Implémentation d'un modèle de scoring

Florence GUILLOU, Février 2021
Parcours data scientist OpenClassrooms – Projet 7

1- Problématique et données

1- Problématique

La société financière « Prêt à dépenser » propose des crédits à la consommation à destination de personnes ayant peu ou pas d'historique de prêt.

Afin d'étayer la décision d'accorder ou non un prêt, l'entreprise souhaite développer un modèle de scoring de la probabilité de défaut de paiement.
Par ailleurs, dans un souci de transparence concernant les décisions d'octroi de crédit, « Prêt à dépenser » décide de développer un dashboard interactif permettant aux chargés de clientèle d'explorer facilement les informations personnelles des clients, de les leur présenter et d'expliquer de manière transparente les décisions concernant les accords de crédit.

Cette note vise à décrire la méthodologie d'entraînement du modèle mise en œuvre, la fonction coût métier proposée, l'algorithme d'optimisation et la métrique d'évaluation, l'interprétabilité du modèle dont les résultats serviront à expliciter les décisions aux clients. Enfin, les limites et améliorations identifiées seront finalement discutées.

2- Données

Le jeu de données mis à disposition recense plus de 300000 prêts décrits par 121 caractéristiques (âge, sexe, revenus, logement, emploi, enfants, informations concernant le crédit) réparties dans 7 tables différentes. Les données sont anonymisées.

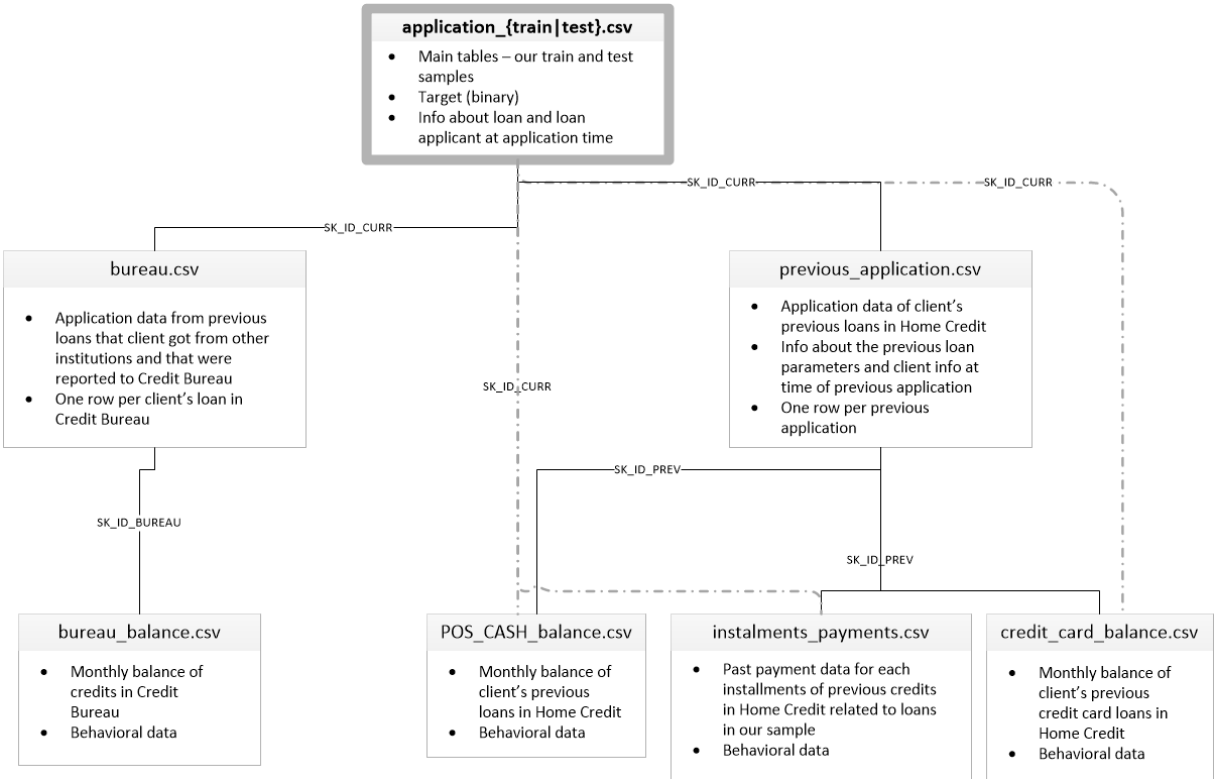
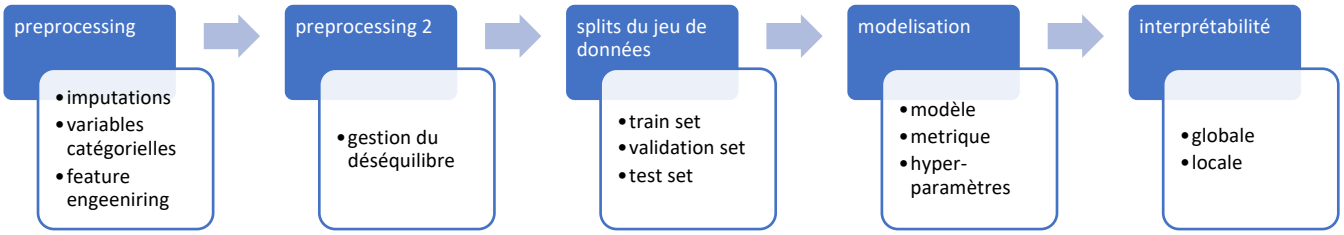


