Implémentez un modèle de scoring

17.08.2021

—

Fatma AIDI
OpenClassroom

Vue d'ensemble et Objectifs

La société financière "Prêt à dépenser" propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêts. Elle souhaite **développer un modèle de scoring afin d'établir la probabilité de défaut de paiement d'un client**.

L'entreprise 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 en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

Elle décide donc de **développer 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, mais également permettre à leurs clients de disposer de leurs informations personnelles et de les explorer facilement.

Grandes étapes

I. Méthodologie d'entraînement

II. Optimisation et métrique d'évaluation

III. Interprétabilité du modèle

IV. Dashboard et API

V. Axe d'amélioration

.

1. Méthodologie d'entraînement

• Préparation des données et Analyse

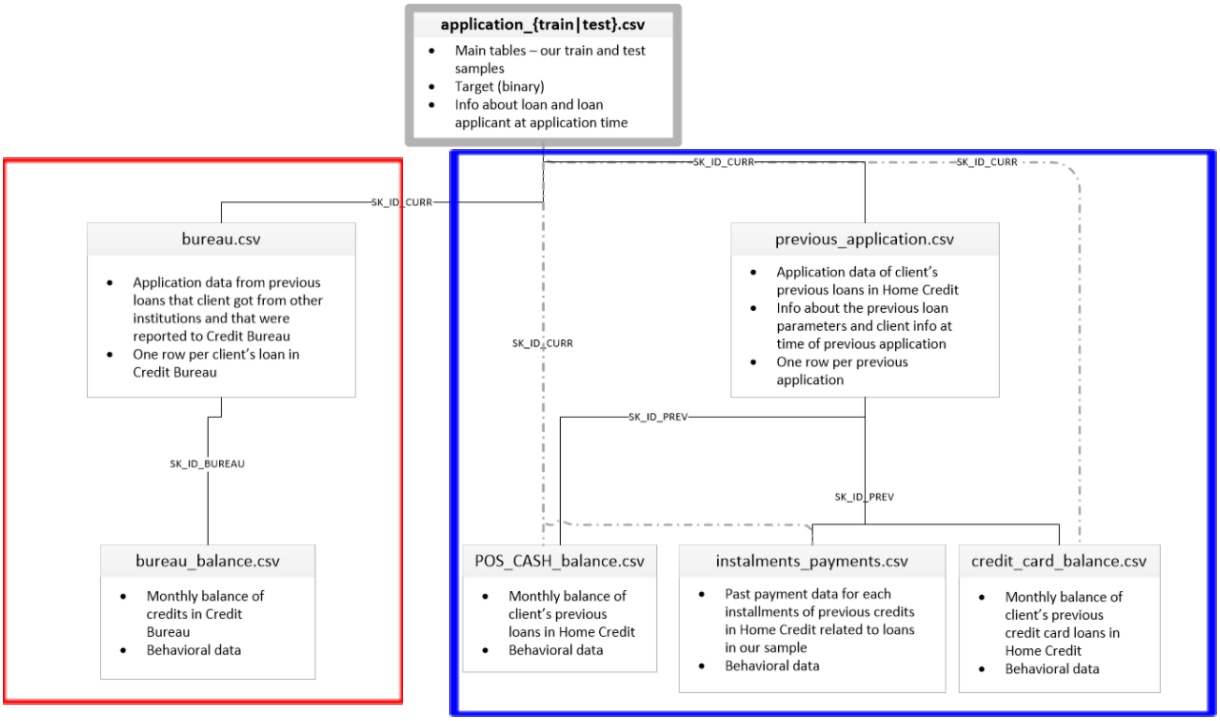


Figure 1: Diagramme des Tables

Les jeux de donnée sont divisé en 3 groupes (figure 1):

