Note Méthodologique

Parcours Data Scientist
Projet 7

Implementez un modèle de scoring

Paul De Ruta

Table des matières

I. Contexte

Ce travail s'inscrit dans le cadre du projet 7 de la formation OpenClassrooms : « Implémentez un modèle de scoring ».

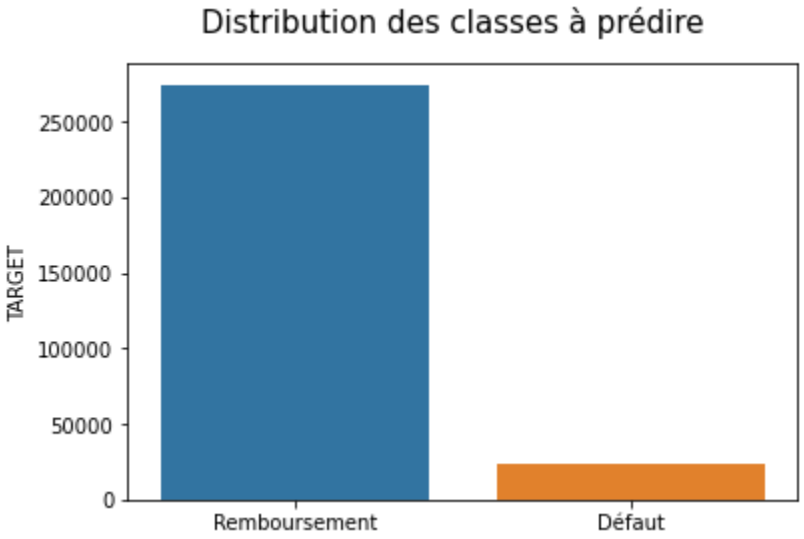
L'objectif est de développer pour l'entreprise « Prêt à dépenser » un outil de scoring permettant d'aider à la décision d'attribution d'un crédit.

Sur les données fournies nous entraînons un modèle capable de générer un probabilité de défaut de paiement par client.

II. Méthodologie

1. Gestion du déséquilibre et recherche du meilleur algorithme

Nous travaillons sur un problème de classification binaire. Ce qui rend la tâche plus délicate ici est la disproportion d'une classe par rapport à l'autre.



Nous avons mis en place une méthodologie spéciale pour faire face à cette situation. En effet, sans traitement particulier, le modèle risque de simplement toujours prédire la classe majoritaire.

Trois outils furent employés pour mitiger ce problème :

Le sous-échantillonage, technique consistant à réduire la quantité de données dans la classe majoritaire.

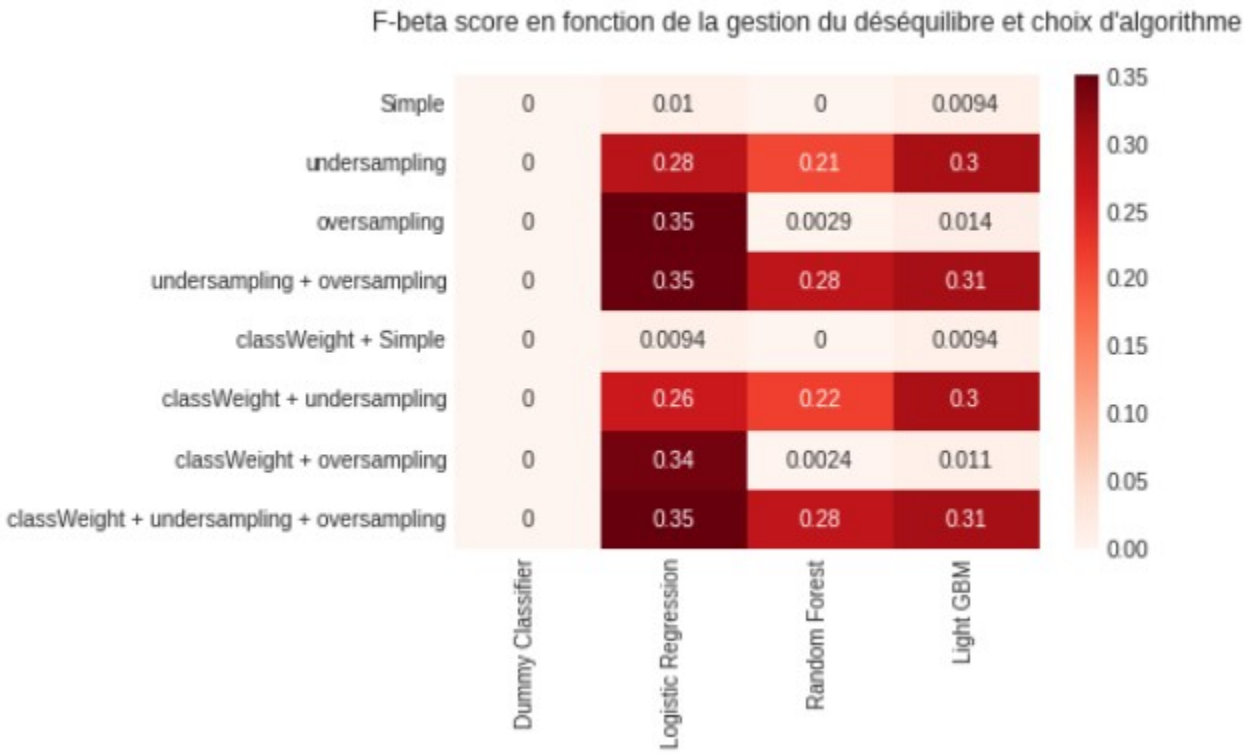
Le sur-échantillonage, implémenté par l'algorithme SMOTE, est une technique qui consiste à augmenter synthétiquement la quantité de données de la classe minoritaire. Celui-ci fonctionne en interpolant des points de données entre les données déjà existantes.

