

A dark blue vertical bar on the left side of the page. A blue arrow points to the right from the bar, containing the date.

15/01/2022

# **Implémentez un modèle de scoring**

Note méthodologique

Several thin, curved lines in dark blue and light grey that sweep upwards from the bottom left towards the center of the page.

**Khalil Henchi**

Openclassrooms Parcours Data Scientist

## Table des matières

I.	Le contexte .....	2
1.	Introduction .....	2
2.	Les données .....	2
II.	Méthodologie et flux de travail .....	3
1.	Prétraitement des données .....	3
2.	Choix de l'algorithme de modélisation .....	3
3.	Métriques d'évaluation des modèles .....	3
i.	Métrique technique .....	4
ii.	Métrique métier .....	5
4.	Entraînement du modèle.....	6
III.	Interprétabilité .....	7
1.	Interprétabilité globale .....	7
2.	SHAP.....	7
IV.	Limites et amélioration.....	8

# I. Le contexte

## 1. Introduction

La société financière nommée "**Prêt à dépenser**" propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt. Cette entreprise souhaite développer un outil de scoring permettant de classer les profils de clients en fonction de probabilité de défaut de paiement. Cet outil aidera ainsi les experts métiers pour appuyer la décision d'accorder ou non un prêt à un client.

L'outil est entraîné avec un jeu de données englobant des types de données variées : âge, salaire, statut professionnel, et également des informations provenant d'autres institutions financières, etc.

Pour faciliter l'utilisation de l'outil et améliorer l'expérience utilisateur, la société décide 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.

## 2. Les données

Le jeu de données est composé de 9 fichiers. Le fichier de descriptions des variables indique que l'ensemble du jeu représente :

- Nombre de colonnes : 335 variables
- Nombre de lignes : 356251

Une préanalyse met en avant que la variable cible est déséquilibrée :

- Classe 0 représente 92% des crédits
- Classe 1 représente seulement 8% des crédits

## II. Méthodologie et flux de travail

### 1. Prétraitement des données

Le prétraitement, preprocessing en anglais, est la première étape généralement dans un workflow data.

Dans notre cas, le travail consiste en :

- Fusion des données : agrégation des différentes tables des données.
- Features engineering : la création de variables statistiques (moyenne, minimum, maximum, etc).
- Data cleaning : on supprime les valeurs aberrantes, les variables avec un taux de valeurs manquantes remarquables
- Imputation des valeurs manquante
- Features encoding : on transforme les variables catégorielles en des variables numériques utilisables par les algorithmes de machine learning
- Préparation des jeux d'entraînement et de la validation

### 2. Choix de l'algorithme de modélisation

Le problème métier consiste à classifier des demandes de crédit en solvables et non solvables. Il s'agit ainsi d'un problème de classification binaire où :

- La classe 0 : représente les demandes de clients acceptés
- La classe 1 : représente les demandes de clients rejetés.

Pour choisir le meilleur algorithme, on utilise différents modèles et on évalue leurs performances en se basant sur nos métriques d'évaluations.

### 3. Métriques d'évaluation des modèles

A travers ce projet, on cherche à augmenter les bénéfices de la société « **Prêt à dépenser** ». En effet, on veut éviter de perdre des potentiels nouveaux clients, bons payeurs, on doit limiter ainsi le nombre de faux positifs (client catégorisé comme mauvais alors qu'il est bon).

On va donc chercher à minimiser le FPR (False Positive Rate), donné par la formule suivante :

$$FPR = \frac{FP}{FP + TN}$$

De même, on cherche à éviter qu'un client soit catégorisé comme bon, alors qu'il est un mauvais (faux négatif). On cherche donc à maximiser le Recall ou TPR (True Positive Rate), donnée par la formule suivante :

$$TPR = \frac{TP}{TP + FN}$$

La matrice de confusion suivante illustre les différents cas possibles.

			Prédiction	
			Positive	Negative
			0	1
Réelle	Positive	0	True positive (TP)	False negative (FN)
	Negative	1	False positive (FP)	True negative (TN)

Vu la nature de nos besoins, on utilise les métriques suivantes :

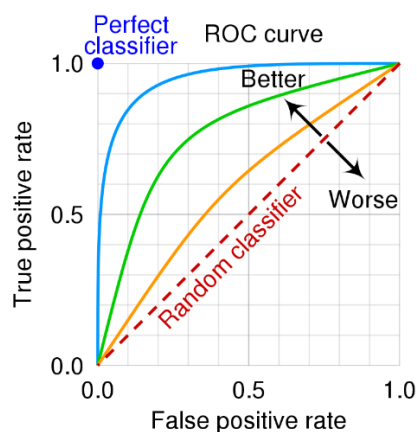
- Métrique technique
- Métrique métier

#### i. Métrique technique

En se basant sur le contexte métier, une métrique technique est intéressante à utiliser pour mettre en relief le compromis entre la détection d'un bon client et une fausse détection d'un mauvais client. Il s'agit de la AUC-ROC : la courbe ROC est un graphique qui montre les performances d'un modèle de classification à tous les seuils possibles (le seuil est une valeur particulière au-delà de laquelle vous dites qu'un point appartient à une classe particulière). La courbe est tracée entre deux paramètres :

- Taux de Vrai Positif TP
- Taux de Faux Positifs FP

La figure suivante illustre la courbe AUC-ROC.



