# Evaluation comparative de modèles d'apprentissage automatique : cas de l'octroi de crédit bancaire

**Par Serigne Dame LO**

Titulaire d'un master en **Economie et Finance Quantitatives, Data scientist**

Serignedame2.lo@ucad.edu.sn

**e-Portfolio :** https://github.com/losdame

***Résumé***

Nous cherchions à effectuer une évaluation comparative entre modèles d'apprentissage automatique en évaluant leur performance par le biais de métriques pertinentes dans le cadre de l'octroi des crédits bancaires. Les modèles utilisés sont l'arbre de décision, la forêt aléatoire, le gradiant boosting, la catégorie boosting et la régression logistique. Le choix de ces modèles est porté par leur performance, leur capacité d'interprétation, leur robustesse et leur popularité dans la détermination du risque. La base de données utilisée dans cet exercice est constituée de 20000 demandes de prêts bancaires des clients d'une banque avec 36 caractéristiques dont l'âge, le score de risque, le taux d'intérêt, le statut de propriété, le niveau d'étude ou le fait qu'un prêt a été approuvé ou non. Sur ces données historiques de demandes de prêts, 76,1 % n'ont pas été approuvées contre 23,9%. Les données sont subdivisées d'une manière aléatoire en données d'entrainement de test et de validation. Les résultats de cette étude montrent que la catégorie boosting est plus performante que les autres en termes de précision (96%), de rappel (97%), d'AUC (98%) et d'exactitude (98%). À partir du modèle champion, nous avons constaté que les variables telles que le score de risque, le ratio total dettes revenus, le taux d'intérêt, le score de crédit, le revenu mensuel et le montant du prêt sont les facteurs les plus influents dans l'octroi de prêts bancaires.

***Mots clés*** : Risque de Crédit, Arbre de Décision, Forêt Aléatoire, Catégorie Boost, gradient boosting. Régression Logistique, Machine Learning

**Abstract**

We aimed to conduct a comparative evaluation of machine learning models by assessing their performance through relevant metrics in the context of bank credit granting. The models used are decision trees, random forests, gradient boosting, CatBoost, and logistic regression. The choice of these models is driven by their performance, interpretability, robustness, and popularity in risk assessment. The database used in this exercise consists of 20,000 loan applications from bank customers, with 36 characteristics including age, risk score, interest rate, property status, education level, and whether a loan was approved or not. Among these historical loan application data, 76.1% were not approved compared to 23.9%. The data is randomly divided into training, testing, and validation datasets. The results of this study show that CatBoost outperforms the others in terms of precision (96%), recall (97%), AUC (98%), and accuracy (98%). From the champion model, we found that variables such as the risk score, total debt-to-income ratio, interest rate, credit score, monthly income, and loan amount are the most influential factors in the granting of bank loans.

*Keywords:* Credit Risk, Decision Tree, Random Forest, CatBoost, Gradient Boosting, Logistic Regression

## 1. Introduction

L'analyse du crédit est un élément central dans le domaine financier, permettant aux institutions et aux prêteurs de mieux comprendre les risques liés à l'octroi de crédits aux divers emprunteurs. Elle est essentielle pour la prise de décision financière car ayant un impact direct sur la rentabilité et la stabilité des institutions financières, ainsi que sur la confiance des investisseurs. Dans ce cadre, les modèles de scoring de crédit se présentent comme des outils indispensables pour évaluer le risque de défaut d'un emprunteur potentiel. En s'appuyant sur des variables tant financières que non financières, ces modèles cherchent à prédire la probabilité de défaut, offrant ainsi aux prêteurs des informations importantes pour éclairer leurs décisions en matière de crédit. Aujourd'hui, face à un environnement financier de plus en plus complexe, marqué par des innovations technologies, par des fluctuations économiques et les changements dans les comportements des acteurs, différentes méthodes alternatives basées sur l'apprentissage automatique sont utilisées dans le but de faciliter la prise de décision mais aussi de la rendre plus efficace. Notre objectif est d'effectuer une évaluation comparative entre modèles d'apprentissage automatique en évaluant leur performance en fonction des métriques pertinentes dans le cadre de l'octroi des crédits. Pour cela, nous avons choisi des modèles tels que la régression logistique, l'arbre de décision, la forêt aléatoire, le gradiant boosting et la catégorie boosting. Le choix de ces modèles est motivé par leur performance, leur robustesse et aussi leur popularité.

## 2. Revue de la littérature

Ces dernières années, plusieurs études ont été menées dans l'application d'une variété de méthodes d'apprentissage automatique dans l'évaluation du risque de crédit, telles que les arbres de décision, les réseaux neuronaux, les machines à vecteurs de support et les algorithmes d'intégration. Dans cette section, nous allons présenter quelques-uns de ces différentes recherches.

