Notes méthodologiques

Costumer credit allocation

Projet 7: OpenClassrooms
Implémentez un modèle de scoring

Par Rouba Yaacoub

Table des matières

A. La mission

I. Introduction

L'entreprise "**Prêt à dépenser**" souhaite **mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité** qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé.

Dans une volonté de transparence auprès de ses clients, l'entreprise souhaite mettre en place un dashboard interactif et facile de présentation afin que les chargés de relation clients puissent exposer les décisions d'octroi ou non de prêt.

II. Data

Les data est télécharge depuis kaggle. La base de données rassemble des informations pour 301 511 clients selon 122 indicateurs. Ces informations sont du type générique comme l'âge, le sexe, les revenus, l'emploi, le logement, les informations de crédit en cours, des notations externes, etc.

B. Méthodologie d'entraînement du modèle

I. Traitement données

Afin d'exploiter au mieux le set de données, des étapes de prétraitement doivent être appliquées. Il s'agit d'étapes de nettoyage des données, d'encodage numérique des variables catégorielles, ainsi que la création de nouveaux indicateurs comme le terme du crédit. Enfin une imputation des valeurs manquantes est réalisée et une mise à l'échelle est appliquée.

II. Algorithme et Sampling

Le modèle de classification choisit pour cette étude est le modèle Light Gradient Boosting Machine ou Light-GBM. Il s'agit d'un modèle robuste, originellement développé par Microsoft, et qui repose sur le principe d'arbres de décisions.

Le **Sample weights** est une méthode directement gérée par le modèle Light-GBM et qui permet de pénaliser les poids associés aux observations sur représentées.

C. Performance du modèle

I. Score métrique

Dans notre travail, nous avons implementé plusieurs scores afin d'évaluer la performance du modèle.
La matrice de confusion étant le meilleur indice pour cette évaluation.

Dans ce calcul de gain, si un prêt est accordé à une personne non solvable, la banque perdra de l'argent. Il s'agit des «**False Negatif**» dans la matrice. De même si un prêt est refusé à une personne solvable, les «**False Positif**», la banque aura perdu un gain d'argent potentiel. Au contraire si l'attribution, «**True Negatif**» ou le refus, «**True Positif**», de prêt est correct la banque gagnera de l'argent ou à défaut n'en perdra pas.

		Actual Value	
		Present	Absent
Predicted Value	Present	TP	FP
	Absent	FN	TN

- FN ➔ Perte d'argent pour la banque : **- 100**
- TP ➔ Refus de prêt, la banque ne perd pas d'argent : **+10**
- TN ➔ Prêt accordé, gain d'argent pour la banque : **+ 10**
- FP ➔ Client potentiel perdu, perte d'argent pour la banque : **- 1**

Tot = FN-100+TP*10+TN*10+FP*-1*
$gain_{max}$ *= (TN+FP)*10 + (TP+FN)*0*
$gain_{min}$ *= (TN+FP)*-1 +(TP+FN)*-100*

