Note méthodologique

Projet 7 : Implémentez un modèle de scoring

ZOUHEIR HMIDI

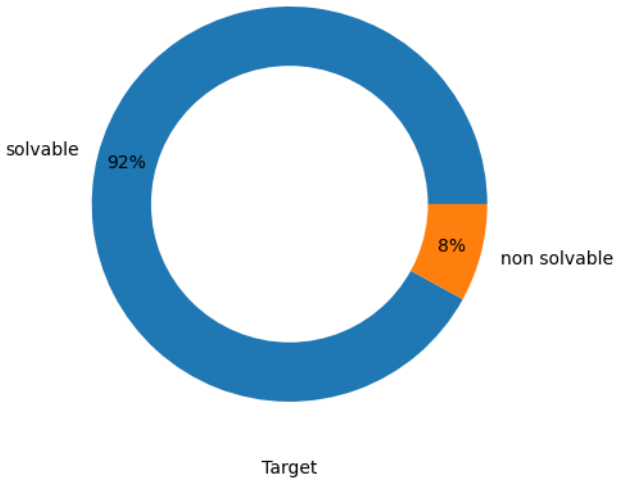
PARCOURS DATA SCIENTIST

1. Contexte

La société financière "Prêt à dépenser" a l'intention de créer un modèle de notation pour évaluer la probabilité de non-remboursement de crédits à la consommation, spécialement conçu pour les individus ayant peu ou pas d'historique de prêt. Cette initiative vise à prendre des décisions éclairées sur l'octroi de prêts aux clients potentiels en se basant sur une multitude de sources de données, y compris des données comportementales et des informations provenant d'autres institutions financières.
En parallèle, l'entreprise développera un tableau de bord interactif qui permettra d'interpréter les prévisions générées par le modèle de notation. Il offrira également une visualisation des données générales relatives à chaque client, ainsi que des comparaisons avec des clients similaires.

2. Données

Pour élaborer cet outil, la société Home Credit, qui se spécialise dans la fourniture de crédits aux personnes non bancarisées, met à disposition une base de données. Cette base de données contient des informations générales sur chaque client, ainsi que des données supplémentaires concernant les prêts contractés par ces clients.
Un premier constat révèle un déséquilibre entre les clients qui n'ont pas connu de défaut de paiement des échéances de crédit et ceux qui ont rencontré des difficultés à respecter leurs obligations financières. Cette disparité est illustrée dans le graphique ci-dessous.



3. Traitement des données

Avant d'utiliser les données à des fins de prévision et pour en tirer le meilleur parti, le prétraitement des données est une étape cruciale qui intervient après l'exploration des données. Il englobe plusieurs étapes visant à nettoyer les données brutes, qui sont souvent peu fiables et incomplètes, pouvant ainsi conduire à des résultats biaisés.

Cependant, de nombreux clients présentent un grand nombre de données manquantes. Dans de tels cas, où les données sont très limitées (moins de 50% de complétion), il est préférable de supprimer ces clients considérés comme peu fiables. Cependant, il est essentiel de veiller à conserver suffisamment d'informations pour maintenir la qualité des données.

Après cette réduction de données, il est nécessaire d'imputer les données manquantes restantes, effectuer une standardisation des données pour finir par un encodage numérique des variables catégorielles.

4. Méthodologie d'entraînement des modèles

4-1 Séparation des données

Le processus d'entraînement d'un modèle requiert une phase de séparation des données, se situant entre la phase de prétraitement des données et la phase de modélisation. Cette séparation vise à créer deux ensembles distincts : un ensemble d'apprentissage (échantillon d'entraînement) représentant 70% des données, et un ensemble de test représentant les 30% restants.Haut du formulaire

4-2 Correction du déséquilibre entre les classes

Comme mentionné précédemment, dans les données il y a un fort déséquilibre entre les clients solvables et non solvables. Ce déséquilibre va avoir un impact sur la performance de l'algorithme et les prédictions seront faussées car le modèle attribuera beaucoup plus fréquemment la classe la plus représentée aux clients. Pour pallier à ce problème de déséquilibrage, la technique SMOTE est appliquée sur les données d'entraînement :

SMOTE est une méthode qui va créer de nouvelles données pour la classe sous-représentée à partir des données existantes (et donc de la variété) pour que chaque classe ait le même nombre de données que la classe surreprésentée à l'origine.