Bahnsen et al. (2016) ont constaté que la régression logistique traditionnelle n'était pas aussi performante que les précédentes lorsque les différentes variables caractéristiques présentaient des relations non linéaires compliquées. Bien que le modèle de régression logistique ne soit peut-être pas aussi bon que le modèle d'apprentissage automatique en termes de précision de prédiction, il présente un fort avantage en termes d'interprétabilité et de stabilité des variables. Par conséquent, certains chercheurs ont amélioré la régression logistique et l'ont appliquée à la prédiction du comportement par défaut de l'emprunteur.

En 1985, pour la première fois, Makowski a utilisé l'approche de l'arbre de décision dans le domaine de l'évaluation du crédit personnel. Carter a prouvé en 1987 que l'approche de l'arbre de décision a une précision de classification élevée dans le domaine de l'évaluation du crédit personnel. Wang Maoguang et al. (2016) ont également constaté que le modèle d'arbre de décision présente les avantages de l'adaptabilité, de la haute précision et d'une forte interprétabilité pour expliquer les raisons du défaut de prêt.

Jiang Cuiqing et al. (2017) ont analysé la relation entre différentes variables d'information douce et les défauts de paiement, et ensuite, en combinant le filtrage des informations douces et dures, en introduisant un algorithme de forêt aléatoire pour construire un modèle de prédiction par défaut intégrant des informations douces et en combinant les données réelles pour une analyse empirique. Les résultats montrent que l'inclusion d'informations douces précieuses dans le modèle de prédiction par défaut des emprunts peut améliorer la précision de la prédiction.

En 2023, Y. Ruicheng et al. ont utilisé un modèle de CatBoost pour déterminer une nouvelle approche de notation de crédit. Les auteurs ont jumelé l'algorithme de recherche Sparrow(SSA) et CatBoost pour améliorer la précision de la classification et de la prédiction pour la notation de crédit. Les résultats expérimentaux montrent que le modèle SSA-CatBoost a une précision idéale dans la classification et la prédiction de la notation de crédit comparé aux autres modèles d'apprentissage automatique.

Pour faire face au problème du scoring de crédit dans le domaine du crédit à la consommation sur le marché chinois, Miaojun Bai et al. (2022) ont fourni une solution avancée en quantifiant la probabilité de survie. Le modèle de survie proposé est basé sur l'algorithme de gradient boosting qui optimise simultanément la probabilité de survie pour chaque période de temps, ce qui peut réduire considérablement l'erreur globale. Ensuite, des ensembles de données de marché réels à grande échelle ont été utilisés pour tester l'applicabilité du modèle. Les résultats montrent que le modèle GBST surpasse les modèles de survie existants mesurés par l'indice de concordance (indice C) ainsi que par l'aire sous la courbe caractéristique de fonctionnement du récepteur (AUC) de chaque période de temps.

## 3. Présentation des modèles

L'apprentissage automatique est une méthode d'analyse de données et une partie de l'intelligence artificielle, qui repose sur la capacité d'un système à apprendre à partir de données antérieures, à identifier des modèles ou des distributions d'ensembles de données et à prendre des décisions. D'une manière générale, il s'agit d'un système qui automatise la création de modèles analytiques avec une intervention humaine minimale.

### 3.1 Decision Tree

Il s'agit d'une méthode d'apprentissage supervisé non paramétrique et d'un outil d'aide à la décision utilisé pour la régression et la classification. En fonction du jeu de données, il peut identifier et dériver une courbe sinusoïdale à l'aide d'une série de règles de décision de type si-alors-sinon (if-then-else). Le modèle est plus ajusté si l'arbre est plus profond, car cela signifie qu'il contient des règles de décision plus complexes.

Cet algorithme utilise la représentation arborescente pour résoudre des problèmes, avec des nœuds internes représentant des attributs et des nœuds feuilles représentant des étiquettes de classe. On commence à la racine de l'arbre, en séparant les échantillons en groupes homogènes selon les diviseurs les plus importants des variables d'entrée. Ce processus est

répété jusqu'à ce que tous les nœuds feuilles soient trouvés dans les branches de l'arbre. Par conséquent, il est facile à visualiser, comprendre et interpréter (Rajesh, 2018).

Cette méthode est utile pour le filtrage des variables et la sélection de caractéristiques, et elle peut traiter efficacement des données de haute dimension. Comparé à d'autres méthodes, elle nécessite peu de nettoyage des données car elle n'est pas affectée par les valeurs aberrantes. De plus, elle peut gérer à la fois des données quantitatives et qualitatives. Cependant, il existe des faiblesses, comme le problème du surajustement. Un arbre de décision est susceptible de sur ajuster les données si aucune limite de profondeur maximale n'est fixée, car l'arbre peut continuer à croître jusqu'à avoir un nœud feuille pour chaque observation, ce qui le rend trop complexe et nuit à sa capacité de classification. Cette méthode manque également de stabilité, car même une petite variation dans les données peut conduire à la génération d'un arbre de décision complètement différent (Avinash, 2018).

