NOTE

METHDOLOGIQUE

**Projet : Implémentez un modèle de scoring**

https://github.com/babi7777/scoring-model-credit-risk

**Badreddine HANNACHE Aout 2023**

* OpenClassrooms **-**

***Résumé***

*Le projet vise à mettre en place un outil de "scoring crédit" pour évaluer la probabilité de remboursement d'un client et classifier les demandes de crédit. En utilisant des données variées, l'entreprise souhaite développer un modèle de classification automatique.*

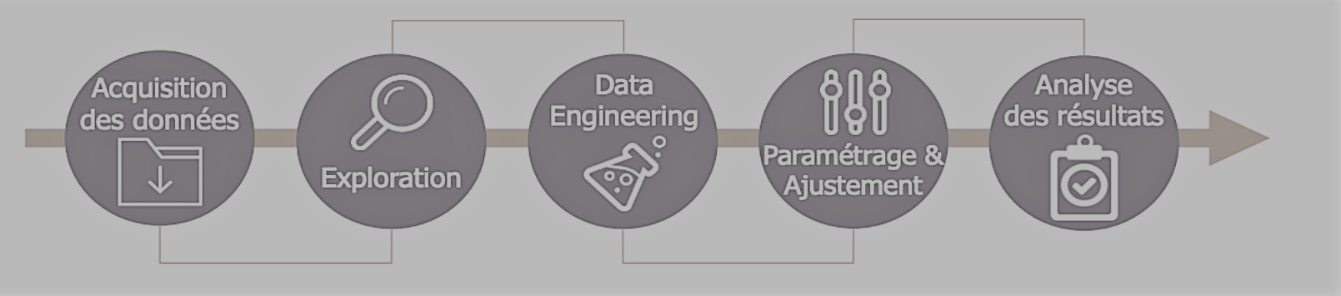
*De plus, afin de répondre à la demande croissante de transparence des clients, un Dashboard interactif sera créé pour expliquer les décisions d'octroi de crédit et permettre l'accès facile aux informations personnelles des clients.*

*Enfin, le modèle de scoring sera déployé via une API, accompagné du Dashboard interactif.*

1. **Méthodologie d'entraînement du modèle**

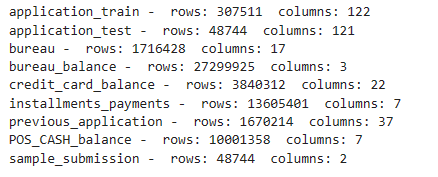
Dans cette section, nous décrirons la méthodologie adoptée pour entraîner notre modèle de scoring de crédit. L'objectif principal est de développer un algorithme de classification capable de prédire la probabilité de remboursement d'un crédit par un client. Pour ce faire, nous nous appuyons sur des sources de données variées, incluant des données comportementales et des informations provenant d'autres institutions financières.

Un projet de Machine Learning évolue à travers différentes étapes successives. La mise en œuvre suit généralement une approche courante, se déployant selon le schéma suivant :



**1.1. Collecte des données**

La première étape de notre méthodologie consiste à collecter les données nécessaires à la construction de notre modèle. Ces données comprennent des informations sur les clients tels que leur historique de crédit, leurs revenus, leur emploi, et d'autres caractéristiques pertinentes.



Datasets utilisés dans le projet

Nous utilisons également des données provenant d'autres sources financières pour enrichir notre jeu de données et améliorer les performances du modèle.

**1.2. Exploration et prétraitement des données**

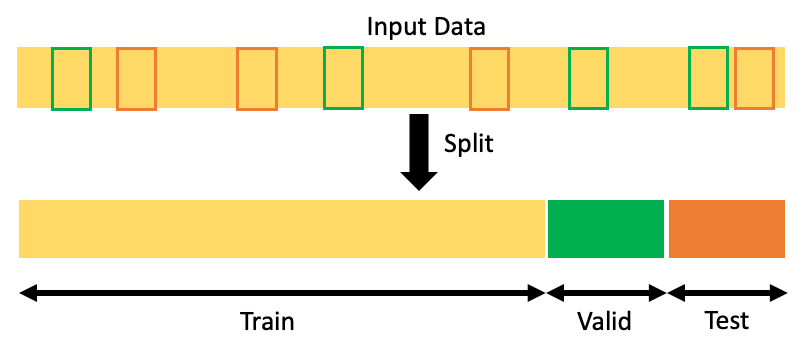
Avant de procéder à l'entraînement du modèle, nous réalisons une étape d'exploration et de prétraitement des données. Cela comprend la gestion des valeurs manquantes, le nettoyage des données aberrantes, et la transformation des variables catégorielles en variables numériques.