Figure 1: Description des données

2- Stratégie

L'objectif est d'entraîner un modèle dont la finalité est de prédire la probabilité de remboursement d'un prêt par un client. Chaque client est décrit par plusieurs dizaines de caractéristiques le concernant personnellement et d'autres caractérisant le prêt. Ces caractéristiques sont utilisées pour entrainer un modèle de classification de manière supervisée grâce à la labellisation des données permettant de distinguer les clients ayant présenté un défaut de paiement des autres.
Les différentes étapes sont recensées dans le schéma ci-dessous et décrites dans les sections suivantes.



3- Preprocessing et gestion du déséquilibre

Un kernel Kaggle[1] a été sélectionné[2] pour faciliter le preprocessing des données.
Différents prétraitements ont été réalisés en vue d'alimenter l'algorithme :

- Ainsi, Les valeurs manquantes ont été imputées par les valeurs médianes.
- Pour les variables catégorielles avec seulement 2 catégories, le label encoding a été utilisé, lequel génère une colonne avec 0 ou 1 selon la catégorie. Dans le cas des variables catégorielles à plus de 2 catégories, le One Hot Encoding a été préféré. Il génère autant de colonnes que de catégories chacune codée 0 ou 1.
- Concernant le feature engineering, des variables composites orientées métier ont été créées :

- le pourcentage du montant du crédit par rapport au revenu d'un client (CREDIT_INCOME_PERCENT),
- le pourcentage de la rente de prêt par rapport au revenu d'un client (ANNUITY_INCOME_PERCENT),
- la durée du paiement en mois (puisque la rente est le montant mensuel dû) (CREDIT_TERM),
- le pourcentage des jours employés par rapport à l'âge du client (DAYS_EMPLOYED_PERCENT).

Finalement, le jeu de données exploité dans le cadre de cette étude est composé de 240 variables.

Gestion du déséquilibre du jeu de données

1- Inégale représentation des classes

S'agissant de la cible, la valeur 0 est attribuée aux clients ne présentant pas de retard de paiement de leur prêt, la valeur 1 aux clients présentant des défauts de paiement.
La classe 0 est sur-représentée par rapport à la classe 1 puisque cette dernière ne représente que 9% des données cibles. Dans une telle situation de déséquilibre, et sans traitement spécifique, les algorithmes auront tendance à prédire systématiquement la classe majoritaire, c'est à dire sans défaut de paiement.
Or, d'un point de vue financier, accorder un crédit à un client n'étant pas en mesure de le rembourser sera plus lourd que de ne pas accorder de crédit à un client capable de le rembourser sans retard.
Il est donc essentiel de traiter ce déséquilibre.