### 3.2 Logistic Regression

La régression logistique est une méthode de régression et statistique puissante pour les problèmes de classification, simple à utiliser et qui peut servir de base pour des problèmes de classification binaire. Ce type de modèle statistique (également appelé modèle logit) est souvent utilisé à des fins de classification et d'analytique prédictive. Comme le résultat est une probabilité, la variable dépendante est comprise entre 0 et 1. Dans la régression logistique, une transformation logit est appliquée aux odds, c'est-à-dire à la probabilité de succès divisée par la probabilité d'échec. Elle est également connue sous le nom de log-odds ou logarithme naturel de l'odds ratio.

Il peut être difficile de comprendre les log-odds dans le contexte d'une analyse des données via régression logistique. Par conséquent, l'exponentiation des estimations bêta est une pratique couramment utilisée pour transformer les résultats en odds ratio (OR), ce qui en facilite l'interprétation. L'OR représente les chances qu'un résultat se produise compte tenu d'un événement particulier, par rapport aux chances que le résultat se produise en l'absence de cet événement. Si l'OR est supérieur à 1, l'événement est associé à des chances plus élevées de générer un résultat spécifique. Inversement, si l'OR est inférieur à 1, l'événement est associé à des chances plus faibles que ce résultat se produise.

Ce modèle présente un grand avantage par rapport à d'autres modèles : il n'est pas seulement un modèle de classification, mais fournit également des probabilités. Cependant, il a ses limitations. L'interprétation est difficile, car les poids ne sont pas additifs mais multiplicatives. De plus, ce modèle souffre de séparation complète. Si une caractéristique peut séparer parfaitement les deux classes, le modèle ne pourra jamais être entraîné, car le poids pour cette caractéristique ne convergera pas et le poids optimal sera infini. Cela pose problème, car cette caractéristique pourrait être très utile. Pour résoudre ce problème, nous pouvons définir une distribution de probabilité a priori des poids (Christoph, 2019).

### 3.3 Random Forest

La forêt aléatoire est une combinaison de prédicteurs en arbre, où chaque arbre dépend des valeurs d'un vecteur aléatoire échantillonné indépendamment et selon la même distribution pour tous les arbres de la forêt (Breiman, 2001). Ces arbres de décision sont assemblés à partir d'un jeu de données d'entraînement et utilise un outil appelé bagging pour effectuer des tâches de classification et de régression. Chaque arbre de décision représente une prédiction de classe, et cette méthode collecte les votes de ces arbres de décision ; la classe ayant le plus de votes est considérée comme la classe finale (Savan, 2017).

Lors de l'entraînement, chaque modèle de base est créé indépendamment en apprenant à partir de différents sous-échantillons aléatoires des données. Les échantillons sont tirés par bootstrap, ce qui signifie que certains échantillons peuvent être utilisés plusieurs fois dans un seul arbre de décision. En utilisant différents sous-échantillons pour entraîner chaque arbre de décision, l'ensemble de la forêt aura une faible variance mais un biais élevé, même si chaque arbre présente une forte variance par rapport à un ensemble particulier de données d'entraînement. Dans la phase test, les prédictions de chaque arbre de décision sont moyennées pour obtenir les prédictions globales. La raison pour laquelle cette méthode surpasse les arbres de décision est que de nombreux arbres de décision non corrélés peuvent se protéger mutuellement des erreurs individuelles pour obtenir des prédictions en ensemble, ce qui réduit ainsi le problème de surajustement et rend les résultats de prédiction inégalés en termes de précision (Tony, 2016).

Ce modèle peut fonctionner efficacement sur un grand nombre de jeux de données et il peut estimer quelles variables sont significatives dans la classification. La forêt aléatoire peut aussi capturer des relations non linéaires entre l'objet et les caractéristiques. Cependant, un inconvénient de cette méthode est qu'elle ne peut pas fonctionner avec des caractéristiques dispersées, car les arbres de décision sont les éléments de base. Il est donc nécessaire de prétraiter les entrées pour les adapter au modèle.

### 3.4 Gradient Boosting

Le gradient boosting est une méthode puissante et efficace pour créer des modèles robustes de classification et de régression. Grâce à sa capacité à combiner plusieurs modèles faibles pour en faire des apprenants forts, il est largement utilisé dans la pratique du machine learning. Cela est réalisé par la généralisation de l'Adaptive Boosting (AdaBoost) grâce aux techniques de pondération et de combinaison adaptatives (Sourav De et al., 2022).

