Note méthodologique

Open Classrooms
Projet 7 Parcours Data Scientist
« Implémentez un modèle de scoring »

Luc Rogers

08/10/2021

I. Introduction

Cette note méthodologique constitue un des livrables du projet 7 du parcours Data Scientist d'Open Classrooms. Elle présente dans les grandes lignes la méthodologie employée pour réaliser le modèle de scoring, son interprétabilité et le fonctionnement du dashboard interactif associé.

Données : https://www.kaggle.com/c/home-credit-default-risk/data

Livrables : https://github.com/lucrogers/Projet-7

Dashboard interactif : https://loan-management-dash.herokuapp.com

II. Problématique

La société « Prêt à dépenser » propose des crédits à la consommation pour des personnes ayant peu ou pas d'historique de prêt. Pour guider la prise de décision d'octroi ou de refus d'un prêt elle souhaite développer un modèle de scoring de la probabilité de défaut de paiement client.

Par souci de transparence, le modèle doit être facilement interprétable par une personne sans connaissance particulière en data science.

De plus, la société souhaite bénéficier d'un dashboard interactif à l'attention des gestionnaires de relation client, afin que ceux-ci puissent expliquer à leurs clients avec le plus de transparence et de clarté possible les raisons du choix d'octroi ou non d'un prêt, en fonctions des informations relatives aux clients.

III. Entraînement du modèle

Les données sont constituées d'une base d'environ 300 000 emprunteurs dont l'historique de remboursement est connu. Chaque individu reçoit un score cible de 0 ou de 1 selon si l'emprunteur est respectivement en règle ou a connu un défaut de paiement à l'issu de son prêt.

1. Pre-processing

- Les informations relatives aux emprunteurs sont divisées en plusieurs datasets qu'il convient d'agréger par individu.