4-3 Choix des algorithmes

Un modèle de prédiction prend en entrée des données en entrée et donne une prédiction en sortie.
Pour déterminer l'algorithme optimal de classification adapté à la problématique, 4 algorithmes ont été testés :
- un algorithme de classification fictif comme baseline (Dummy Classifier)
- un algorithme de LGBMClassifier
- un algorithme AdaBoostClassifier
- un algorithme GradientBoostingClassifier

5. Évaluation et choix de l'algorithme

5-1 Métriques utilisées pour le choix de l'algorithme

Pour analyser les résultats produits et ainsi démontrer l'efficacité de chaque modèle, les métriques suivantes ont été privilégiées :
- L'AUC (l'aire sous la courbe ROC) est la mesure de la capacité d'un classificateur à distinguer les classes et correspond au résumé de la courbe ROC. Plus l'AUC est élevée, plus la performance du modèle à distinguer les classes positives et négatives est bonne.
- L'Accuracy (ratio entre le nombre total de prédictions correctes et le nombre total de prédictions), compte tenu du déséquilibre des classes
Par conséquent, le modèle choisi devra maximiser l'AUC et la fonction de revenu net normalisée : cela correspond au modèle Light Gradient Boosting Classifier .

5-2 Optimisation des hyper-paramètres

Il est nécessaire d'effectuer une recherche des hyperparamètres par GriSearchCV afin d'optimiser le modèle en fonction des données. L'échantillon d'apprentissage est l'élément central qui permet au modèle de régler ses prédictions grâce à une validation croisée. Au cours de cette étape, cet ensemble d'apprentissage est subdivisé en plusieurs sous-ensembles sur lesquels les algorithmes s'entraînent pour ajuster le modèle.

5-3 Fonction coût : métrique métier

Une fonction de coût personnalisée a été développée pour l'entreprise Prêt à dépenser, afin de créer un nouveau score métier visant à sanctionner de manière plus rigoureuse les Faux Négatifs. Les Faux Négatifs correspondent aux clients qui ont effectivement fait défaut mais qui ont été incorrectement prédits comme étant sans risque. Cette erreur a des conséquences financières préjudiciables pour l'entreprise, car elle entraîne une perte de capital.

Nous basons donc notre score sur la matrice de confusion

		Classe prédite	
		Positive (1)	Négative (0)
Classe réelle	Positive (1)	TP	FN
	Négative (0)	FP	TN

TN : Refuser le prêt à un client incapable de le rembourser ne génère ni pertes ni gains.
FN : Accorder un crédit à un client qui ne peut pas le rembourser par la suite entraîne des pertes.
TP : Accorder un crédit à un client qui le remboursera par la suite représente un gain.
FP : Refuser le prêt à un client qui aurait pu le rembourser équivaut à une perte de client et, par conséquent, d'argent.

Nous définissons notre fonction coût de la manière suivante :

$gain\_total = (TN * coeff\_tn + FP * coeff\_fp + FN * coeff\_fn + TP * coeff\_tp)$
$gain = (gain\_total - gain\_min)(gain\_max - gain\_min)$

Avec :
$gain\_\text{min} = (TN+FP) * coff\_fp + (TP+FN) * coeff\_fn$
$gain\_\text{max} = (TN+FP) * coeff\_tn + (TP+FN) * coeff\_tp$

Coefficients arbitraires possibles en respectant la métrique métier :

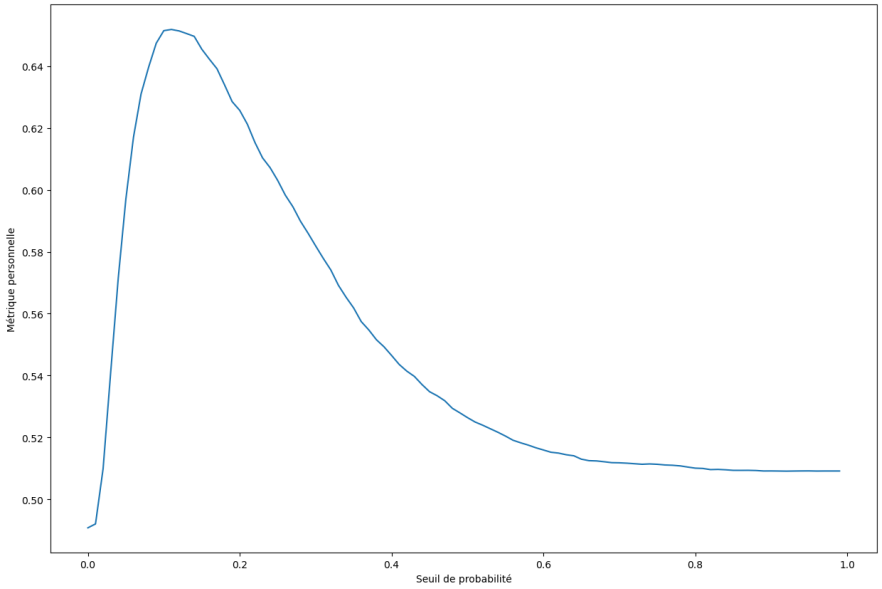
- 0 pour les vrais négatifs
- -1 pour les faux positifs
- +1 pour les vrais positifs
- -10 pour les faux négatifs

Cette fonction coût a été implémentée afin de pénaliser l'impact des erreurs sur la décision d'octroi de crédit.