Dans ce processus, chaque nouveau modèle est entraîné pour minimiser la fonction de perte, telle que l'erreur quadratique moyenne ou l'entropie croisée, du modèle précédent en utilisant la descente de gradient. À chaque itération, l'algorithme calcule le gradient de la fonction de perte par rapport aux prédictions de l'ensemble actuel, puis entraîne un nouveau modèle faible pour minimiser ce gradient. Les prédictions du nouveau modèle sont ensuite ajoutées à l'ensemble, et le processus est répété jusqu'à ce qu'un critère d'arrêt soit atteint.

### 3.5 CatBoost

CatBoost ou "Category Boosting" est une implémentation du gradient boosting, spécialement conçue pour gérer les caractéristiques catégorielles en utilisant des arbres de décision binaires comme prédicteurs de base (Liudmila Prokhorenkova et al., 2018). Développé par Yandex, CatBoost se distingue par sa capacité à travailler efficacement avec des variables catégorielles sans nécessiter de prétraitement approfondi.

Le Category Boosting fonctionne sur le principe du gradient boosting, où le modèle est construit de manière progressive. Il commence avec un modèle simple et l'améliore progressivement en ajoutant de nouveaux modèles qui corrigent les erreurs des précédents. Cependant, il introduit plusieurs innovations clés qui le distinguent des autres méthodes de gradient boosting telles que :

- la gestions des variables catégorielles sans nécessiter de les transformer au préalable ;

- l'utilisation d'un ordre aléatoire pour le traitement des données ;

- l'utilisation de l'information de cible qui donne une meilleure performance sur les données avec des distributions déséquilibrés ;

- l'apprentissage par niveaux pour traiter les erreurs de manière efficace.

## 4. Métriques

Table 1 :

| | Positif | Négatif |
|---|---|---|
| Vraix | TP | TN |
| Faux | FP | FN |

True negatives (TN) : Nombre d'observations négatives que le modèle a correctement prédit.

True positives (TP) : Nombre d'observations positives que le modèle a correctement prédit.

False positves (FP) : Nombre d'observations négatives que le modèle a incorrectement prédit comme positives.

False negatives (FN) : Nombre d'observations positives que le modèle a incorrectement prédit comme négatives.

### a) Precision

La précision est la proportion des observations positives correctement prédites par rapport à la somme des prédictions positives effectuées par le modèle. Elle est plus favorisée dans les cas où les faux positifs peuvent avoir des conséquences significatives. Une précision élevée indique que la majorité des prédictions positives sont correctes. Le cas contraire entraine une faible précision. Elle est déterminée comme suit :

$$\text{Precision } = \frac{TP}{TP + FP}$$

**b) Recall**

Le rappel est le rapport des observations positives correctement prédites par rapport à toutes les véritables observations positives. Elle est plus favorisée dans les cas où les faux négatifs peuvent avoir des conséquences significatives. La valeur du rappel varie de 1 à 0 et sa formule est donnée comme suit :

$$\text{Recall } = \frac{TP}{TP + FN}$$

**c) Accuracy**

La précision globale représente le rapport des observations correctement classifiées par le modèle par rapport au nombre total d'observations. Une accuracy élevée montre qu'une grande partie des observations est bien prédite. Le cas contraire indique une faible accuracy. Elle peut être déterminée ainsi :

$$\text{Accuracy } = \frac{TP + TN}{\text{Total Prédictions}}$$

**d) F1 Score**

Le F1 peut être interprété comme la moyenne harmonique de la Précision et du Recall. Le F1 score atteint sa meilleure valeur à 1 et son pire score à 0. Il permet d'obtenir des informations sur la qualité des prédictions, surtout dans les contextes où il est crucial de trouver un compromis entre précision et rappel. La formule du F1 score est la suivante :

$$\text{F1 Score } = \frac{2 * TP}{2 * TP + FP + FN}$$

**e) La courbe de ROC**

La courbe de ROC (Receiver Operating Characteristic) est un puissant outil de visualisation pour évaluer la performance d'un modèle de classification binaire. Cette courbe permet d'observer comment le taux de faux positifs et le taux de vrais positifs évoluent ensemble aux différents seuils, qui sont des limites permettant de différencier la classe positive de la classe négative. Dans un modèle idéal, il existerait un seuil à partir duquel le taux de vrais positifs est élevé et le taux de faux positifs est faible. Donc, plus la courbe ROC épouse le coin supérieur gauche du graphique, mieux le modèle classe les données.