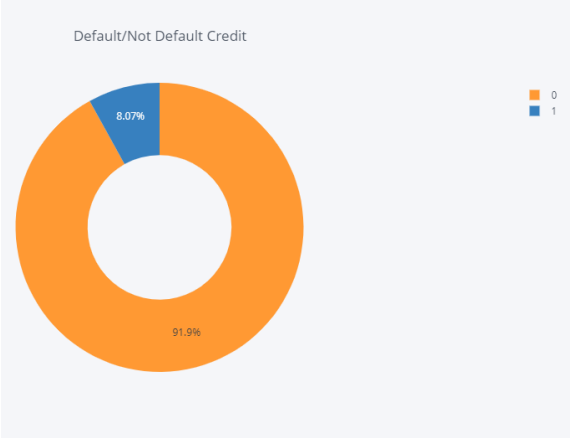
- groupe 1: des informations générales sur le client (âge, revenu, nombre d'enfants..) et les données concernant le crédit(prix du bien, annuité..).
- groupe 2(bleu): des données concernent des prêts antérieurs de chaque client auprès Home Credit, ainsi que des données mensuelles de ses cartes crédit.
- groupe 3(rouge): des données concernent des prêts antérieurs de chaque client auprès de toutes les institutions financières autres que Home Credit.

Dans cette partie on a rassemblé toutes ces jeux dans une seule table une table en passant par:

- **Jointure** et **agrégation** des tables par client unique
- Création des variables (Features engineering): Âge, durée du travail, nb de crédit en cours..

- Analyse exploratoire pour mieux orienter la modélisation
- Nettoyage et supprimer les variables dont le taux de valeurs manquante >30%.

Nous obtenons finalement un tableau de 307511 observations (dossier client) et 101 variables avec 92% de client n'ont pas un défaut de remboursement de leur crédit.

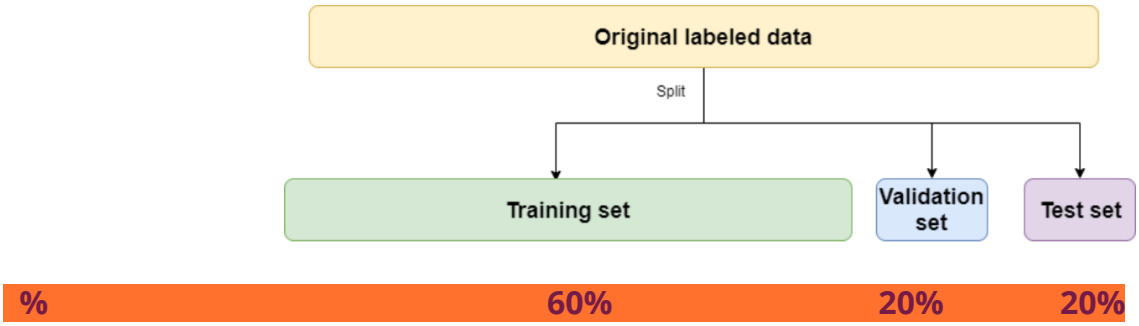


Avant de passer à la modélisation, il faut préparer les données qui seront entraînées par un modèle de classification. Les étapes de pré-processing sont:

- Encodage des données catégoriques (méthode utilisée: TargetEncoder).
- Imputation de données manquantes (méthode utilisée: IterativeImputer).
- Mise à l'échelle et normalisation des données.
- Sélection 40 variables les plus pertinentes (méthode utilisée: SelectKBest) pour simplifier la modélisation.

• Modélisation

Pour l'apprentissage supervisé d'un modèle de classification binaire, on devrait séparer 3 sous-ensembles de ces données étiquetées d'origine : les ensembles d'apprentissage (60%), de validation (20%) et de test (20%). Il s'agit d'une étape importante pour évaluer les performances de différents modèles et l'effet du réglage des hyperparamètres et pour éviter l'**overfitting**.



*Équilibrage des classes :Downsampling

Puisque l'ensemble de données est déséquilibré (91% des étiquettes est 0) nous avons procédé à l'équilibrage des classes par sous-échantillonnage (downsampling) de la classe majoritaire pour train-set et validation-set en utilisant **RandomUnderSampler**.

*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**

* Métrique d'évaluation:

Pour évaluer le meilleur modèle on a choisi deux métriques d'évaluation avec deux stratégies d'apprentissage différentes sensibles aux coûts et pondérées par classe.

* 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.**

* Entraînement du modèle choisi sur val_set et test_set

Pour choisir le meilleur modèle à retenir pour notre projet, on doit comparer les résultats de de classification sur les jeux de validation et de test.

2. Optimisation et métrique d'évaluation

Pour évaluer le risque de non paiement de crédit par client, on doit déterminer la probabilité liée à ce risque et le seuil toléré par la banque à partir duquel le dossier est classé défaillant(donc la demande du crédit est rejetée). Si la probabilité est inférieure au seuil, on considère que le crédit sera non-risqué, le dossier est classé négatif (0). Inversement, si la probabilité est supérieure au seuil, on considère que le crédit est risqué, le dossier est classé positif (1).