**1.3. Sélection et évaluation du modèle**

Dans cette étape, nous sélectionnons un algorithme de classification approprié pour notre problème. Nous avons opté pour un algorithme de gradient boosting LightGBM

Pour évaluer les performances de notre modèle, nous utilisons des métriques pertinentes telles que la précision, le rappel, la F1-score, et la courbe ROC.

**1.4. Entraînement et réglage du modèle**

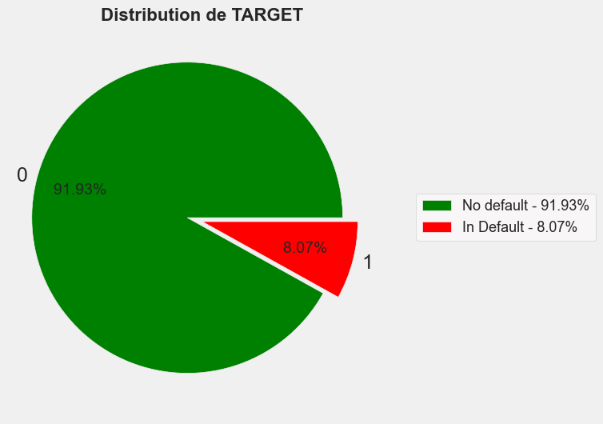
Une fois le modèle sélectionné, nous divisons notre jeu de données principalement en ensembles d'entraînement et de test.

**Figure 1**: Découpage des données

Nous utilisons l'ensemble d'entraînement pour ajuster les paramètres du modèle et l'ensemble de test pour évaluer sa performance. Les hyperparametres optimaux ont été trouvés par ‘’Bayesian optimization’’ sur <https://www.kaggle.com/tilii7/olivier-lightgbm-parameters-by-bayesian-opt/code>. Elles ont été préférés pour leur efficacité prouvée dans la recherche productive d'hyperparametres, épargnant ainsi des efforts et du temps comparé à une recherche manuelle.

Pour prévenir tout surajustement, une validation croisée avec StratifiedKFold de sklearn a été employée.

1. **Le traitement du déséquilibre des classes**

Le défi en *Machine Learning* réside dans la classification binaire. Il est crucial d'assurer une distribution équilibrée des classes 0 et 1 de chaque côté.

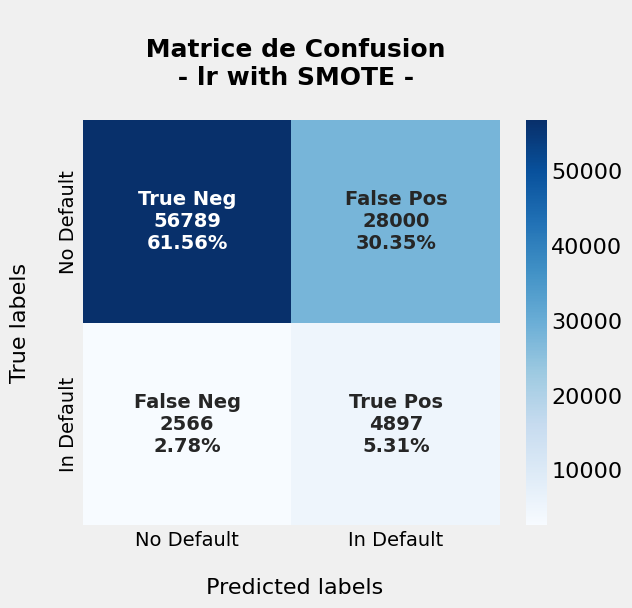
L'exploration des données a révélé un déséquilibre significatif de précision entre la classe 0 et la classe 1. L'échantillon de données présente un déséquilibre, avec une prédominance de 92% pour la classe 0 et seulement 8% pour la classe 1. Pour corriger cette disparité, nous avons utilisé une technique de sur-échantillonnage (Oversampling) qui ajuste la distribution des classes de manière à atteindre une répartition plus équilibrée.