**Les composants de la courbe de ROC sont** :

- **True Positive Rate (TPR) ou Taux de Vrais Positifs**

Le TPR est également appelé recall, il se détermine comme suit :

$$TPR = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR) ou Taux de Faux Positifs**

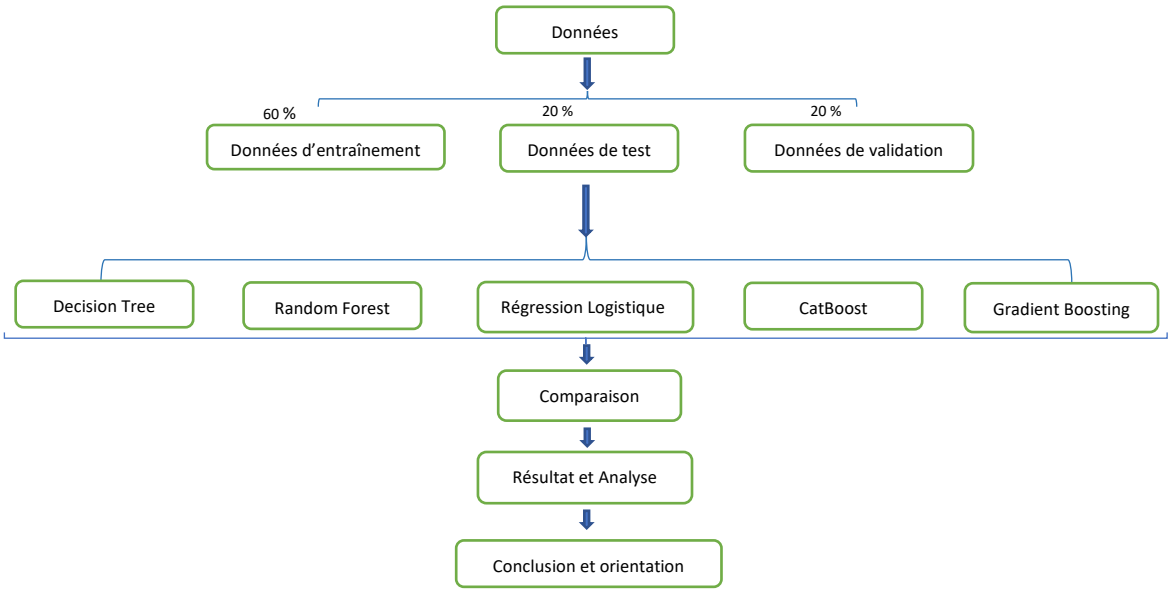
Représente le rapport entre les faux positifs et le nombre total d'observations prédites comme fausses. La formule est la suivante :

$$FPR = \frac{FP}{FP + TN}$$

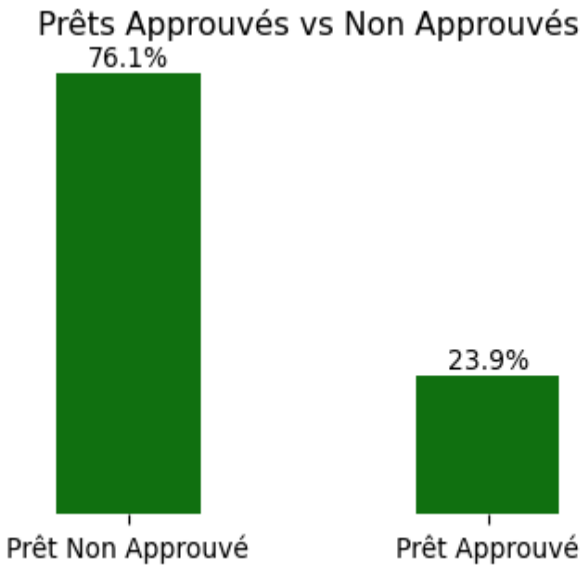
L'aire sous la courbe ROC appelée AUC (Area Under the Curve) fournit une mesure globale du rendement pour tous les seuils de classification possibles. Plus l'AUC est élevée, plus le modèle est bon. L'AUC permet de comparer facilement plusieurs modèles de classification car, non seulement elle garde une certaine indépendance par rapport au seuil de décision, mais elle est robuste face aux déséquilibres de classe.

## 5. Données

**Figure 1** : dispositif expérimental



Dans cet exercice, nous avons utilisé des données bancaires qui regroupent des demandes de prêts des clients. La base contient 20000 enregistrements (clients) et 36 variables parmi lesquelles l'âge, le score de risque, le taux d'intérêt, le statut de propriété, le niveau d'étude ou le fait qu'un prêt a été approuvé ou non. Nous avons effectué des traitements sur les variables catégorielles. Dans le but de rendre plus efficace la modélisation, nous avons scindé les données en 3 parties d'une manière aléatoire : une part de 60% pour l'entraînement et deux parts de 20% pour le test et la validation des modèles. Les données révèlent que seuls 23,9% des demandes de prêts ont été octroyés contre 76,1%.