[1] Will Koehrsen - Boston, Massachussets
https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

[2] Le notebook a été sélectionné selon le langage (Python) et le nombre de votes.

2- Traitement

Différents traitements ont été mis en œuvre pour gérer le déséquilibre des données :

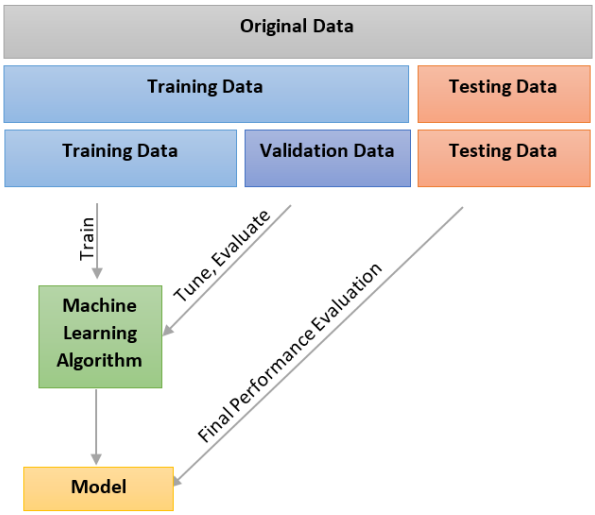
- le sous-échantillonnage, ou undersampling, consiste à abaisser le nombre de prêt sans défaut de paiement au nombre de prêts avec, et ainsi égaliser les classes en termes d'effectifs. L'inconvénient réside dans la perte d'informations liée à la suppression d'une importante quantité d'informations.
- le sur-échantillonage, ou oversampling, vise au contraire à créer des données synthétiques de sorte qu'ici encore les deux catégories présentent des effectifs similaires. Cette méthode a cependant tendance à produire des données similaires à celles existantes.
- le cost sensitive learning consiste à redéfinir la fonction de coût du modèle en tenant compte des poids. Scale-pos-weight est un hyper-paramètre de XGBoost Classifier, implémenté ici. Il est utilisé pour redimensionner les erreurs faites par le modèle pendant l'entraînement sur la classe minoritaire et encourage le modèle à les sur-corriger.
  Sample weight, paramètre de la méthode fit, permet de spécifier un poids différent pour chacun des cas et ainsi tenir compte du montant du crédit.

Enfin, une fois l'algorithme entraîné le seuil de probabilité maximisant le score est recherché.
Ainsi, si ce seuil est fixé à 0.36 et que les probabilités associées estimées sont respectivement de 0.62 et 0.38 pour les classes 0 et 1, le prêt sera classifié 1.

4- Approche de la modélisation

1- Train/ Validation/ Test split

Le jeu de données est d'abord partagé en un jeu d'entraînement et un jeu de test. Le jeu d'entraînement est ensuite lui-même partagé en un jeu d'entraînement et un jeu de validation destiné à ajuster les paramètres du modèle. Le jeu de test sert à l'évaluation final du modèle.



- Pour chacun des traitements de gestion du déséquilibre, undersampling, oversampling, scale-pos-weight et sample weight, le modèle est entraîné sur le jeu d'entraînement avec une gamme de valeurs pour différents hyper-paramètres (n_estimators, learning rate, max_depth).

Le score est finalement évalué sur le jeu de validation.

2- Choix du score et fonction coût métier

1- Score

La matrice de confusion permet de mesurer la qualité de la classification. Dans le cas de la problématique présente, elle peut être représentée ainsi :

	Prédits sans défaut (0)	**Prédits en défaut (1)**
Réellement sans défaut (0)	Vrais Négatifs	Faux Positifs
Réellement en défaut (1)	**Faux Négatifs**	Vrais Positifs