Dans cet exemple, le seuil qui maximise la fonction est aux alentours de 0,11 et donc qu'un crédit sera accordé si la probabilité qu'un client fasse défaut est inférieure à 11%.

5-3 Ajustement du seuil de probabilité

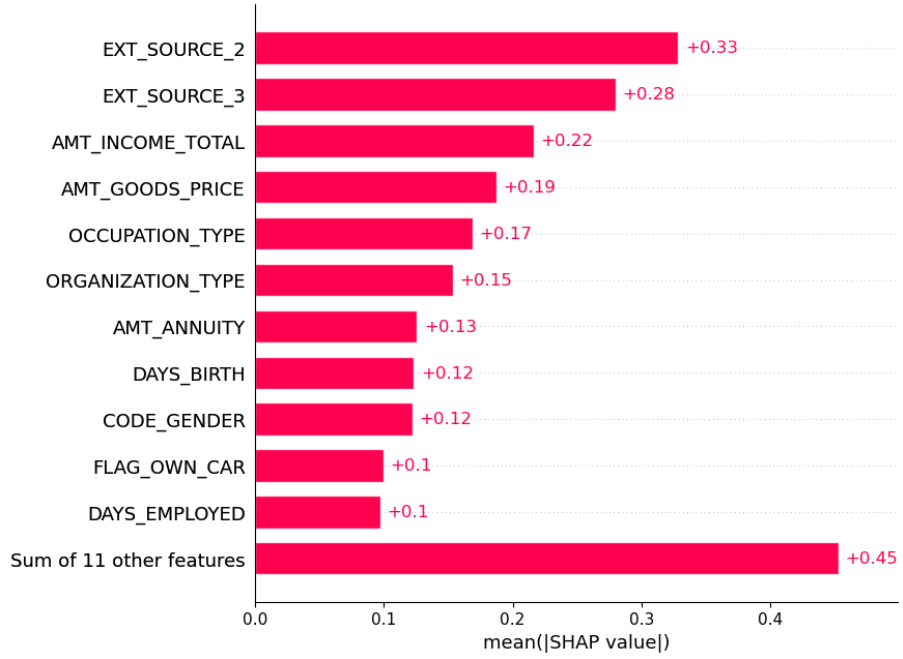
Pour optimiser les gains de la banque et établir le seuil de probabilité à partir duquel un client est considéré comme solvable ou non, il est nécessaire de maximiser la métrique personnelle "gain". Habituellement, les modèles choisissent un seuil de probabilité de 0,5 pour classer les individus. Cependant, en examinant la courbe de la métrique gain (avec un coefficient de -10), il apparaît que le seuil de 0,5 n'est pas l'optimum pour maximiser les bénéfices de la banque.