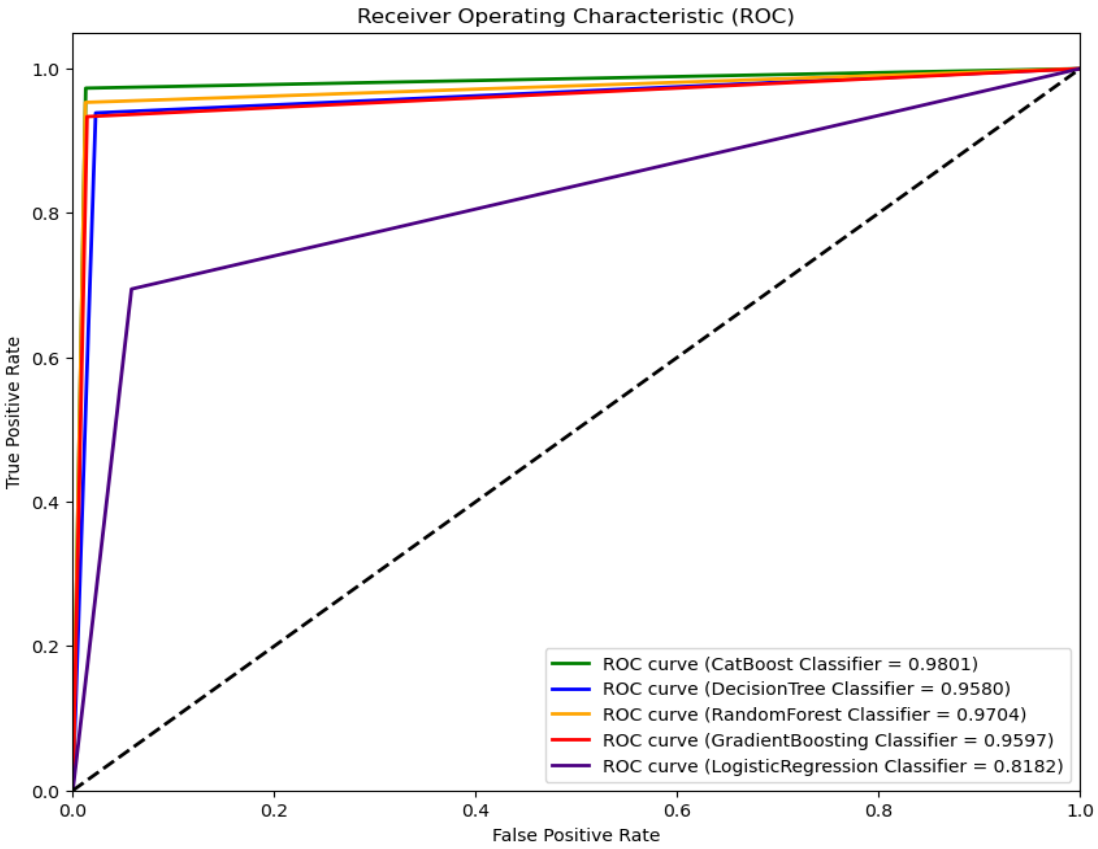

## 6. Résultats et analyse

**Table 2** : *Les performances des modèles utilisés*

| Modèles | AUC | Precision | Accuracy | Recall | F1 score |
|---|---|---|---|---|---|
| **Decision Tree** | 0,9580 | 92,91 % | 96,80 % | 93,87 % | 93,39 % |
| **Random Forest** | 0,9701 | 96,03 % | 97,92 % | 95,33 % | 95,67 % |
| **CatBoost** | 0,9801 | 96,00 % | 98,38 % | 97,30 % | 96,65 % |
| **Gradient Boosting** | 0,9597 | 95,44 % | 97,32 % | 93,35 % | 94,38 % |
| **Logistic Regression** | 0,8182 | 79,08 % | 88,22 % | 69,47 % | 73,96 % |