D'un point de vue métier, il s'agit de minimiser les faux négatifs, c'est à dire les personnes prédites comme étant sans défaut de paiement et qui pourtant présentent un défaut de paiement.
D'un point de vue métrique, cela reviendrait à optimiser le rappel qui correspond au rapport du nombre de personnes prédites en défaut sur le nombre de personnes total réellement en défaut :

$$rappel = \frac{vrais\ positifs}{vrais\ positifs + faux\ négatifs}$$

Toutefois, en maximisant uniquement le rappel, le risque est d'augmenter significativement les faux positifs, c'est à dire le nombre de personnes sans défaut de paiement prédites comme étant en défaut.
Il convient donc également de rechercher à maximiser la précision, c'est à dire le rapport entre le nombre de personnes prédites en défaut sur le nombre de personnes totales prédites en défaut :

$$précision = \frac{vrais\ positifs}{vrais\ positifs + faux\ positifs}$$

Finalement, la problématique revient à trouver le meilleur compromis entre ces 2 métriques.
Le F1-score est une moyenne pondérée de la précision et du rappel.

$$F1\ score\ = \frac{2 * precision * rappel}{precision + rappel}$$

Il est retenu comme métrique pour optimiser le modèle. Il s'agira de le maximiser.

2- Fonction coût métier

Si on considère que l'entreprise présente un gain de 30% du montant du crédit, sous forme d'intérêts, lorsqu'un prêt est accordé et remboursé (vrais négatifs), un coût d'opportunité de 30% pour un crédit refusé, correspondant à ces mêmes intérêts manqués, alors qu'il aurait été remboursé (faux positifs), et un coût de 80% du montant du crédit pour un crédit accordé mais partiellement remboursé (à hauteur de 80% en moyenne) (faux négatifs), la fonction de coût métier simplifiée ci-dessous peut être proposée :

$$Gain = 0.3 * vrais\ négatifs - 0.3 * faux\ positifs - 0.8 * faux\ négatifs$$

Ce gain est également à maximiser.

3- Visualisation des scores

Le modèle XGBoost Classifier a été utilisé pour traiter cette problématique de classification supervisée.

Les courbes ci-dessous recensent les F1-scores et gains obtenus selon la méthode de gestion du déséquilibre du jeu de données et l'ajustement du seuil de probabilité.

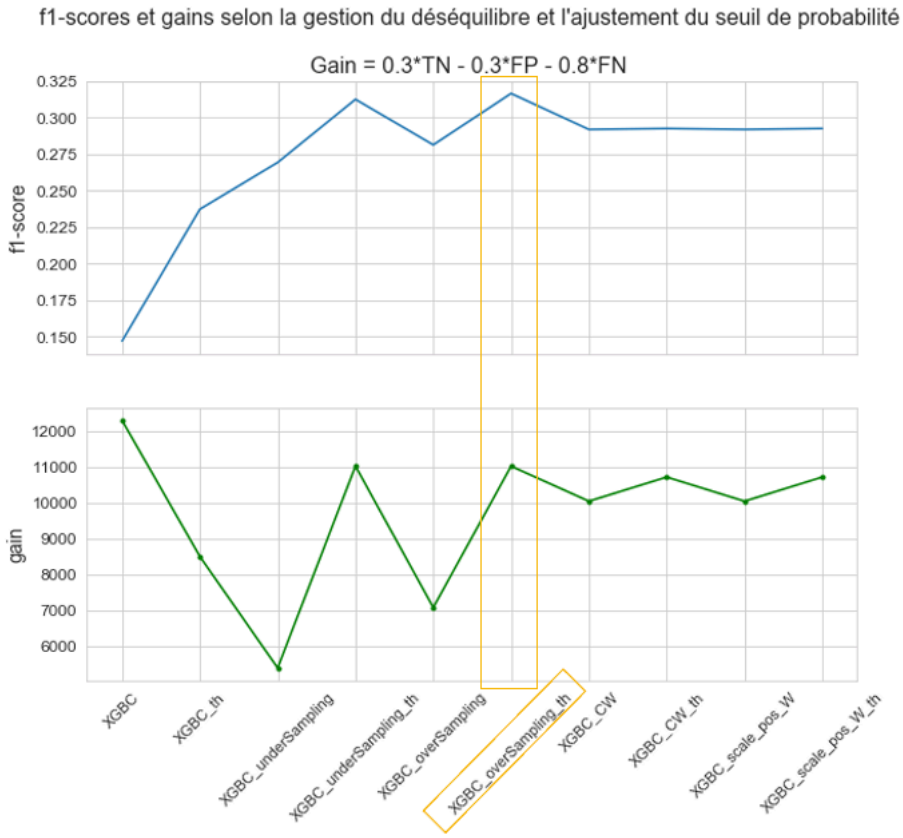


Figure 2: f1-scores et gains selon la gestion du déséquilibre du jeu de données (classe majoritaire versus classe minoritaire) et l'ajustement du seuil de probabilité