## ii. Métrique métier

Dans le contexte Business :

- Un faux positif : client accepté considère comme rejeté : Client ayant un bon profil à qui on rejete un prêt
- Un faux négatif : client rejeté considère comme accepté : Client ayant un mauvais profil à qui on accorde un prêt

Il est clair qu'un faux négatif est plus coûteux qu'un faux positif pour la société. Ainsi, on doit donner plus de poids à la minimisation des faux positifs. La fonction à minimiser est donc Beta-score donnée par la formule suivante :

$$F_{\beta} = (1 + \beta^2) \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

En effet, le F-beta score permet de varier la pondération de la precision et du recall, et donc aux faux positifs et aux faux négatifs. Pour résumer :

- Pour  $\beta \geq 1$ , on accorde plus d'importance au recall (autrement dit aux faux négatifs).
- Pour  $\beta \leq 1$ , on accorde plus d'importance à la precision (autrement dit aux faux positifs).
- Pour  $\beta = 1$ , on retrouve le F1-score, qui accorde autant d'importance à la precision qu'au recall.

Dans ce projet, on choisit  $\beta = 1$ , choix arbitraire pour essayer de maximiser les gains.

#### 4. Entraînement du modèle

La méthode de validation croisée a été utilisée pour l'entraînement des modèles afin d'obtenir un modèle qui arrive à généraliser le problème. Cette validation croisée est de plus stratifiée pour avoir un nombre de classe égales entre les jeux de données séparés. En effet, on a implémenté la validation croisée à 10 blocs, « 10-fold cross-validation » : on divise l'échantillon original en 10 échantillons (ou « blocs »), puis on sélectionne un des 10 échantillons comme ensemble de validation pendant que les 9 autres échantillons constituent l'ensemble d'apprentissage. Après apprentissage, on peut calculer une performance de validation. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les blocs prédéfinis. À l'issue de la procédure nous obtenons ainsi 10 scores de performances, un par bloc. La moyenne et l'écart type des 10 scores de performances peuvent être calculés pour estimer le biais et la variance de la performance de validation. La figure suivante illustre un des résultats trouvés.

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F10Score
0	0.6786	0.6910	0.6273	0.1490	0.2408	0.1260	0.1788	0.6080
1	0.6930	0.7029	0.6237	0.1549	0.2481	0.1357	0.1880	0.6056
2	0.6772	0.7028	0.6466	0.1515	0.2455	0.1312	0.1874	0.6263
3	0.6906	0.6999	0.6309	0.1550	0.2489	0.1362	0.1898	0.6123
4	0.6920	0.7029	0.6219	0.1541	0.2470	0.1343	0.1863	0.6038
5	0.6847	0.6864	0.6051	0.1479	0.2376	0.1232	0.1721	0.5871
6	0.6893	0.6995	0.6183	0.1522	0.2443	0.1310	0.1824	0.6001
7	0.6802	0.6786	0.5912	0.1435	0.2310	0.1153	0.1617	0.5735
8	0.6780	0.6873	0.6153	0.1467	0.2369	0.1217	0.1723	0.5964
9	0.6763	0.6972	0.6373	0.1495	0.2422	0.1275	0.1821	0.6174
Mean	0.6840	0.6948	0.6218	0.1504	0.2422	0.1282	0.1801	0.6031
SD	0.0063	0.0081	0.0150	0.0036	0.0054	0.0064	0.0085	0.0143

Suite à cette étape, une étape d'optimisation est obligatoire afin de trouver les meilleurs hyperparamètres répondants à nos besoins métier. On a testé des optimisations par des méthodes de Random Grid Search pour trouver les meilleurs hyperparamètres.

### III. Interprétabilité

#### 1. Interprétabilité globale

Les classifieurs basés sur la notion des arbres de décisions ont des méthodes permettant de trouver le poids et l'importance de chaque variable dans le résultat final du modèle. Ces méthodes permettent d'avoir une interprétation globale des résultats.

#### 2. SHAP

La librairie SHAP, SHAP (SHapley Additive exPlanations), est une approche théorique des jeux pour expliquer la sortie de tout modèle d'apprentissage automatique. Il explique la prédiction en locale en utilisant les valeurs de Shapley classiques de la théorie des jeux pour montrer la contribution de chaque variable à l'output final du modèle. Comme le montre la figure suivante :











## IV. Limites et amélioration

Le flux d'un projet de data passe à travers différentes étapes de transformation, de traitement et d'analyses des données. Ainsi, pour optimiser notre flux, nous devons investiguer et interroger chaque étape pour trouver les meilleurs paramètres et les meilleures combinaisons de méthodes possibles afin d'optimiser nos résultats.

Ainsi, on peut améliorer nos résultats en travaillant les points suivants à titre d'exemple :

-  Nettoyage de donnée : Pour optimiser nos résultats, on peut envisager autres approches pour nettoyer le jeu de données :
  - Suppression des colonnes en fonction d'un taux de remplissage optimal
  - Suppression des outliers en utilisant différentes méthodes
-  Imputation des variables : Pour optimiser nos résultats, on peut envisager autres approches pour imputer le jeu de données :
  - Utilisation de modes à la place de mean
  - Utilisation des algorithmes d'imputation multivariés comme : IterativeImputer, etc.
-  Sélection des variables : Pour sélectionner les meilleures variables à utiliser, on peut se référer aux experts métier en vue de définir les techniques à utiliser et marquer les variables les plus importantes.
-  Equilibrage des données : Pour équilibrer les données, nous pouvons utiliser d'autre approches comme la création des modèles d'ensemble ou de bagging entraîné chacun sur un set équilibré de données, également, on peut toujours collecter plus de données pour résoudre ce problème.
-  Fonction d'évaluation métier : Les règles et les métriques d'évaluation métier doivent être définies et approuvées par les experts métiers en vue de développer un modèle adapté à notre problème métier.
-  Optimisation des hyperparamètres : Une recherche des meilleurs combinaisons d'hyperparamètres plus poussée et avec d'autres algorithmes bayésien peut donner des meilleurs résultats.