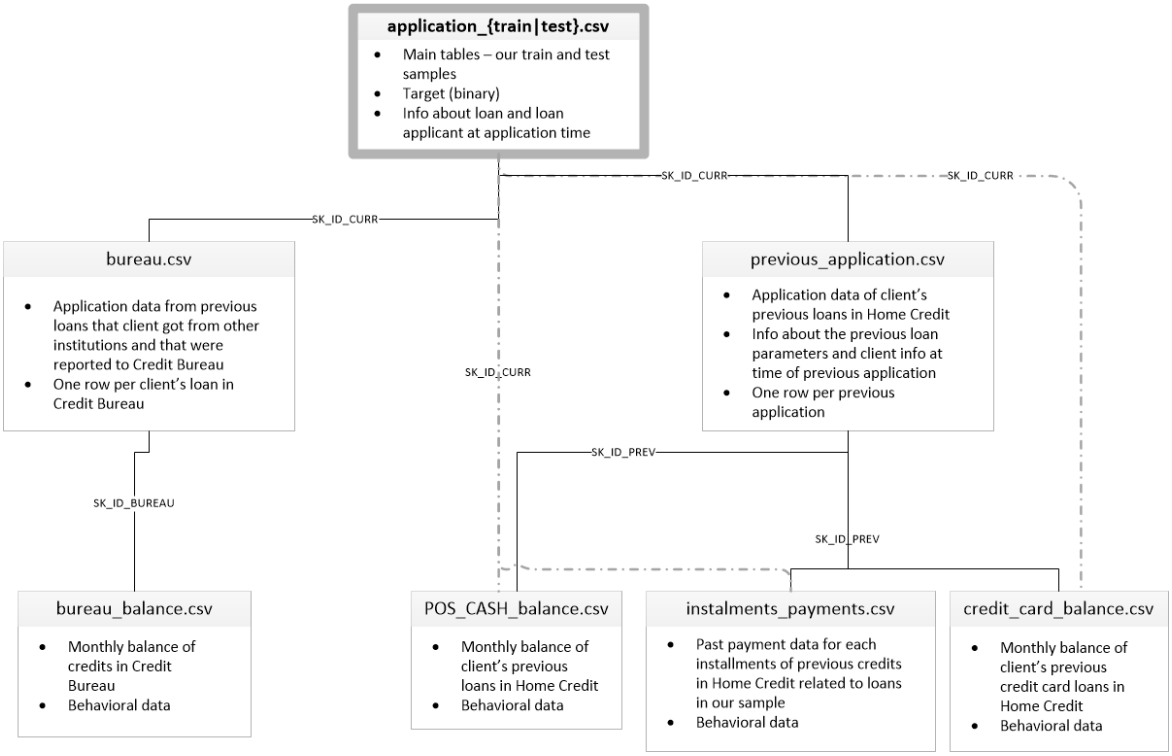

Figure 1 Structure du jeu de données

- Les valeurs aberrantes sont supprimées.
- Les valeurs manquantes sont remplacées par la moyenne ou par le mode selon si elles sont de nature quantitative ou catégorielle.
- Les variables catégorielles sont ensuite encodées avec un encodage de type one hot encoding.

2. Modélisation

La partie modélisation étant assez secondaire dans ce projet, le modèle retenu a été choisi parmi les meilleures entrées de la compétition Kaggle associée. Il s'agit d'un modèle LightGBM entraîné avec validation croisée sur 10 folds. Les hyperparamètres associés ont été trouvés grâce à une optimisation Bayesienne avec TPE (Tree Parzen Estimator).

Le score AUC associé est de 0,80.

IV. Fonction coût et scoring

1. Scoring

A ce stade, notre modèle donne une probabilité de défaut de paiement par client, qui constitue son score. Le graphe ci-dessous représente la densité des groupes d'individus en règle ou en défaut de paiement en fonction de ce score.

0 correspond à un remboursement en règle et 1 à un défaut de paiement.

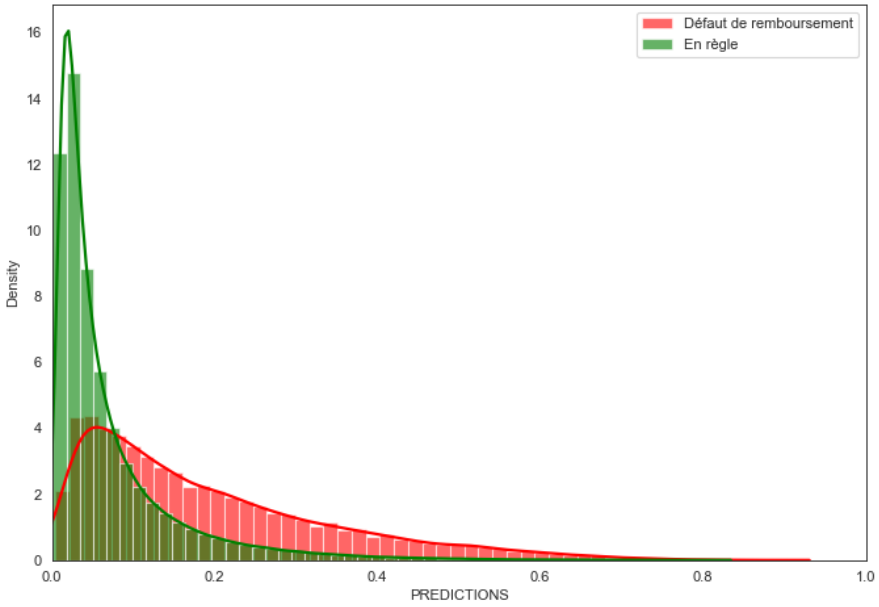

Figure 2 Probabilité de défaut de paiement selon la cible étudiée

La distribution des défauts de paiement est plus étalée vers la droite. Le fait que les modes de ces deux distributions soient si rapprochés et proches de 0 vient du fait que les dossiers des emprunteurs ont été triés au préalable et jugés aptes au remboursement.

2. Fonction coût

Il reste maintenant à définir un seuil sur le score client à partir duquel la décision de refus sera envisagée.

Si le seuil est trop élevé, c'est-à-dire si les crédits sont accordés trop facilement, l'entreprise connaîtra beaucoup de défauts de paiements et va perdre de l'argent.