La méthode de gestion du déséquilibre retenue est l'oversampling avec un ajustement du seuil de probabilité pour la prédiction des classes.

5- Interprétabilité

Un des objectifs de cette étude est de produire un tableau de bord à destination des chargés de clientèle, lesquels devront être capables d'interpréter le score fournit par l'algorithme et de justifier leur décision d'octroi de crédit auprès du client.
Par ailleurs, compte tenu des enjeux liés à un accord de crédit pour un individu, il convient de mettre en œuvre un outil d'interprétabilité tel que le framework SHAP.

SHAP (SHapley Additive exPlanations TreeExplainer) permet d'accéder à une approximation des valeurs de Shapley dans le cas d'arbres de décision ou d'ensemble d'arbres.
Les valeurs de Shapley calculent l'importance d'une variable en comparant ce qu'un modèle prédit avec et sans cette variable. Cela se fait dans tous les ordres possibles de sorte que les fonctionnalités soient comparées équitablement. Cette approche est inspirée de la théorie des jeux.

Tree Explainer permet de calculer l'importance globale des variables (interprétabilité globale) ainsi que les effets des variables pour chaque exemple du jeu de données (interprétabilité locale).

1- Interprétabilité globale

Le summary plot ci-dessous permet de voir les variables les plus importantes et l'amplitude de leur impact sur le modèle.