***Source : auteur***

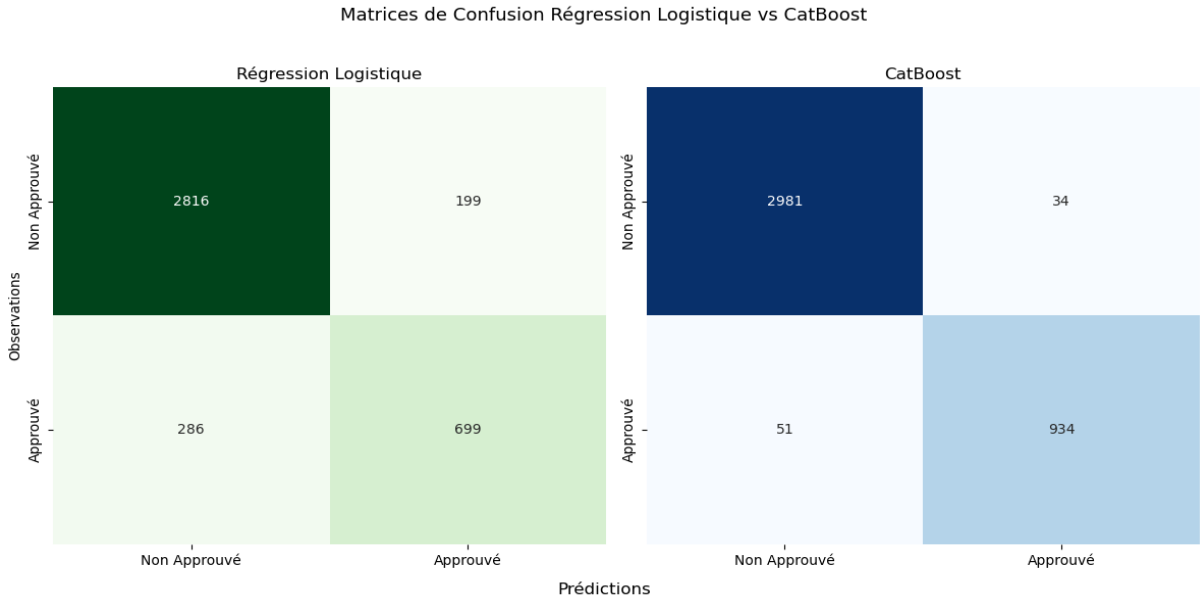
**Figure 2** : Courbes de ROC




***Source : auteur***

L'analyse des résultats obtenus montre que le CatBoost se démarque comme le modèle le plus performant suivi de près par le Random Forest. Par contre, la régression logistique affiche une performance nettement inférieure.

**Figure 3** : Matrices de Confusion LR vs CatBoost

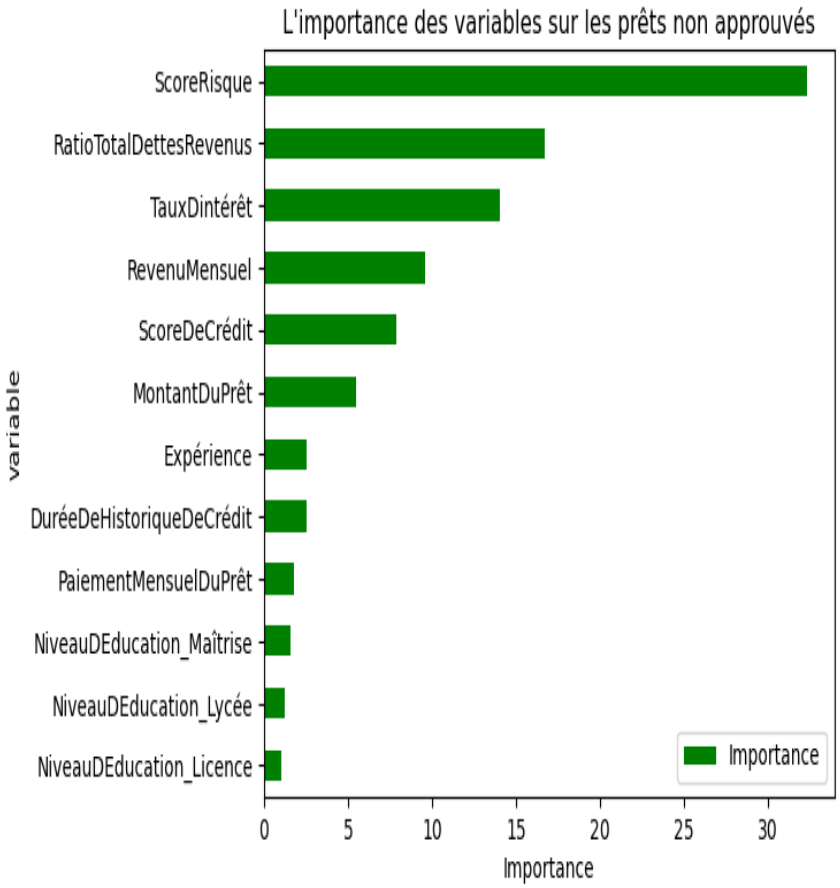



***Source : auteur***

Dans l'optique de mieux énumérer les différences de performances, nous avons choisi deux modèles, le plus et le moins performant en termes de AUC qui sont respectivement le CatBoost et la régression logistique. Ensuite, nous avons appliqué les données de validation dans chacun de ces modèles. Les résultats confirment la performance affichée par chaque modèle.

Parmi 3015 observations non approuvées le modèle de RL a prédit que 199 sont approuvées (alors qu'elles ne pas sont approuvées), tandis que le Catboost en a prédit que 34. De même que pour les 985 observations approuvées, le modèle RL a prédit que 286 ne sont pas approuvées, contre 51 pour le modèle CatBoost. Ces résultats sur les données de validation consolident la différence significative de performance des modèles.

**Figure 4** : L'influence des variables sur l'octroi de prêts




Suite aux performances affichées par le CatBoost dans l'approbation ou non du crédit, nous avons examiné plus en détail les données pour déterminer les variables qui possèdent le plus grand pouvoir prédictif dans le modèle d'apprentissage automatique champion. Au total des 36 variables prédictives utilisées le score de risque, le ratio total dettes revenus, le taux d'intérêt, le score de crédit, le revenu mensuel et le montant du prêt ont été identifiés comme les facteurs les plus importants dans l'octroi de prêts bancaires.