Si le seuil est trop bas, trop peu de prêts seront accordés et l'entreprise connaîtra un manque à gagner.

Le choix de ce seuil résulte donc d'un compromis qui peut se calculer en fonction du profit et du coût respectivement engendrés par un client en règle ou en défaut. Notre fonction cible est donc le bénéfice réalisé par l'entreprise.

En considérant en première approche simple qu'un montant moyen est accordé à chaque client, elle est reliée au score client par la définition suivante :

$$Bénéfice = [\ (1 - p(seuil)\ ) * profit + p(seuil) * coût\ ] * n(\text{seuil})$$

Avec :

- p(seuil) : la proportion de clients en défaut de paiement
- profit : le profit moyen réalisé par un prêt en règle (en %)
- coût : le coût moyen associé à un défaut de paiement (en %)
- n(seuil) : le nombre de prêts accordés

p et *n* ne dépendent que du seuil choisi et sont connus en tous points.

On peut alors calculer le bénéfice réalisé par l'entreprise en fonction du seuil choisi sur le score client :

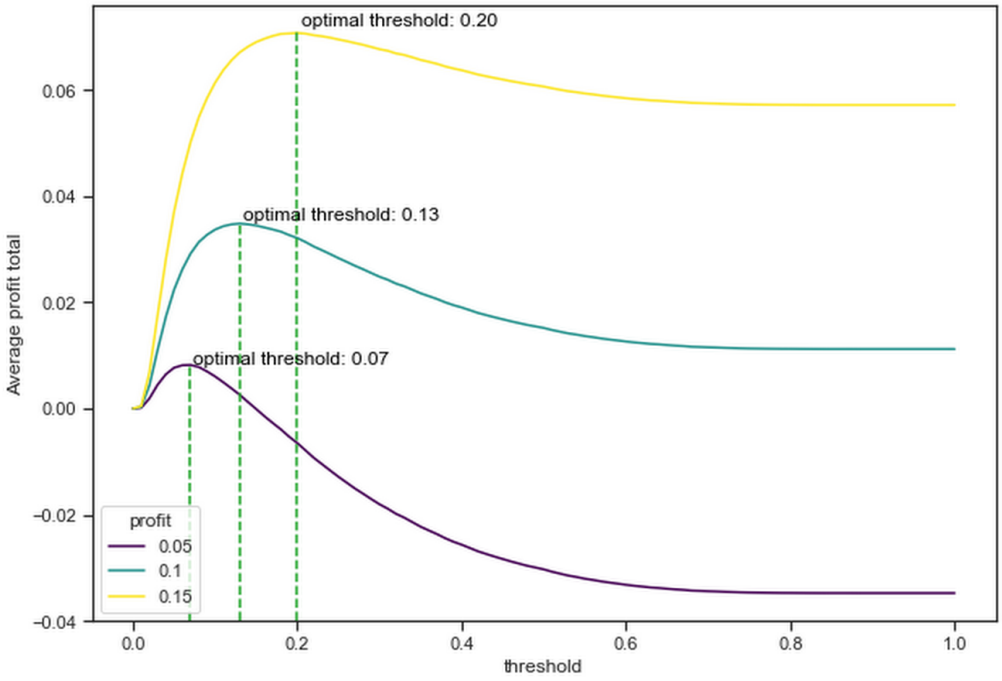

Figure 3 Bénéfice (%) en fonction du seuil sur le score client

Pour la démonstration du phénomène, on considère ci-dessus que le coût engendré par un défaut de paiement est de -100% et le profit est compris entre 5 et 15% de la valeur du montant moyen d'un prêt accordé. Ces valeurs sont à préciser avec le corps métier.

Si le profit engendré par un crédit remboursé est trop faible (5% , courbe violette), on est rapidement déficitaire en acceptant des crédits risqués. En revanche, si le profit est plus important (15%, courbe jaune), on gagne à accepter des clients plus à risque.

Pour la suite du projet et dans le dashboard, le seuil de 0,20 sur le score client sera retenu.