Le « class-weight » qui permet de rendre certaines erreurs plus coûteuses que d'autres. Ainsi, toujours prédire la classe majoritaire ne donnera plus un bon score.

En combinaison avec ces méthodes de gestion du déséquilibre, nous avons testé la performance des algorithmes suivants :

- dummyClassifier comme baseline
- Régression logistique
- Random Forest
- Light GBM

Les résultats de ces combinaisons sont consultables dans la figure suivante :



Nous avons sélectionné la régression logistique qui a montré des performances supérieures. En ce qui concerne le pré-traitement, notre choix s'est porté sur class-weight + sous-échantillonage + sur-échantillonage. Bien que similaire sur le score du F-beta(beta = 2), le score de rappel était supérieur pour ce pré-traitement (0.66).
Les détails des métriques de performances seront explicités dans la partie III.

2. Optimisation des hyper-paramètres

Une fois le modèle de régression logistique sélectionné nous avons optimisé son ratio de régularisation entre norme L1 et L2 par pas de 0.5 %.
L'optimisation des hyper-paramètres s'est effectuée par validation croisée. Les données on étés coupées en 5 sur un total de 4 fois.

Le meilleur score (f-beta2 = 0.365) fut obtenu avec un ratio de régularisation L1 = 0.736

III. Performances

1. Métrique d’évaluation

Lors de nos classifications, deux erreurs sont possibles : prédire à tort qu’un client sera en défaut de paiement (faux positif) et attribuer un crédit à un client qui sera effectivement en défaut (faux négatif).
La proportion de faux positifs est mesurée par la **précision**

$$\mathbf{precision} = \frac{\mathbf{vrais\ positifs}}{\mathbf{vrais\ positifs} + \mathbf{faux\ positifs}}$$

La proportion de faux négatifs est mesurée par le **recall**

$$\mathbf{recall} = \frac{\mathbf{vrais\ positifs}}{\mathbf{vrais\ positifs} + \mathbf{faux\ negatifs}}$$

Pour maximiser ces deux mesures, nous les combinons avec un score **f-beta**

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Où β est un paramètre nous permettant de donner plus d’importance à l’une des deux mesures.

Nous choisissons pour notre modèle un β = 2 pour privilégier le rappel sur la précision.
Cette décision est motivée par le fait qu’un prêt non remboursé est plus grave qu’un prêt non attribué.