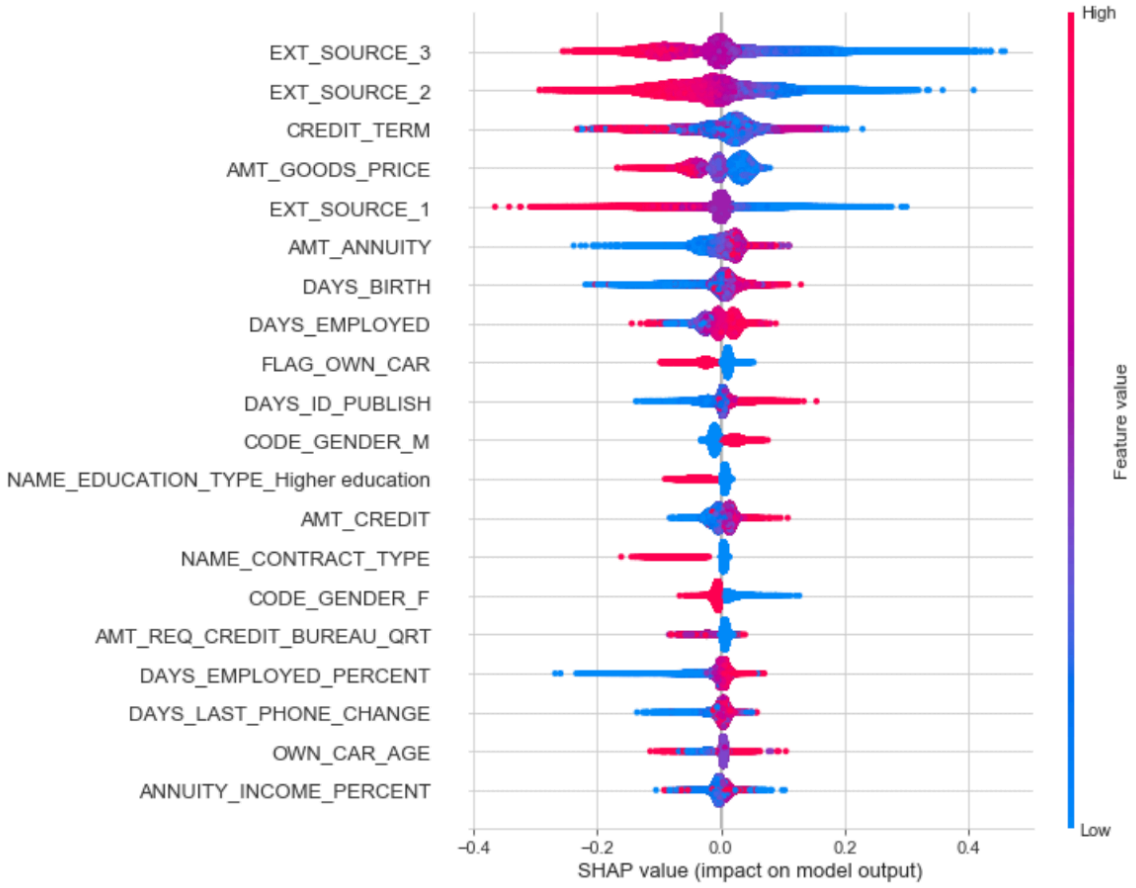


Figure 3: Diagramme de l'importance des variables - interprétabilité globale

Caractéristiques	Description
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_EDUCATION_TYPE	Level of highest education the client achieved
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
OWN_CAR_AGE	Age of client's car
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
AMT_ANNUITY	Annuity of the Credit Bureau credit
NAME_CONTRACT_TYPE	Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application
AMT_ANNUITY	Annuity of previous application
AMT_CREDIT	Final credit amount on the previous application. This differs from AMT_APPLICATION in a way that the AMT_APPLICATION is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT_CREDIT
AMT_GOODS_PRICE	Goods price of good that client asked for (if applicable) on the previous application

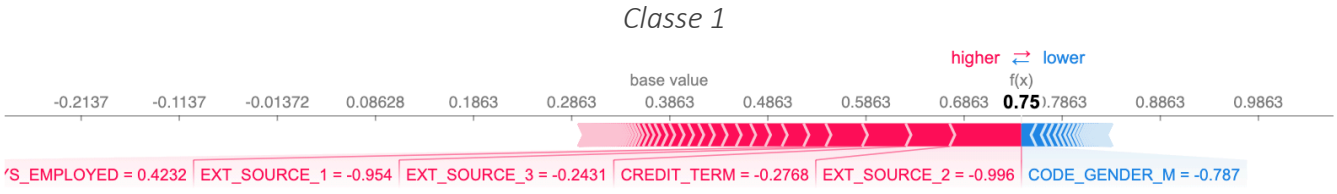
Figure 4: Description des paramètres les plus influents sur le score de probabilité de faillite des clients.

Ce diagramme de l'importance des variables a été élaboré à partir de tous les points des données du jeu d'entrainement. Il nous renseigne sur différents points :
- l'importance de la fonctionnalité : les variables sont classées par ordre décroissant. Ainsi la variable 'EXT_SOURCE_3' apparaît comme étant la plus influente.
- l'impact : l'emplacement horizontal indique si l'effet de cette valeur est associé à une prédiction supérieure ou inférieure.
- la valeur d'origine : la couleur indique si cette variable est élevée (en rouge) ou basse (en bleu) pour cette observation. Ainsi, un niveau élevé de EXT_SOURCE_3 (couleur rouge) a un impact négatif.