Il s'agit donc d'un problème de classification dont le résultat binaire dépend du paramètre: **seuil de risque** (fixé par la banque)**.**

Pour comparer les résultats des modèles, on établit la matrice de confusion qui compare les valeurs prédites avec les valeurs réelles, en calculant le nombre des valeurs correctement prédites et les fausses prédictions.

Pour notre problématique :

- Les défaillants forment la classe positive(1): classe minoritaires(8%),
- Les non-défaillants forment la classe négative(0): classe majoritaire(92%).

	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

Pour minimiser les pertes d'argent, nous devons minimiser le nombre de faux négatifs (prédit non-défaillant mais client défaillant: perte des intérêts de la somme prêtée) et faux positifs (prédit défaillant mais client non-défaillant: perte d'un pourcentage de la somme prêtée).

Pour cela on détermine quelques indicateurs qui vont nous aider à résoudre ce problème et orienter nos modèles vers un meilleur résultat de classification :

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

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

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

$$Spécificité = TN/(TN + FP)$$

- ## 1 ère stratégie: Youden's J statistic

Dans la première stratégie on suppose que TP et TN ont la même importance donc le même poids, donc **Youden's index** est la meilleure métrique pour évaluer le modèle. Cette métrique consiste à minimiser les FN et FP donc minimiser la perte de la banque .

l'équation de Youden's index est(https://en.wikipedia.org/wiki/Youden's_J_statistic) :

$$J_{index} = Recall + Spécificité - 1$$

Résultat: La classifieur répondant le mieux à l'ensemble des critères est **XGBClassifier** avec le score **0.39** sur les deux set validation et test.

- ## 2ème stratégie: Cost-Sensitive Learning

Puisque l'ensemble de données est déséquilibré, nous utiliserons un apprentissage sensible aux coûts et pondéré par classe. Dans l'apprentissage sensible aux coûts, une fonction de coût pondérée est utilisée.

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

Les valeurs de la matrice des coûts sont déterminées par l'argent que la banque perdra ou gagnera pour chaque mauvaise ou bonne valeur cible prédictive.(https://towardsdatascience.com/model-performance-cost-functions-for-classification-models-a7b1b00ba60)

	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).

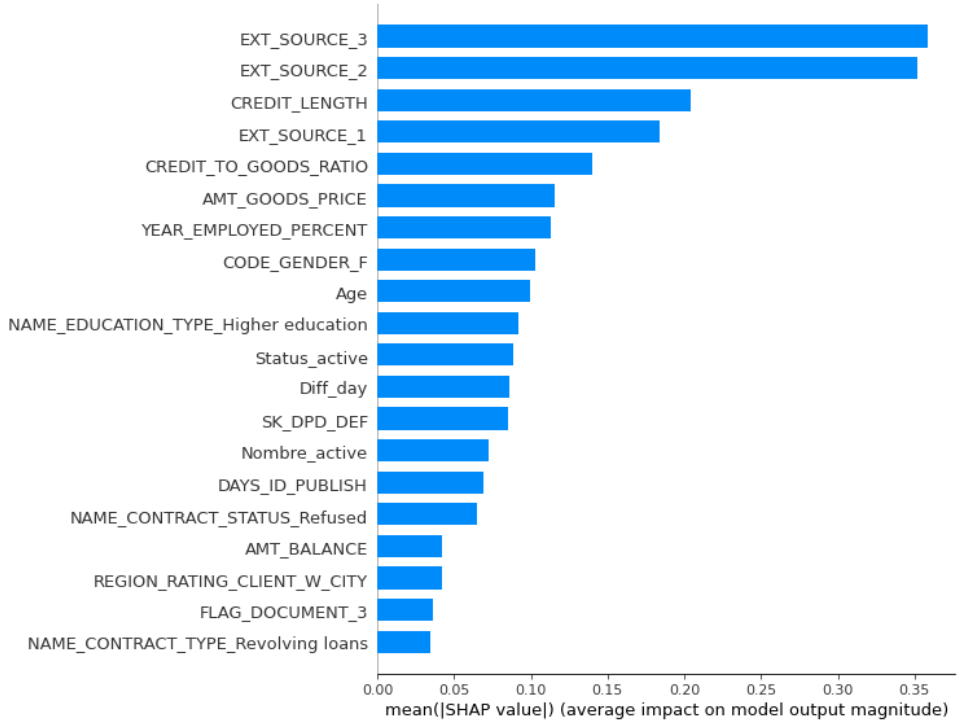
- **Ctp** : la banque ne gagne pas de l'argent mais elle perd les frais administratifs pour l'étude du dossier du client et qui est un service gratuit pour le client: valeur forfaitaire à définir avec la banque(300).