$$gain = \frac{(Tot - gain_{min})}{(gain_{Max} - gain_{min})}$$

gain = **91%**

D. Interprétabilité

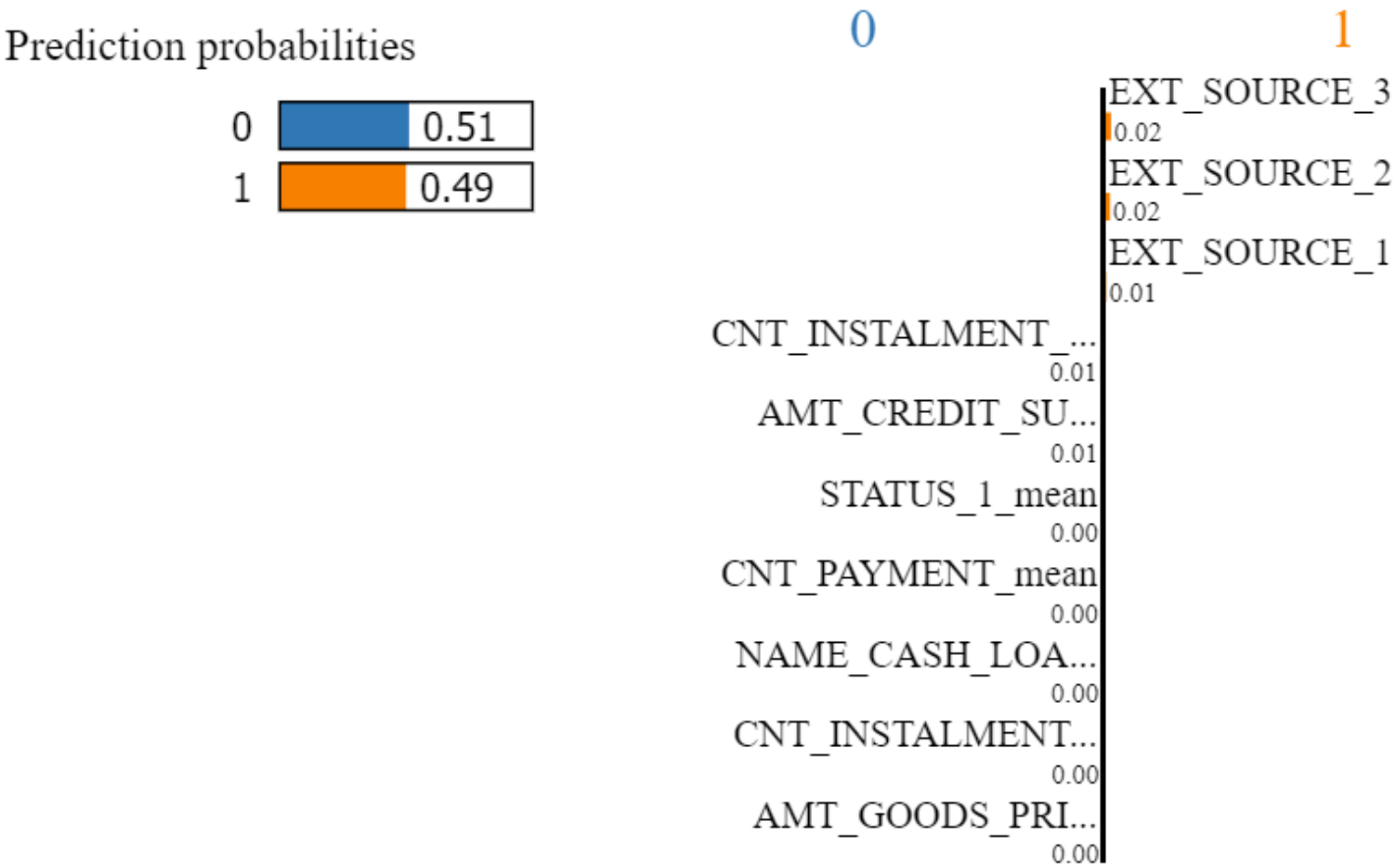
Pour garantir la transparence auprès de ses clients, un **dashboard interactif** est réalisé afin de comprendre comment la décision d'attribution ou de refus de prêt est prise. Donc pour interepreter le modèle, nous allons utiliser la l'algorithme Lime.

Résultats sur : https://score-credits.herokuapp.com/

I. L'algorithme Lime

L'algorithme **Lime** permet d'observer l'évolution des prédictions en fonction de l'importance de certains critères. Grâce à Lime nous pouvons expliquer facilement les raisons pour lesquels chaque individu a été scoré de telle ou telle manière. [Figure en page 6]