**Figure 2** : Déséquilibre des classes

1. **Modélisation de la probabilité de non-paiement par le client**

Le modèle de notation qui permettra de prédire la probabilité de défaillance d'un client a été mis en œuvre en utilisant une approche de "classification".

La classification implique la détermination des catégories auxquelles appartiennent de nouveaux objets en se basant sur des exemples déjà connus. Dans le cadre spécifique de ce projet, la classification est binaire, avec une sortie divisée en deux classes : l'approbation du crédit ou le rejet du crédit. À la différence d'un problème multi-classes, où une variable cible peut comprendre plusieurs catégories distinctes.

À la différence des modèles statistiques traditionnels qui nécessitent la vérification d'hypothèses sur la distribution des données, le Machine Learning possède d'impressionnantes capacités prédictives grâce à son approche non paramétrique et à la capacité de ses algorithmes à apprendre directement des données sans préjugés a priori. Parmi ces avancées, nous avons évalué en particulier l'algorithme de Gradient Boosting "LightGBM". Parallèlement, nous avons établi une référence en utilisant la Régression Logistique, un modèle statistique classique.

**Figure 3 :** Résultats Régression Logistique :

Matrice de confusion

1. **Approche de Gestion des Pertes Financières via une Fonction de Coût personnalisée**

L'entreprise "Prêt à Porter" fait face au défi des défauts de paiement de la part de ses clients, cherchant à éviter les pertes financières (telles que les frais de recouvrement et les pertes).

**Terminologie :**

Faux Positifs (FP) : Ces cas se produisent lorsque la prédiction est favorable, mais la réalité est défavorable, pouvant entraîner une opportunité manquée si un crédit est injustement rejeté malgré sa capacité à être remboursé.

Faux Négatifs (FN) : Dans ce cas, la prédiction est négative, mais la réalité est positive. Cela entraîne une perte réelle lorsque des crédits accordés aux clients sont acceptés, mais finissent par ne pas être remboursés comme prévu.

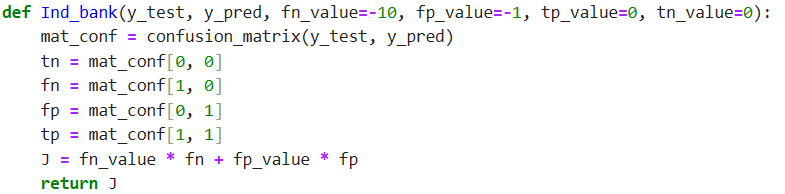
Vrais Positifs (TP) : Ce sont les situations où la prédiction d'acceptation est correcte, et en conséquence, le crédit accordé au client est remboursé avec succès.

Vrais Négatifs (TN) : Dans cette catégorie, les prédictions de refus sont correctes, et ainsi, les crédits clients sont rejetés de manière appropriée pour éviter les risques de non-remboursement.

Dans le cadre de notre projet visant à évaluer la solvabilité des clients, nous avons élaboré une approche méthodologique centrée sur l'application de la fonction de coût personnalisée "Ind\_bank". Cette fonction a été conçue pour quantifier les pertes financières potentielles liées aux décisions de classification, en mettant l'accent sur les erreurs de classification et leurs conséquences financières.