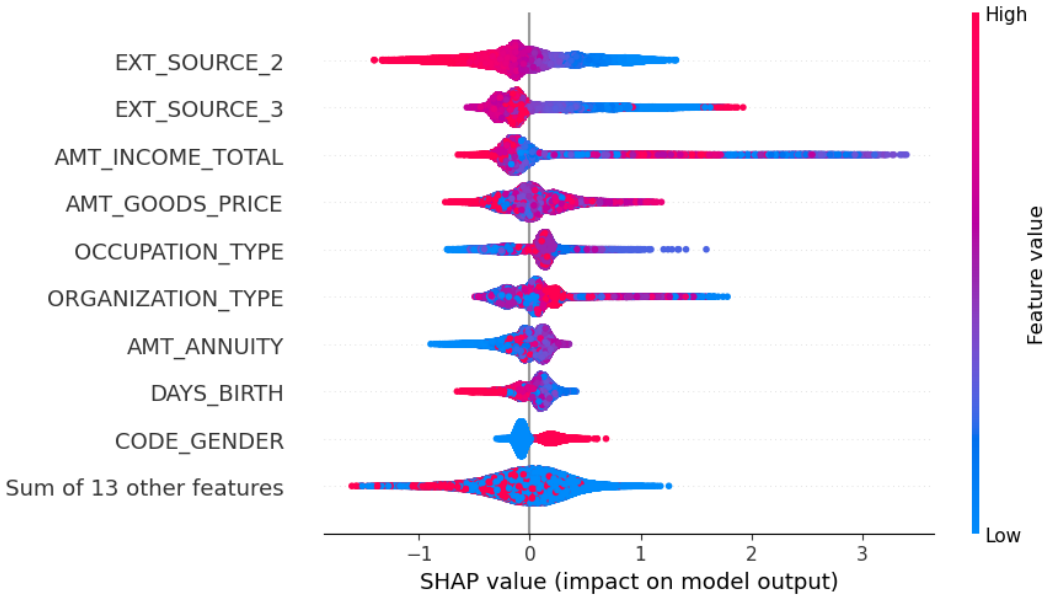


Dans cet exemple, le seuil qui maximise la fonction est aux alentours de 0,11 et donc qu'un crédit sera accordé si la probabilité qu'un client fasse défaut est inférieure à 11%.

6. Interprétabilité de l'importance des variables

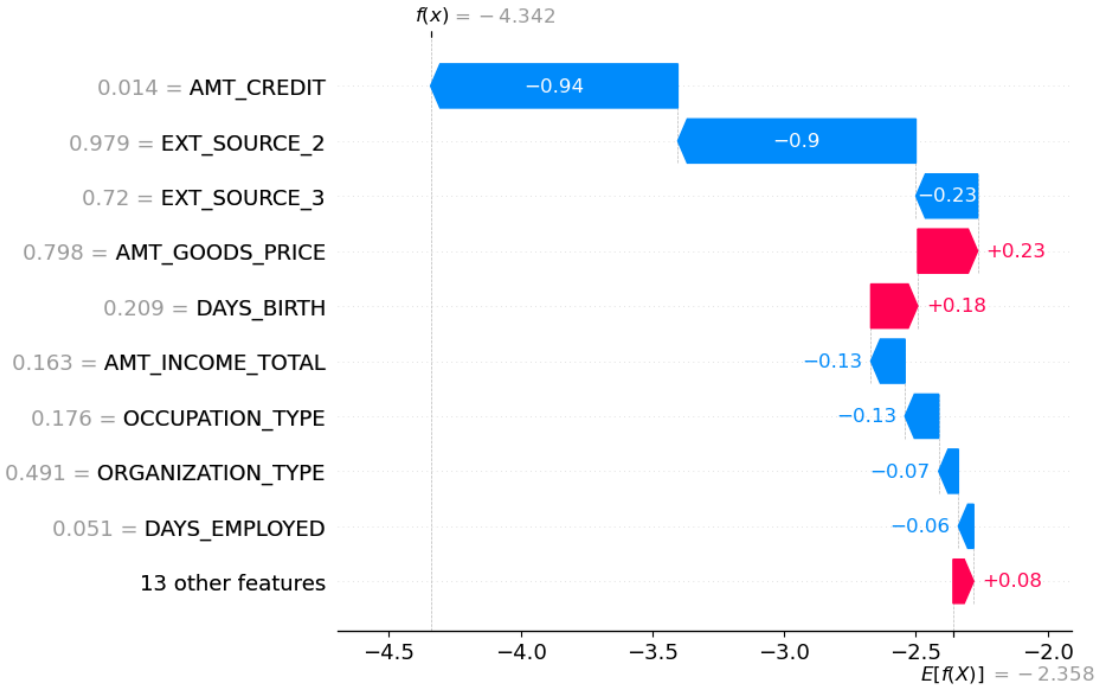
6-1 Interprétabilité globale





En se basant sur cette illustration, on peut conclure que les caractéristiques les plus influentes dans la prédiction de l'approbation d'un prêt sont les sources extérieures 2 et 3, qui sont des scores normalisés générés à partir de données externes. Ensuite, viennent le revenu total du client et la valeur du bien détenu par le client.

6-2 Interprétabilité locale



7. Limites et axes d’améliorations

La modélisation a été effectuée sur la base d'une métrique personnelle créée pour répondre au mieux au besoin de gain d'argent d'une banque. Les coefficients de cette métrique ont été choisis arbitrairement selon le bon sens. L'axe principal d'amélioration serait donc de définir plus précisément ces coefficients associés à chaque combinaison classe prédite/classe réelle car le modèle déterminé ici ne sera pas obligatoirement le meilleur.

Compte tenu des contraintes sur les ressources système, je n’ai pas pu affiner le réglage des hyper-paramètres et utiliser le méthode d’échantillonnage des données pour la correction du déséquilibrage des données. Ainsi, la performance du modèle final est peut-être sous-estimée.

Enfin, le Dashboard interactif pourra être amélioré pour répondre au mieux aux attentes et besoins des conseillers clients.