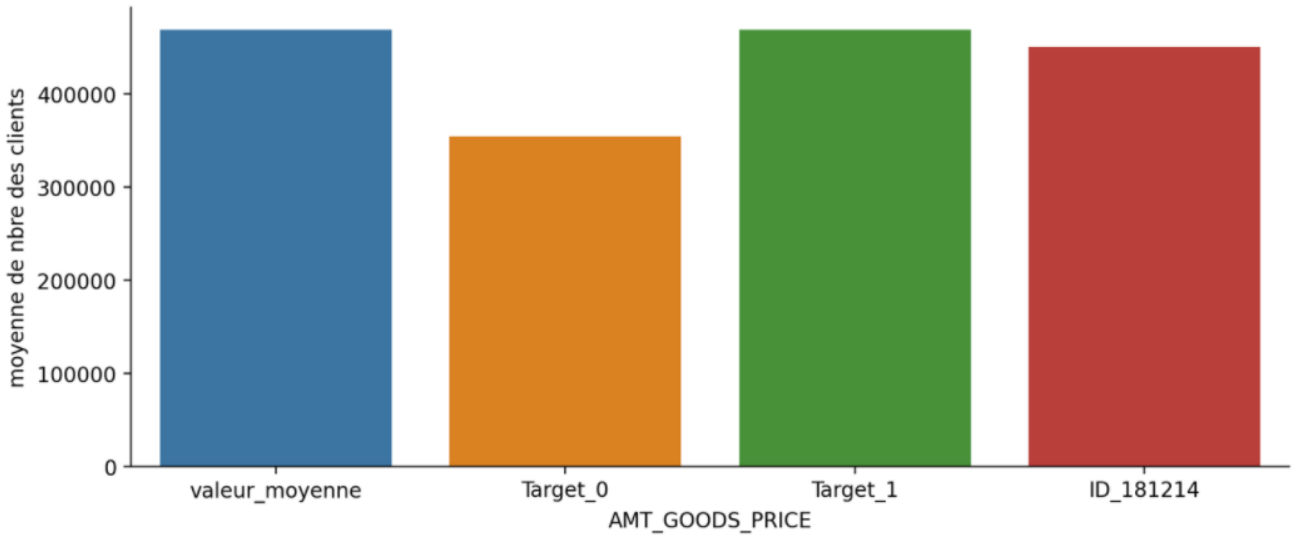
Pour ce client, les indicateurs «AMT_GOODS_PRICE_mean» et «CNT_INSTALLEMENTS_FUTURE_mean» ont faire tendre le modèle vers l'attribution du prêt bancaire au contrare des "EXT_SOURCE" 1,2 et 3 qui ont augmenter le risque de défaut de remboursement.