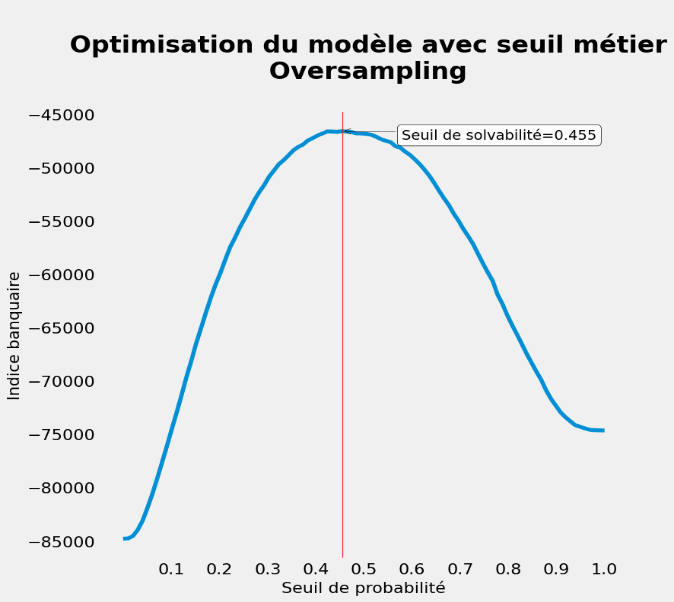
**Fonction de Coût Métier Personnalisée**

Notre approche repose sur la fonction de coût "Ind\_bank", que nous avons développée pour mesurer l'impact financier de nos prédictions de solvabilité. Cette fonction attribue des valeurs spécifiques aux Faux Positifs (FP) et aux Faux Négatifs (FN), représentant respectivement les cas où un client non solvable est classé comme solvable et vice versa. La formule de "Ind\_bank" est la suivante :



Cette fonction nous permet d'évaluer les pertes financières potentielles découlant de nos décisions de classification.

**Optimisation des Seuils de Solvabilité à l'Aide de la Métrique "Indice Bancaire" :**

****Pour trouver les seuils de solvabilité optimaux qui minimisent l'Indice Bancaire, nous avons mis en place une approche d'optimisation. Nous avons développé la fonction "gain\_seuil" qui évalue l'Indice Bancaire à différents seuils de solvabilité. Cette fonction parcourt une gamme de seuils et ajuste les prédictions en conséquence. Nous utilisons cette fonction pour visualiser la relation entre le seuil de solvabilité et l'Indice Bancaire, ce qui nous permet de choisir le seuil optimal pour notre modèle.

**Figure 4 :** Optimisation du modele avec seuil métier

(fn\_value = 10 \* fp\_value)

1. **Un tableau de synthèse des résultats**

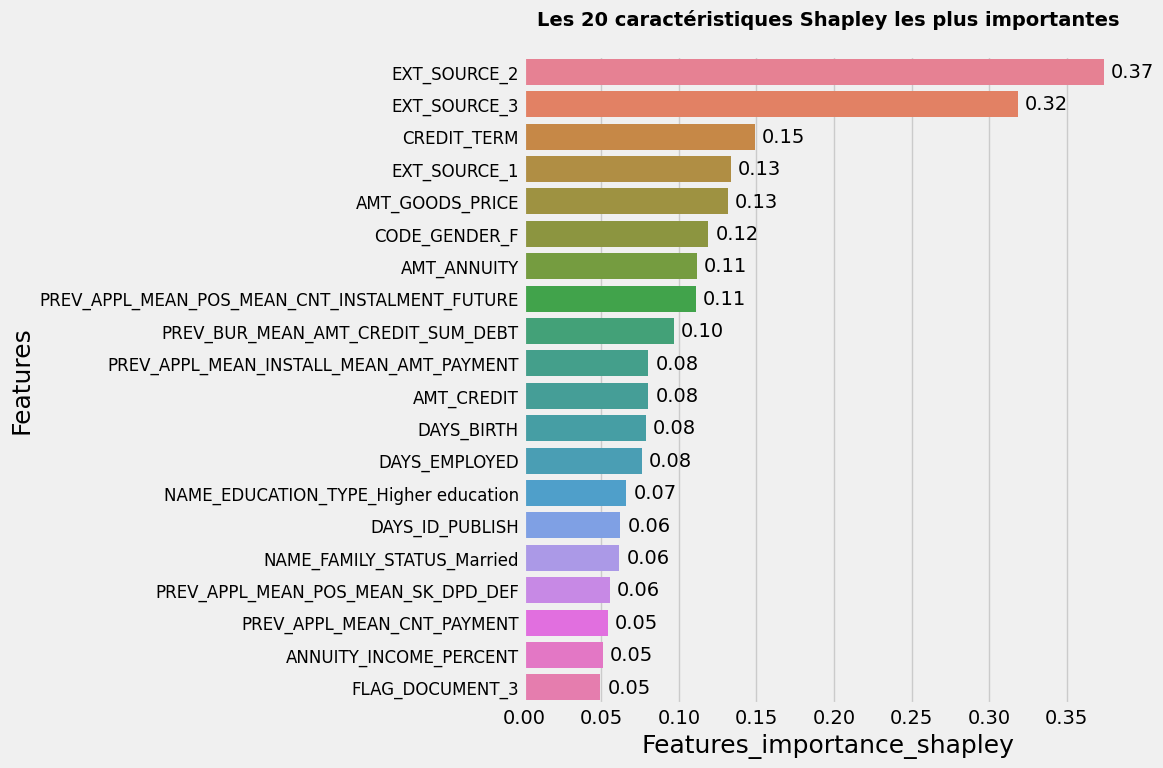
|  |  |  |
| --- | --- | --- |
| **Métriques** | **Score initial** | **Score optimal** |
| **Precision/Recall/**  **Accuracy/F1 Score/**  **ROC AUC** | **C:\Users\hbcon\Desktop\download.png** | **C:\Users\hbcon\Desktop\download.png** |
| **Matrices de confusion** | **C:\Users\hbcon\Desktop\download.png** | **C:\Users\hbcon\Desktop\download.png** |

