Openclassrooms Parcours Data Science Projet 7

Dashboard sur les données bancaires pour évaluer la solvabilité client

Nicolas Martinez

# Introduction

Cette note a pour but d’expliquer les différentes étapes du projet.

Le but du projet est de mettre à disposition un tableau de bord permettant d’évaluer la solvabilité de particulier et ainsi proposer une recommandation au secteur de la banque pour l’attribution de prêt bancaire à ce potentiel client. Ce tableau de bord interroge une API (Application Programming Interface).

Ces deux objets sont hébergés sur des plateformes cloud afin d’être accessible partout sur le web.

Pour la partie modélisation des données, il s’agit d’un problème de classification. Les résultats sont classés dans les 4 états suivants :

* Vrais positifs : les clients que le modèle estime solvable et qui remboursent leur prêt sans délai.
* Vrais négatifs : les clients que le modèle estime à risque et qui ont eu des retards avec leurs échéances.
* Faux positifs : les clients que le modèle estime solvable mais qui ont eu des retards avec leurs échéances.
* Faux Négatifs : les clients que le modèle estime non solvable mais qui le sont en réalité.

Le cœur du sujet se situe sur les 2 derniers aspects (faux positifs & faux négatifs) car ce sont ceux qui coutent à la banque par leur retard de paiement ou représentent un manque à gagner car ils auraient été de bons clients.

Le modèle que l’on construit est donc paramétré en tenant compte de ce point clé.

# La méthodologie d'entraînement du modèle (2 pages maximum)

L’entrainement du modèle se déroule en 5 étapes :

1. Appropriation des données

Cette partie se rapporte à la première partie du Notebook P7\_Scoring\_00\_Données\_MN

Le jeu de données est une série de fichier CSV contenant des données bancaires et financières de particuliers.

Dans cette partie on effectue les opérations suivantes :

* Chargement des données csv dans des dataframes Pandas
* Analyse univariée des données
* Remplissages des données manquantes par le mode pour ne pas modifier la distribution et subir les effets de points aux valeurs extrêmes (outliers). Des opérations différentes sont utilisées pour les données textes, catégories et numériques discrètes d’une part, et les valeurs numériques continues d’autre part.
* Fusion des différents dataframes

1. Préparation des données
2. Feature engineering
3. Recherche du modèle le plus efficace
4. Optimisation du modèle

# La fonction coût métier, l'algorithme d'optimisation et la métrique d'évaluation (1 page maximum)

# L’interprétabilité globale et locale du modèle (1 page maximum)

# Les limites et les améliorations possibles (1 page maximum)

Modele : Pipeline

Dashboard : ajout des variables