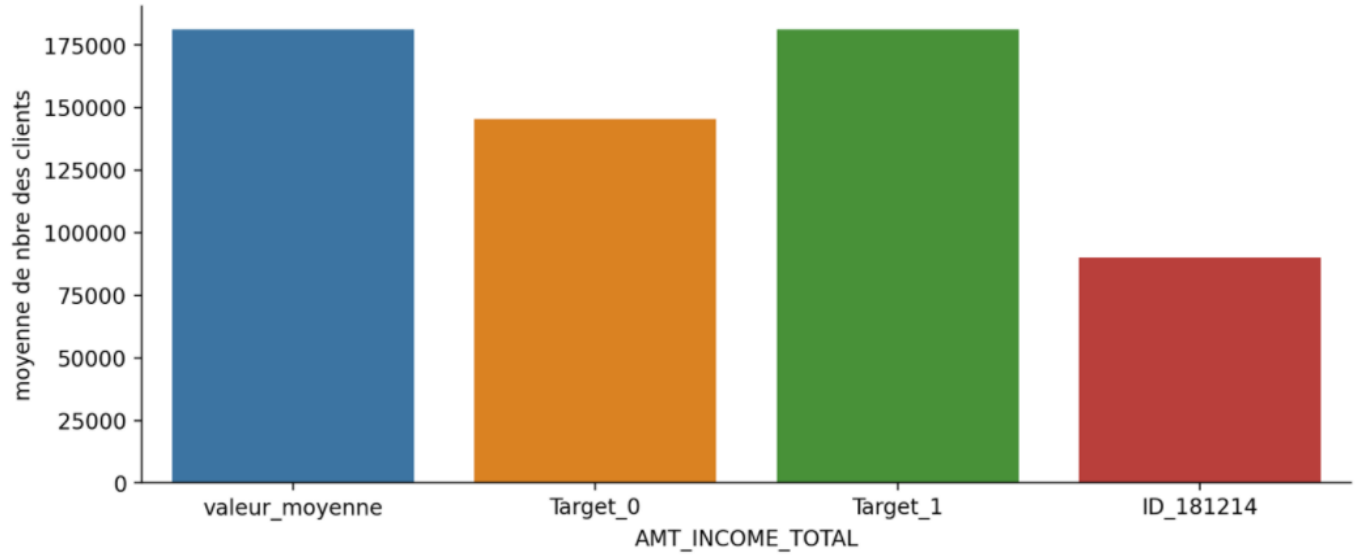


Feature	Value
EXT_SOURCE_3	0.10
EXT_SOURCE_2	0.26
EXT_SOURCE_1	0.00
CNT_INSTALMENT_FUTURE_mean	37.78
AMT_CREDIT_SUM_DEBT_mean	719811.00
STATUS_1_mean	5.00
CNT_PAYMENT_mean	56.00
NAME_CASH_LOAN_PURPOSE_Repairs	3.00
CNT_INSTALMENT_mean	46.28
AMT_GOODS_PRICE_mean	1050000.00

II. Approche global

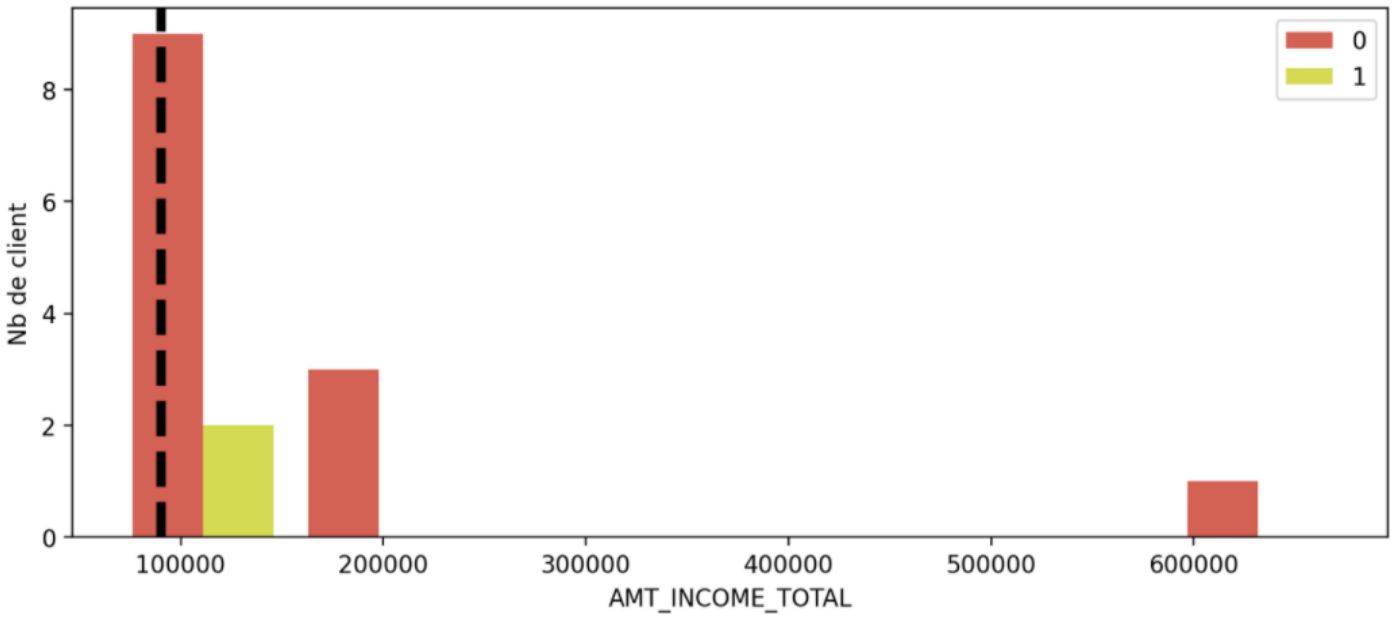
A partir des indicateurs identifiés par la librairie Lime, il est alors possible de comparer l'information d'un client pour ces indicateurs par rapport aux données des autres clients de la base de données. Les distributions de la totalité des clients ainsi que celles des clients solvables et non solvables sont affichées sur le dashboard.