1. **Interprétabilité du modèle**

Pour assurer la transparence et la compréhension des décisions de notre modèle de notation de crédit, nous avons adopté une approche axée sur l'interprétabilité. Nous comprenons que nos utilisateurs incluent des professionnels du crédit qui ne sont pas nécessairement des experts en science des données. Notre objectif est de fournir des explications claires et accessibles sur les facteurs qui influencent les décisions du modèle.

**Caractéristiques Shapley les Plus Importantes :**

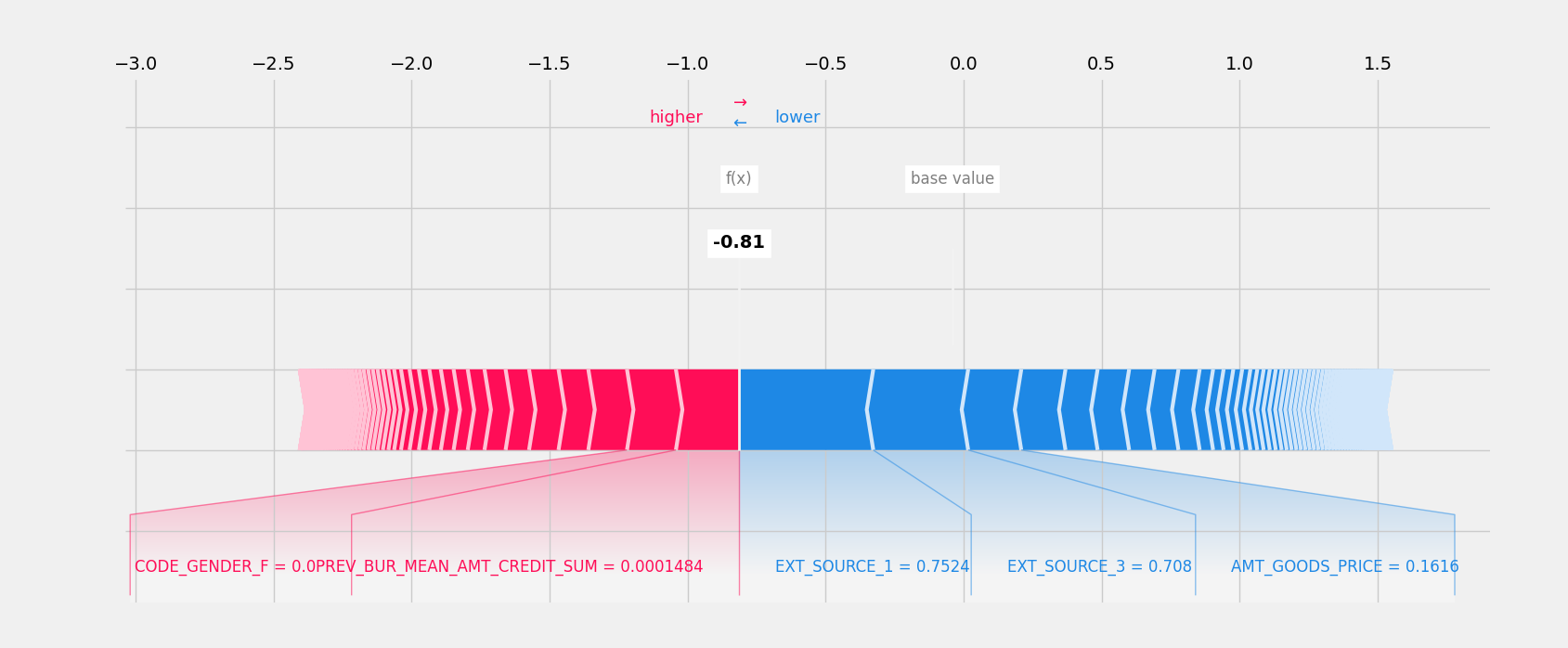
Nous avons identifié les 20 caractéristiques Shapley les plus importantes qui ont le plus grand impact sur les prédictions de notre modèle. Ces caractéristiques ont été sélectionnées en utilisant la méthode Shapley, qui attribue à chaque caractéristique une valeur d'importance relative