Résultats: La classifieur répondant le mieux à l'ensemble des critères est XGBClassifier avec le score 0.7 sur les deux set validation et test.

3. Interprétabilité du modèle

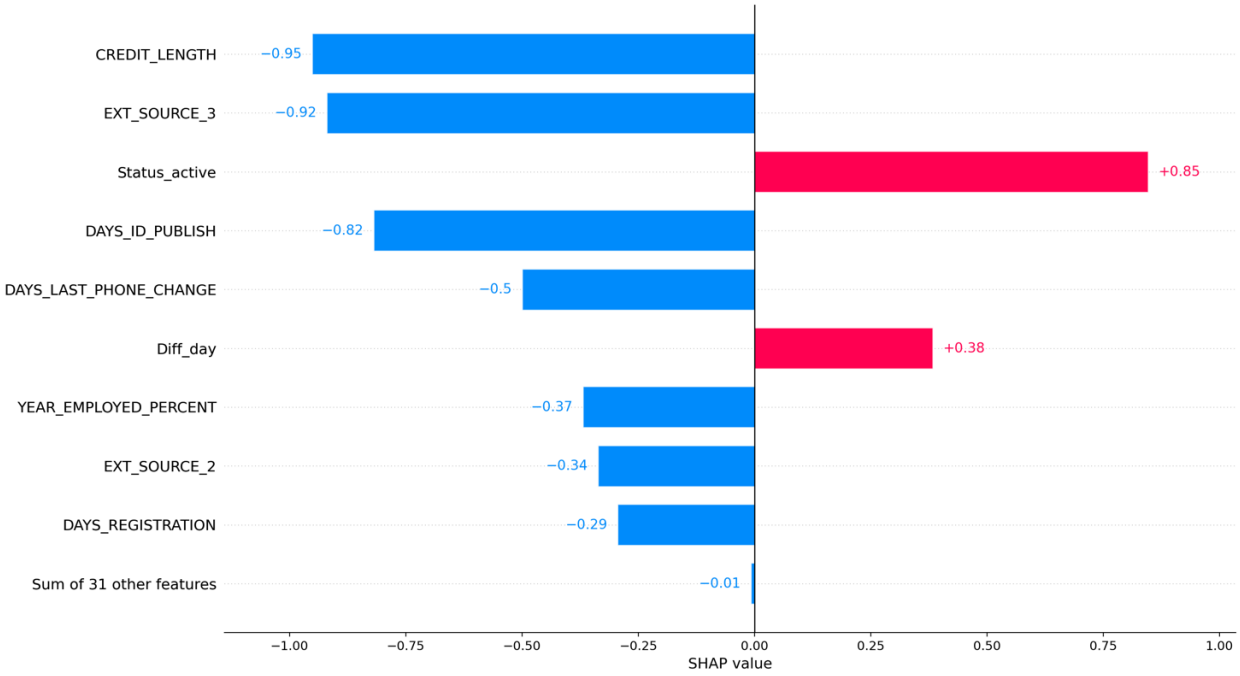
Pour que les chargés de relation client puissent expliquer les décisions d'octroi de crédit, il faut leur fournir un module d'explication simple à lire et interpréter . Donc fournir les principales causes d'accord ou refus de crédit nous semble la bonne approche . La méthode SHAP (SHapley Additive exPlanations) nous permet de calculer la moyenne de l'impact d'une variable sur la valeur prédite.

Le graphe ci-dessous représente les variables importantes sur la prédiction.