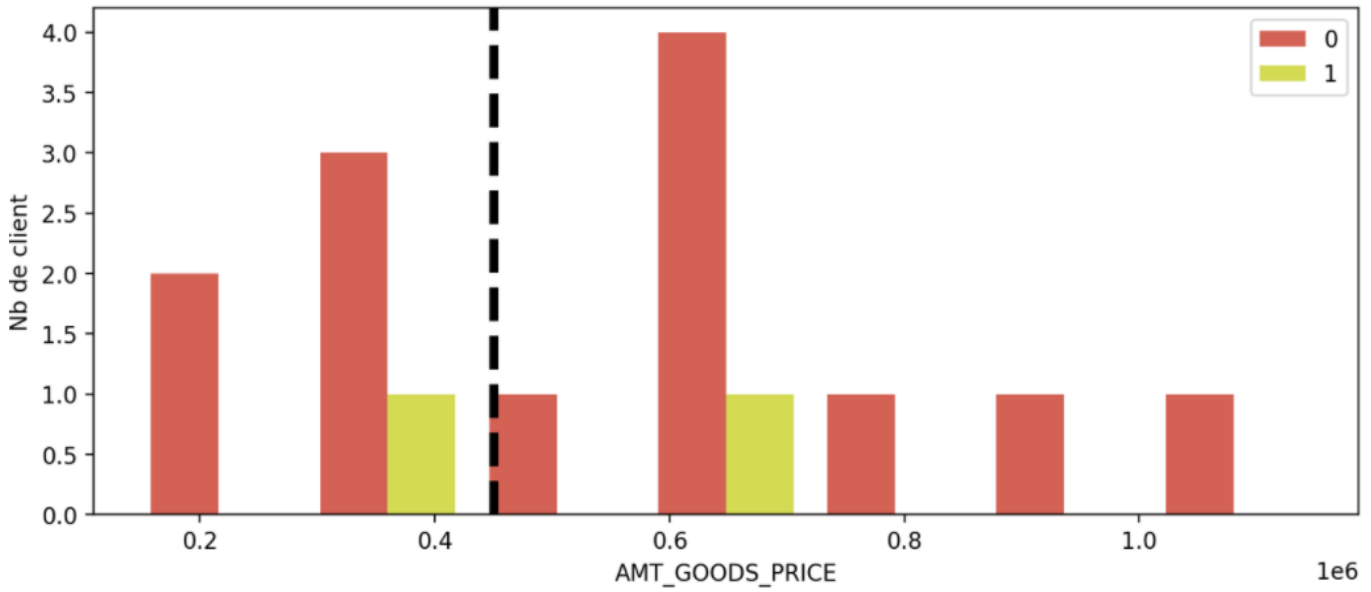




III. Profils similaires

Enfin, il est également possible d'afficher ou se situe un client par rapport aux clients de même profil. Le profil est définit selon des critères simples comme l'âge, le sexe, le nombre d'enfant ou encore le code région d'habitation.





E. Limites et améliorations possibles

Les paramétres utilisés dans le GridSearchCV peuvent etre grandement améliorés.

Une seconde version du modèle pourra être mise en œuvre en considérant les méthodes de gestion de déséquilibrage des données, notamment l'over sampling, under sampling, afin d'éviter le sur apprentissage du modèle sur des données fictives qui peuvent être erronées.

Les algorithmes des fonction de gain peuvent etre modifiable selon la société bancaire.