****

**Figure 5 :** Caractéristiques Shapley les plus importantes

**Graphique Shapley Force Plot :**

Pour donner un aperçu visuel des explications individuelles des prédictions, nous utilisons les graphiques Shapley Force Plot. Ces graphiques illustrent comment chaque caractéristique a contribué à la prédiction pour un échantillon spécifique.



L'interprétabilité de notre modèle est un pilier essentiel de notre approche de notation de crédit. En mettant en avant les caractéristiques Shapley les plus importantes et en utilisant des graphiques Shapley Force Plot, nous offrons une explication transparente et visuelle de la manière dont notre modèle évalue la solvabilité des clients. Cette approche assure une meilleure compréhension de nos décisions de notation de crédit, tant pour nos clients que pour notre équipe interne.

1. **Les limites et les améliorations possibles**

**Limites Actuelles :**

* Hypothèses Sous-jacentes : Nos résultats sont basés sur les hypothèses que nous avons formulées lors du développement du modèle. Il est essentiel de noter que ces hypothèses pourraient ne pas être validées par les équipes métier. Nous reconnaissons ainsi la nécessité d'une collaboration continue pour ajuster et valider les hypothèses.
* Prédiction Optimale des Coûts : Bien que notre modèle intègre une fonction de coût spécifique pour équilibrer les pertes, il est important de noter que les coûts réels peuvent varier en fonction du contexte économique et des décisions métier. Notre modèle vise à optimiser les coûts en fonction des données disponibles, mais il peut y avoir des variations imprévues.

**Perspectives d'Amélioration :**

# Raffinement des Paramètres : Nous continuons à explorer la possibilité d'affiner davantage les paramètres de notre modèle pour obtenir des performances optimales. Cela peut impliquer des ajustements dans la manière dont nous traitons les caractéristiques, calculons les poids ou fixons les seuils de solvabilité.

# Validation des Résultats : Pour garantir la pertinence de nos résultats, nous nous efforçons de valider nos prédictions en collaborant étroitement avec les équipes métier. L'obtention de leurs retours et de leur expertise nous permettra d'ajuster et d'affiner le modèle en conséquence.

# Exploration de Nouvelles Caractéristiques : Nous envisageons d'étendre notre modèle en incluant de nouvelles caractéristiques pertinentes, ce qui pourrait améliorer la performance prédictive. L'exploration de sources de données additionnelles peut également contribuer à la précision de nos prédictions.

# Évaluation des Risques Systémiques : Nous reconnaissons la possibilité d'étendre notre modèle pour tenir compte des risques systémiques potentiels qui pourraient influencer les prédictions de solvabilité.

# Amélioration de l'Interprétabilité : Nous travaillons continuellement sur des méthodes plus avancées pour rendre les décisions du modèle plus compréhensibles et explicables. Cela peut inclure des techniques telles que l'analyse de sensibilité et la visualisation des contributions des caractéristiques.

1. **L’analyse du Data Drift**