***Source : auteur***

## 7. Conclusion

Dans cette étude, nous cherchions à effectuer une évaluation comparative entre modèles d'apprentissage automatique en évaluant leur performance par le biais de métriques pertinentes dans l'octroi de crédit bancaire. Les modèles utilisés sont la régression logistique, l'arbre de décision, la forêt aléatoire, le gradiant boosting et la catégorie boosting. Le choix de ces modèles est porté par leur performance, leur capacité d'interprétation, leur robustesse, leur adaptabilité mais aussi leur forte utilisation. La base de données utilisée dans cet exercice

est constituée de 20000 demandes de prêts bancaires des clients d'une banque avec 36 caractéristiques dont l'âge, le score de risque, le taux d'intérêt, le statut de propriété, le niveau d'étude ou le fait qu'un prêt a été approuvé ou non. Parmi ces données historiques des demandes, 76,1 % des demandes n'ont pas été approuvées contre 23,9%. Les données sont scindées en 3 parties d'une manière aléatoire : une part de 60% pour l'entrainement et deux parts de 20% pour le test et la validation des modèles. Nos résultats ont montré que, dans la modélisation de l'approbation ou non du crédit, les méthodes de CatBoost et de la forêt aléatoire se démarquent par leur performance contrairement à la régression logistique qui affiche une performance nettement inférieure. Après avoir déterminé la performance de chaque modèle par rapport aux données de test, nous avons pris le modèle le plus performant (CatBoost) et celui le moins performant (la régression logistique) pour voir leur réaction face aux données de validation. Les résultats confirment la performance affichée par chaque modèle durant la phase test. La régression logistique a effectué une erreur de prédiction sur les observations non approuvées de 6,6% contre 1,12% pour le catboost. Pour les observations approuvées, les erreurs de prédictions pour la RL et le Catboost sont respectivement 29 % et 5%. Nous avons constaté également que les variables telles que le score de risque, le ratio total dettes revenus, le taux d'intérêt, le score de crédit, le revenu mensuel et le montant du prêt sont les facteurs les plus importants dans l'octroi de prêts bancaires selon le modèle Catboost. En fin, les résultats montrent que les modèles de machines learning peuvent devenir des outils nécessaires voire incontournables dans le processus de traitement des demandes de crédits ; mais il est essentiel d'identifier le modèle le plus approprié face à une situation donnée pour obtenir les meilleurs résultats possibles.

## 8. *Références*


Avinash, M. B., & Govindaraju, T. (2018). Architectonics: design of molecular architecture for functional applications. *Accounts of chemical research*, *51*(2), 414-426.

Bahnsen, Alejandro Correa, et al. (2016): "Feature engineering strategies for credit card fraud detection." *Expert Systems with Applications* 51, 134-142.

Bai, M., Zheng, Y., & Shen, Y. (2022). Gradient boosting survival tree with applications in credit scoring. *Journal of the Operational Research Society*, *73*(1), 39-55.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Haschka, J., Englbrecht, M., Hueber, A. J., Manger, B., Kleyer, A., Reiser, M., ... & Rech, J. (2016). Relapse rates in patients with rheumatoid arthritis in stable remission tapering or stopping antirheumatic therapy: interim results from the prospective randomised controlled RETRO study. *Annals of the rheumatic diseases*, *75*(1), 45-51.

Hemann, E. A., Gale Jr, M., & Savan, R. (2017). Interferon lambda genetics and biology in regulation of viral control. *Frontiers in immunology*, *8*, 1707.

Jiang, Cuiqing, et al. (2017). "Capturing helpful reviews from social media for product quality improvement: a multi-class classification approach." *International Journal of Production Research* 55,12: 3528-3541.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Rajesh, R. (2018). On sustainability, resilience, and the sustainable–resilient supply networks. *Sustainable Production and Consumption*, *15*, 74-88.

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., & Thrash, J. C. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*, *13*(12), 3126-3130.

WANG, Maoguang, GE & al 2016 Research on the Innovation Closed-loop Internet Financial Service Model of Traditional Industry.*International Conference on Computer Science and Electronic Technology*. Atlantis Press, p. 242-245.

Wang, Y., Sourav, S., Malizia, J. P., Thompson, B., Wang, B., Kunz, M. R., ... & Fushimi, R. (2022). Deciphering the mechanistic role of individual oxide phases and their combinations in supported Mn–Na2WO4 catalysts for oxidative coupling of methane. *ACS Catalysis*, *12*(19), 11886-11898.

https://www.ibm.com/fr-fr/topics/logistic-regression

Documentation ML : https://scikit-learn.org/stable/