** (disponible ici : https://oc-app-fatma7.herokuapp.com/)



Numéro du dossier

Seuil toléré par la banque

Définition de chaque variable

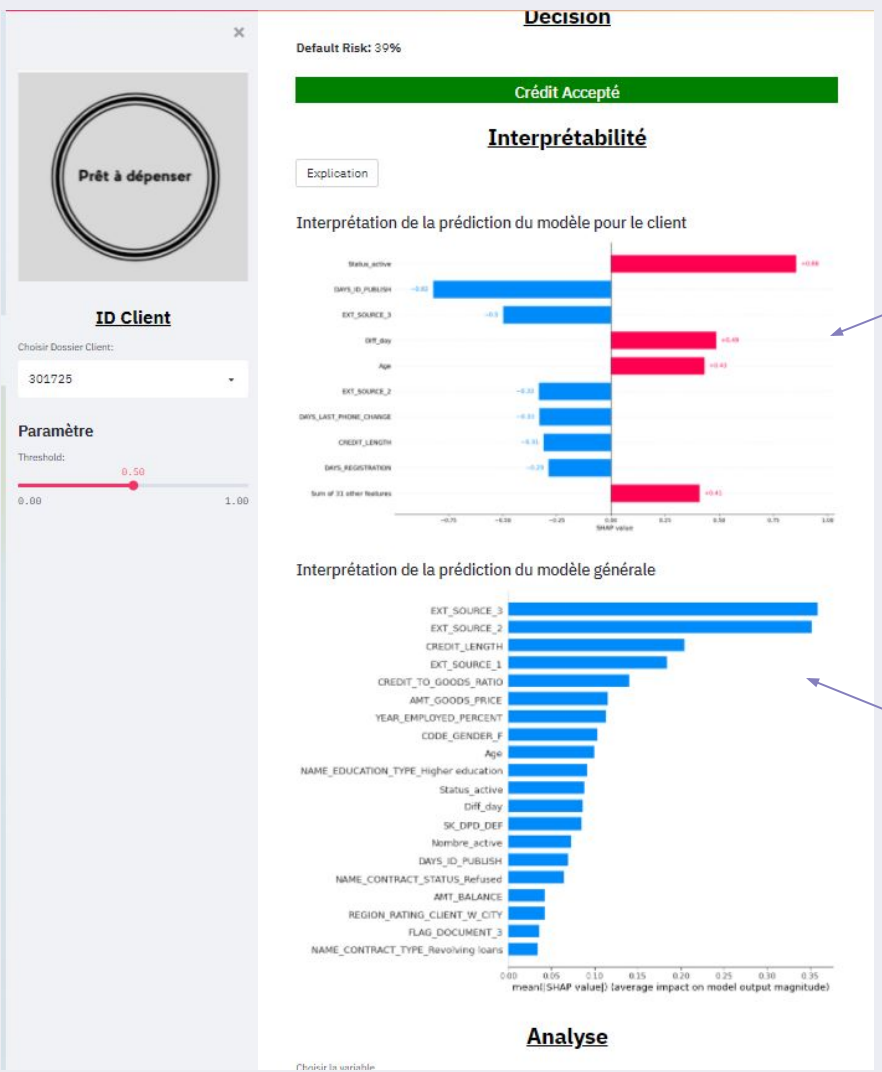
Données de chaque dossier choisi

Décision et probabilité du risque

Interprétabilité de la décision

Distribution de chaque variable et la position de la valeur du client.

Comparaison de la valeur moyenne de la variable pour les dossiers à défaut de paiement ou non avec la valeur du client.



Montrer les variables qui ont le plus d'impact positif ou négatif sur la prédiction du modèle sur ce dossier

Montrer les variables Les plus importante sur la prédiction du modèle

Axe d'amélioration

Modélisation :

- Un modèle plus performant
- Essayer d'autres modèle de classification
- Adapter plus la métrique d'évaluation et la fonction coût vers les besoins de métier
- Plus de Features engineering adaptés au métier
- Optimisation et paramétrage avec une optimisation automatique(TPOTClassifier).

Dashboard

- Plus de Graphes interactifs orientés métier, à regarder avec les chargés de client
- Plus de fonctionnalité

Merci

Fatma AIDI

Github : https://github.com/AIDIF84/DS_project


slidesgo