V. Interprétabilité du modèle

Le modèle doit permettre aux gestionnaires de justifier auprès de leurs clients leur prise de position sur l'approbation ou le refus du crédit. Pour cela nous utilisons l'approche théorique SHAP (Shapley Additive exPlanations) qui permet de quantifier l'effet de chaque variable dans l'attribution du score final.

Sur Python, on utilise la librairie *shap* et on fit l'explainer shap à notre modèle lgbm, préalablement enregistré avec *pickle*.

Le résultat par client, présenté sous forme de graphique en cascade, permet de quantifier de façon claire et simple l'impact des features qui ont joué le rôle le plus important dans la note finale du client. En rouge, les variables augmentent la probabilité de défaut de paiement et en bleu elles augment la probabilité de remboursement.

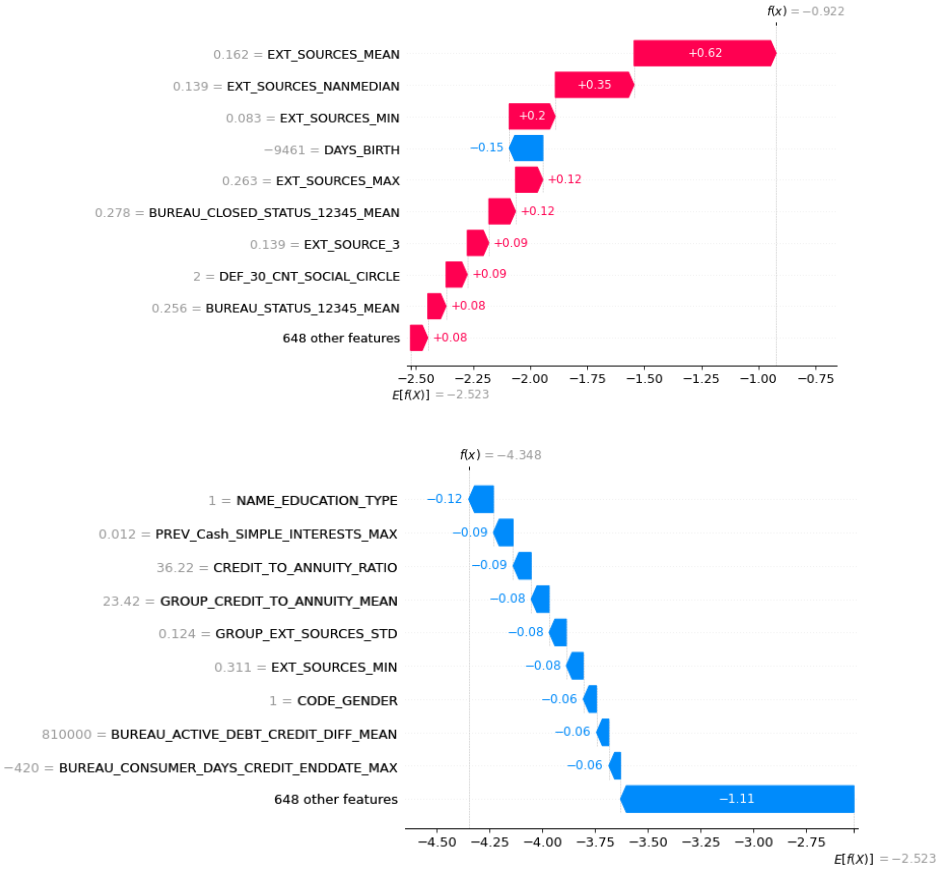

Figure 4 Exemples de valeurs de Shapley pour un client en défaut et un client en règle

Dans notre dashboard interactif, sous ce graphique, un tableau permet de rappeler à l'utilisateur la définition des variables qui ont eu le plus d'impact dans la notation du client. Ce tableau est mis à jour de façon dynamique lorsqu'un client est sélectionné, de façon à afficher les variables qui lui correspondent.

Le remplissage des définitions des 657 variables du modèle est hors spectre de ce projet, les définitions affichées dans le dashboard sont donc pour l'instant vides.