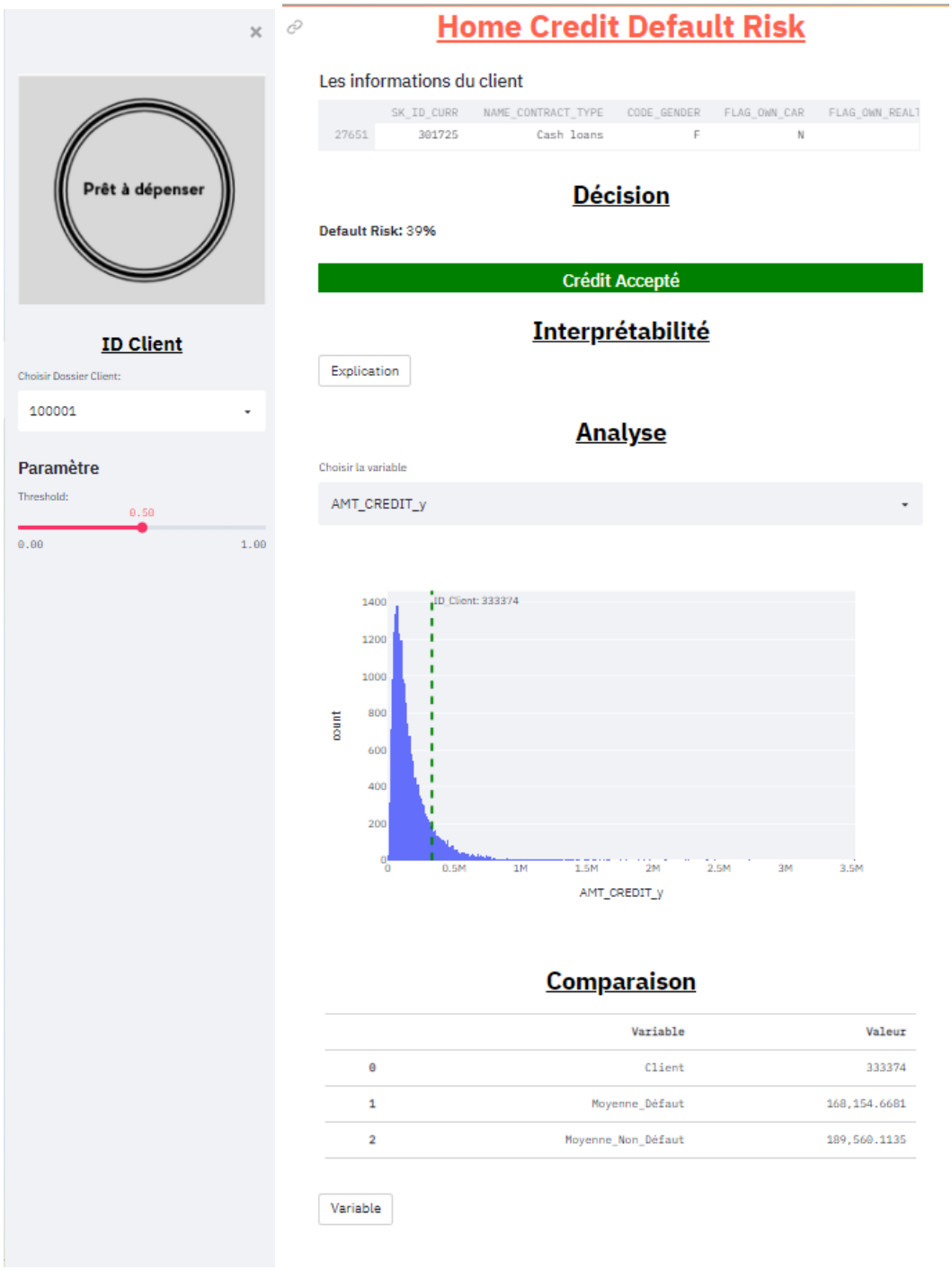
Les caractéristiques qui poussent la prédiction à être négative (vers la gauche, client n'est pas en défaut) sont affichées en bleu, et celles poussant la prédiction à être positive sont en rose (client à défaut).

Les deux graphes ci-dessous représentent les impacts positif ou négatif des valeurs des variables importantes sur la prédiction.



4. Dashboard et 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/)



5. Limites et 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.

- **Dashboard**

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

6. Références

https://towardsdatascience.com/model-performance-cost-functions-for-classification-models-a7b1b00ba60

https://machinelearningmastery.com/cost-sensitive-learning-for-imbalanced-classification/

https://towardsdatascience.com/model-performance-cost-functions-for-classification-models-a7b1b00ba60

https://www.kaggle.com/c/home-credit-default-risk/discussion/58209