2. fonction coût métier

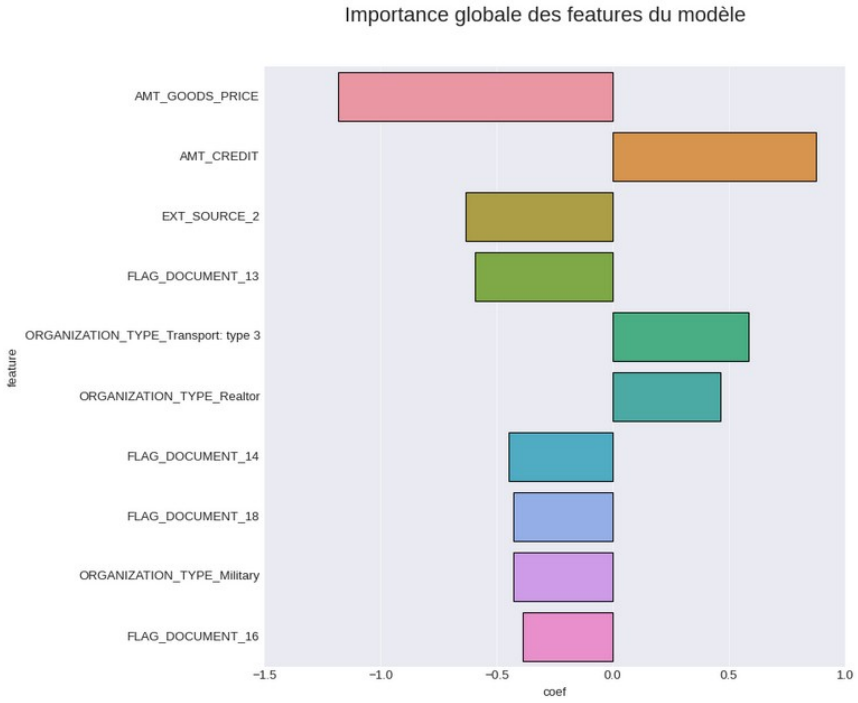
Perdre la somme d’un prêt n’est pas désirable. Refuser chaque crédit par peur de ce scénario n’est pas non plus viable pour l’entreprise. Nous avons recherché sur un échantillon la probabilité de risque permettant de maximiser les gains (cf. annexe):
D’après notre estimation, **en dessous de 68 % de risque de défaut, le crédit devrait être accepté**. Au delà de ce seuil, l’argent perdu en prêts non remboursés entame les bénéfices des intérêts perçus sur une plus large population (plus à risque).

3. Algorithme d’optimisation

Après avoir sélectionné l’algorithme et la méthode de traitement du déséquilibre la plus performante sur un sous échantillon des données, la combinaison la plus performante a été entrainée sur 80 % des données. Le reste a été utilisé pour mesurer la performance. La décision finale d’attribution de crédit a été étalonnée sur un seuil conclu par une étude métier.

IV. Interprétation du modèle

A un niveau global, nous avons profité de la facilité d'interprétation qu'offre un modèle linéaire. En observant les coefficients nous avons pu déterminer les facteurs les plus déterminants pour le modèle.



Un défaut de paiement est encodé par un 1 et un remboursement par un 0.
Pour l'interprétation de ce graphique cela veut dire que les features négatifs sont désirables pour obtenir un prêt alors que ceux positifs vont peser contre une attribution de crédit.

L'interprétation locale des prédictions a été mise en place par la librairie SHAP. (cf annexe)
L'intérêt d'utiliser cette méthode est sa nature « model-agnostic ». Si à l'avenir de nouveaux modèles plus performants viennent remplacer le modèle actuel, le framework d'interprétation n'aura pas à changer.

V. Limites et perspectives d’amélioration

Le script kaggle de pré-traitement des données a très peu été modifié.
Construire un pipeline reliant pré-traitement des données et performances des modèles pourrait aider à optimiser l’exploitation des informations.

La recherche du meilleur modèle et l’optimisation des hyperparamètres n’a pu se faire que sur un petite portion des données.
Plutôt que de chercher par force brute, l’emploi d’optimisation bayésienne pourrait permettre de tester sur plus de données une plus grande dimensionnalité d’hyper-paramètres.