La simplicité d'utilisation et de compréhension du dashboard permet ainsi d'être aussi facilement utilisé par le gestionnaire que par le client, ce qui va notamment dans le sens des valeurs de transparence revendiquées par l'entreprise.

VI. Limites du modèle et pistes de réflexion

- En premier lieu, commençons par avertir que le modèle présenté ici ne peut se suffire à lui-même et se substituer à une étude préalable du dossier client par le prêteur.

En effet, un point important à souligner est que notre base de données est constituée uniquement d'**individus pour lesquels un prêt a été accordé** puisque tous les individus jugés trop à risque par les organismes de prêt ont été préalablement exclus. En d'autres termes, la distribution des individus sur laquelle est basé notre modèle ne reflète pas la distribution des demandeurs de prêt.

Ne pas faire ce tri préalable engendrerait une distribution différente des clients ayant les capacités ou non de remboursement, ce qui aurait deux effets majeurs : le premier étant que le modèle ne serait plus adapté à cette nouvelle population, ce qui pourrait changer drastiquement l'importance des différentes variables ;

le second étant que la fonction coût, calculée au chapitre IV de cette note, serait faussée puisque elle repose en partie sur l'observation empirique du taux de personnes en défaut de paiement dans notre base de données, pour un seuil donné.

Le fait que cette base de données soit constituée de différentes sources correspond déjà à une approximation en soi puisque chaque banque a ses propres règles d'attribution de prêt, et certaines sont fatalement plus efficaces que d'autres.

Pour palier à ce problème et afin d'arriver à un algorithme de classification auto-suffisant, il faudrait refaire cette même étude sur une population pour laquelle un crédit est accordé sans considération préalable au dossier client. En estimant le nombre de demandeurs de prêt potentiellement à risque dans la population générale, on pourrait estimer le coût d'une telle étude et réfléchir au bénéfice qu'elle apporterait. On pourrait être tenté de réaliser une telle étude avec des petits montants de prêts, ce qui en minimiserait le coût, malheureusement ce faisant on introduirait un facteur de confusion puisque les personnes qui empruntent des petits montants ont un profil différent des personnes qui empruntent de gros montants.

- La fonction coût est ici calculée en faisant l'approximation d'un montant de prêt moyen par client, ce qui repose sur l'hypothèse que le montant accordé ne dépend pas du niveau de risque du client. On peut légitimement penser que les prêteurs sont plus réticents à prêter de plus gros montants lorsque le client a un profil plus risqué. Ce qui aurait pour effet de surestimer le bénéfice de l'entreprise en calculant un seuil de cette manière. On pourrait imaginer affiner le calcul en pondérant *n* par la valeur moyenne du montant associé au risque client.

- La fonction coût repose sur l'aspect purement financier de rentabilité pour l'entreprise. Si cette dernière souhaite défendre certaines valeurs (comme les banques éthiques, solidaires, etc…), et puisqu'il y a souci de transparence dans l'accord des prêts, il faudra envisager une marge pour les clients un peu plus à risque mais dont le projet s'inscrit dans les valeurs ouvertement défendues par l'entreprise. Ce qui se traduirait in fine par le troc d'une partie de la rentabilité au profit du bien commun, tel que perçu par la direction.

- Le dernier point n'est pas une limite au modèle en soi, mais il convient de se rapprocher du corps métier afin d'évaluer les coûts et profits définis dans la formule de la fonction coût (§IV). Notamment car dans notre approche, on considère que le moindre défaut de paiement engendre systématiquement la perte du montant du prêt, ce qui met au même niveau l'emprunteur qui n'a eu qu'un incident sur une seule mensualité et celui qui n'a jamais rien remboursé. Une plus grande finesse de ce paramètre permettrait d'optimiser les bénéfices, tout en accordant plus de crédits.

- Une amélioration possible du dashboard serait de prévoir un espace simulation qui permettrait de modifier les données clients en mettant à jour le score associé, de façon à proposer des objectifs concrets dans l'obtention d'un crédit.