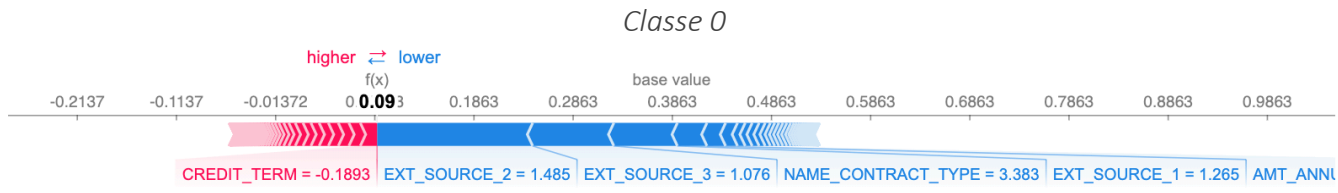
Globalement, des valeurs faibles des 3 paramètres 'EXT_SOURCE' caractérisent des clients ayant des difficultés de remboursement.
Un niveau d'annuités (AMT_ANNUITY) élevé est plutôt favorable aux difficultés de paiement. Elles pourraient être abaissées par un allongement de la durée du crédit par exemple pour limiter le risque.
S'agissant de l'âge (DAYS_BIRTH), la variable est exprimée en jours par rapport à la date de contraction du prêt. Un client jeune aura donc une valeur plus élevée pour cette variable qu'un client plus âgé. Ce sont donc plutôt des clients jeunes qui rencontrent le plus de difficultés de paiement.
Les 3 variables EXT_SOURCE_1, EXT_SOURCE_2 et EXT_SOURCE_3 apparaissent comme étant les plus influentes. Selon la documentation, ces fonctionnalités représentent un « score normalisé à partir d'une source de données externe ». Il pourrait s'agir d'une cote de crédit cumulative établie à l'aide de multiples sources de données. Leur signification devra être précisée avec l'expertise métier pour une meilleur compréhension.

2- Interprétabilité locale

Les force plots ci-dessous représentent l'influence des paramètres, intensité et effets, sur la prédiction de la probabilité de faillite d'un client. Deux cas sont ici présentés : un client risquant de présenter un défaut de paiement et un autre qui devrait rembourser régulièrement son prêt.
Les valeurs des variables EXT_SOURCE_1, EXT_SOURCE_2, et EXT_SOURCE_3 sont également présentées pour chacun des cas. Le client de la classe 0 présente des valeurs plus élevées pour chacune de ces 3 variables.



	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
132241	0.348161	0.260721	0.427657



	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3
192547	0.665102	0.776166	0.676993

Figure 5: diagramme des valeurs SHAP individuelles - interprétabilité locale (en haut: avec risque de défaut de paiement, en bas: peu de risque de défaut de paiement)

6- Limites et améliorations

Le déséquilibre des classes constitue un obstacle pour l'optimisation de la prédiction. Il conviendrait donc, dans la mesure du possible, d'alimenter la classe minoritaire.

Concernant le feature engineering, des variables composites pourraient être créées ou ajustées avec l'expertise métier afin d'améliorer la qualité des données et donc de l'information fournie à l'algorithme.

Le pre-processing pourrait être amélioré avec par exemple une réflexion sur les imputations, éventuellement une révision des catégories des variables catégorielles. D'autres données, externes, pourraient venir compléter le dataset.

La fonction de coût métier ici proposée devra également être discutée et ajustée avec l'expertise métier.
Les variables EXT_SOURCE, a priori les plus influentes, devront être décrites.

Une attention particulière devra être accordée à l'éthique pour ce projet.
L'analyse du jeu de données montre que les personnes présentant un défaut de crédit sont davantage des personnes jeunes, il conviendra de ne pas stigmatiser cette catégorie de clients.
Un autre point d'attention identifié est l'utilisation du sexe de l'emprunteur. En effet, selon les jeux de données utilisées pour entraîner les algorithmes, des biais peuvent naître notamment liés au sexe de la personne. On peut citer l'exemple du programme de l'Apple card dont l'algorithme a attribué une limite de crédit 20 fois plus élevée à l'homme d'un couple marié qu'à sa femme déclarant pourtant leurs impôts conjointement.
Cet événement met par ailleurs en évidence l'importance de la transparence et de l'interprétabilité.
Les conseillers clientèle devront veiller à ne pas fonder leur décision d'octroi de crédit uniquement sur la probabilité fournie par le modèle développé.
Rappelons qu'en Europe, le Règlement Général sur la Protection des Données spécifie dans son article 22 :
« La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire ».

Finalement, si des améliorations peuvent être apportées, il va s'agir de trouver le meilleur compromis entre performance de l'algorithme et interprétabilité.