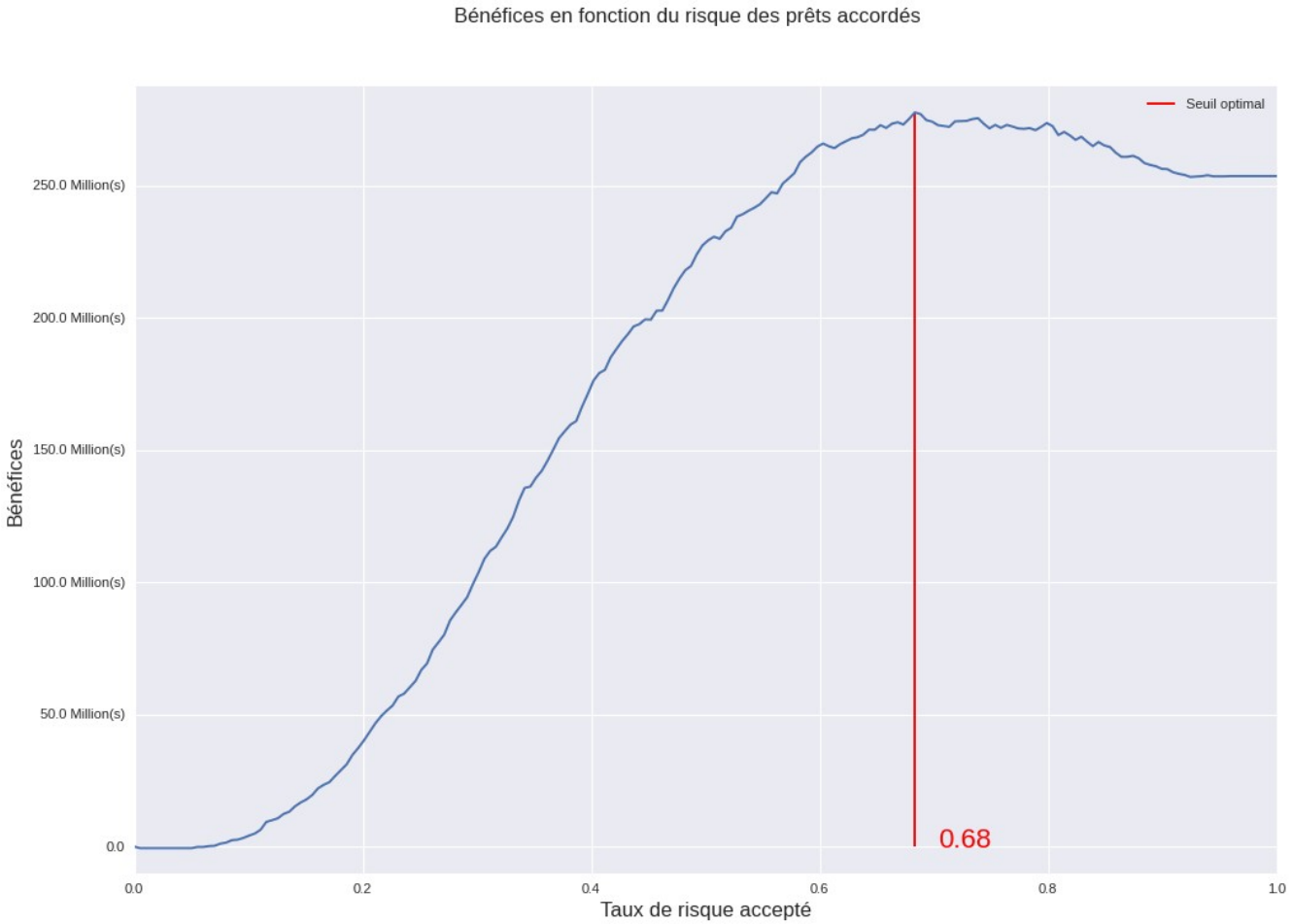
Le seuil de risque optimal calculé s’est basé sur un échantillon significativement différent du reste des données.
Un modèle de génération de données synthétiques aurait pu être construit sur la base des relations entre les différentes variables et le taux d’intérêt.
Le seuil optimal quant à lui ne devrait pas rester fixe mais s’actualiser en fonction des données de gain de l’entreprise.

L’interprétation globale est basée sur les coefficients d’un modèle linéaire. Ces derniers sont utiles pour comprendre le fonctionnement du modèle mais à prendre avec précaution pour interpréter la réalité. Par exemple notre modèle considère l’age comme un facteur positif, plus de défauts de paiements sont observés chez les jeunes. D’après son interprétation, il vaudrait mieux accorder un prêt à une personne de 105 ans qu’à une personne de 40 ans.

L’interprétation locale basée sur SHAP n’est pas non plus parfaite. Il est possible que certaines valeurs se retrouvent anormalement importantes à cause des corrélations présentes dans le dataset. Utiliser des méthodes plus poussées comme treeSHAP et les compléter par des analyses sur les données elles mêmes peuvent aider à une interprétation plus fidèle à la réalité.

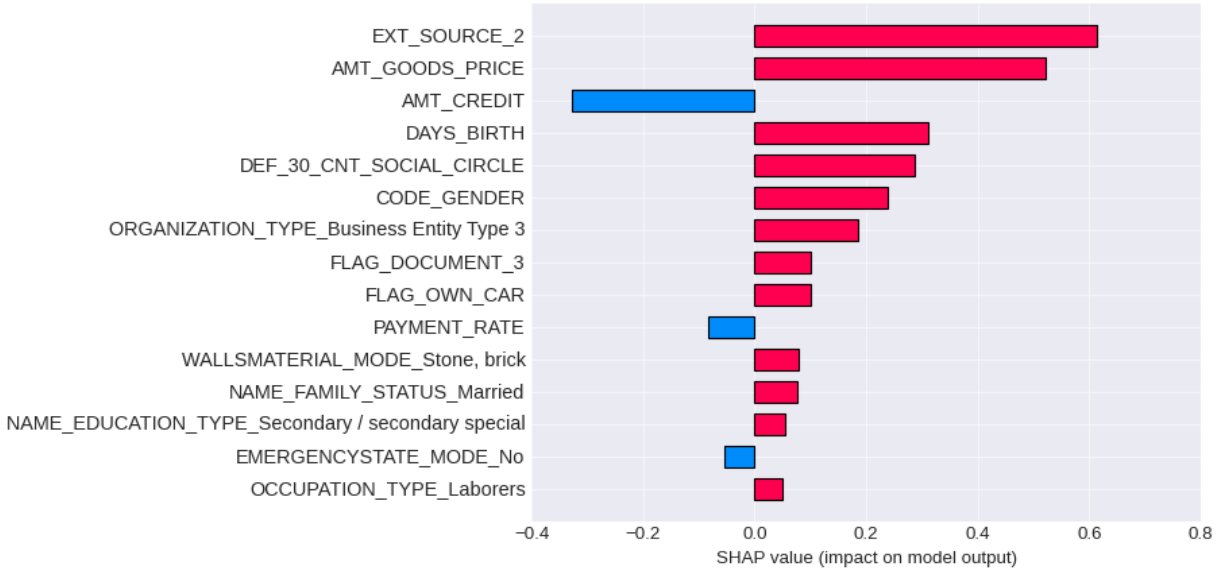
Annexes :

1. Recherche d’un seuil de risque optimal

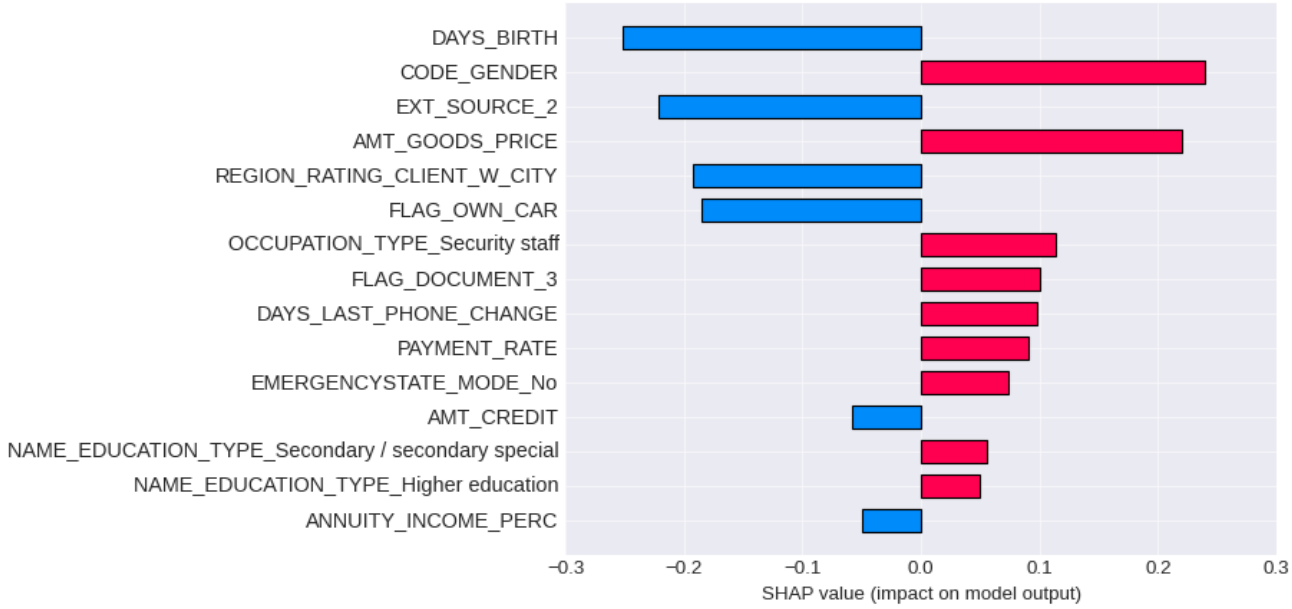


2. Interprétation locale d’un client refusé

Le bleu représente les variables qui facilitent la décision de prêt et le rouge celles qui vont à l’encontre.



3. interprétation locale d’un client accepté



4. Matrice de confusion

La matrice de confusion représente les différents issues des prédictions de notre modèle.

Ici les faux négatifs sont coloriés en rouge afin de mettre en évidence l'issue la plus grave des prédictions du modèle : Accorder un prêt à un client qui ne le remboursera pas.

	Client prédits en défaut	**Clients prédits sans défaut**
Clients réellement en défaut	Vrais positif	Faux négatifs
Clients sans défaut	Faux positif	Vrais négatifs