ANALYSE PREDICTIVE DE L'ACCEPTATION DES PRETS PERSONNELS : CONCEPTION ET EVALUATION DE MODELES DE CLASSIFICATION

Auteur : SEGBEDJI KOKUVI Barnabé

Email : sebarnabe90@gmail.com

INTRODUCTION

Dans un environnement économique caractérisé par une concurrence intense parmi les établissements financiers et des marges d'intérêt de plus en plus étroites, les banques sont obligées de réviser radicalement leurs tactiques commerciales pour maintenir leur profitabilité et consolider leur présence sur le marché. Cette conjoncture les contraint non seulement à séduire de nouveaux clients, mais aussi à optimiser la valeur de leur portefeuille actuel en misant sur une diversification astucieuse des produits et services offerts.

Les prêts personnels représentent un élément clé de l'offre bancaire parmi ces produits. Source de revenus significatifs pour les institutions financières, grâce aux gains d'intérêts périodiques qu'elles produisent, tout en favorisant l'instauration d'un lien durable et fiable entre la banque et sa clientèle. Cependant, la mise sur le marché de ces articles représente une difficulté, surtout à cause des dépenses importantes liées aux campagnes de marketing et du faible taux de réussite constaté lorsque les propositions ne sont pas assez ciblées.

La banque, à l'instar de nombreuses institutions financières, dispose d'une base de données clients vaste et diversifiée, majoritairement composée de déposants. Cependant, malgré ce potentiel important, la proportion de clients ayant effectivement souscrit à un prêt personnel demeure relativement faible. Cette situation limite la capacité de la banque à exploiter pleinement son portefeuille clients et à maximiser ses revenus d'intérêts.

Face à ce constat, la banque cherche à identifier de manière plus précise les clients présentant une forte probabilité d'accepter une offre de prêt personnel. L'objectif est de passer d'une approche commerciale généraliste à une stratégie de ciblage plus fine, fondée sur l'analyse des caractéristiques socio-économiques et comportementales des clients.

Afin d'atteindre ce but, La banque envisage d'utiliser des méthodes modernes de prévision des enjeux de l'apprentissage automatique. Ces techniques facilitent l'utilisation judicieuse des données passées pour approfondir la connaissance des profils de clientèle, identifiant des configurations complexes qui échappent aux méthodes statistiques conventionnelles et établirdes modèles aptes à anticiper les comportements futurs des clients.

La banque espère conseiller et optimiser ses campagnes marketing, minimiser les coûts liés à des campagnes inefficaces et augmenter l'efficacité des offres commerciales, en augmentant la probabilité qu'un client empruntera de l'argent. En d'autres termes, il s'agit de prendre des décisions basées sur les données data-driven decision making, qui étaient une tendance très importante ces dernières années. La banque est l'institution en expansion. Idéalement, l'une des tendances ambitieuses à maîtriser est la commercialisation des services de crédit aux consommateurs.

Au cours d'une précédente campagne, on a noté un taux de conversion approximatif de 9 % parmi les clients qui ont déposé. Ce résultat est prometteur, toutefois, il reste insuffisant pour assurer un développement pérenne. De plus, les campagnes de marketing globales engendrent des coûts importants sans nécessairement viser les clients qui sont vraiment susceptibles d'accepter une proposition de prêt.

Ainsi, la banque est confrontée à un double enjeu, notamment Déterminer avec exactitude les profils de clients qui sont le plus susceptibles d'accepter un crédit personnel et Affiner les campagnes promotionnelles pour diminuer les dépenses superflues tout en améliorant le taux de conversion.

Dans ce cadre, la question clé qui émerge est la suivante :

Comment prédire efficacement les clients ayant une forte probabilité d'accepter une offre de prêt personnel ?

Ce projet vise ainsi à proposer une approche de classification basée sur l'apprentissage automatique afin d'aider la banque à cibler efficacement ses clients et à améliorer ses performances commerciales.

L'objectif général de ce projet est de développer un modèle de classification prédictif permettant d'identifier les types de clients ayant une forte probabilité d'accepter un prêt personnel.

Plus spécifiquement :

- Décrire le profil des clients selon leur statut d'acceptation de prêt,
- Identifier ls principaux facteurs influençant la probabilité d'acceptation d'un prêt personnel
- Construire et évaluer un modèle de classification prédictif permettant d'estimer la probabilité qu'un client accepte une offre de prêt.

Hypothèses :

- **H1 :** Les clients ayant un revenu annuel élevé présentent une probabilité plus forte d'accepter un prêt personnel,
- **H2 :** Le niveau d'éducation et l'usage des services bancaires numériques (internet, carte de crédit) influencent positivement la décision d'accepter un prêt personnel,
- **H3 :** Les clients ayant une dépense moyenne mensuelle élevée sur leur carte de crédit sont plus susceptibles d'accepter une offre de prêt personnel,

Chapitre 1 : Définition des concepts et Revue de littérature

1.1. Définitions et concepts clés

Le crédit scoring (ou notation du crédit) désigne l'ensemble des méthodes statistiques et algorithmiques employées pour classer les emprunteurs potentiels en « bons » et « mauvais » risques, l'objectif est d'estimer la probabilité d'un individu (ou une entreprise) rembourse un prêt, ou comme dans notre ca la probabilité qu'un client accepte une offre de prêt. Le problème est de formalisé comme une classification binaire : : $Y \in \{0,1\}$, où $Y = 1$ représente l'évènement d'intérêt qui est l'acception d'un prêt, défaut etc.

Les approches traditionnelles reposent largement sur la régression logistique et les scorecards, tandis que les approches récentes exploitent les méthodes d'apprentissages automatique (arbre de décision, forêts aléatoires, boosting, et les réseaux de neurones) pour améliorer la capacité prédictive.

1.1.1. Prêt personnel

Un prêt personnel est une forme de crédit accordée par une institution financière à un particulier pour répondre à des besoins personnels, tels que des dépenses familiales, des projets de rénovation ou des urgences financières, Contrairement aux prêts hypothécaires ou automobiles, il n'est pas adossé à une garantie spécifique.

Dans le contexte bancaire, le prêt personnel constitue un produit stratégique important, Il permet d'augmenter la rentabilité des banques grâce aux intérêts générés sur la durée du prêt, Cependant, ce produit présente également un risque de non-remboursement, incitant les banques à évaluer rigoureusement la solvabilité et le profil socio-économique de chaque client avant d'accorder un prêt, Cette vérification inclut la consultation des fichiers d'incidents de paiement et la collecte d'informations sur les revenus, charges et autres engagements financiers de l'emprunteur, Ce contrôle étau. à limiter le risque de défaut tout en optimisant l'offre commerciale[1],

1.1.2. Propension à accepter une offre de prêt

La Tendance à accepter une proposition de prêt La tendance à approuver une offre de crédit représente la possibilité qu'un client accepte une proposition commerciale, notamment une proposition de crédit personnel. Toutefois, cette notion provenant du marketing analytique est cruciale dans la stratégie socio-économique des clients. En effet, les banques peuvent déterminer les profils les plus enclins à réagir favorablement à une campagne d'offre de crédit. Cette méthode facilite l'optimisation des ressources marketing et l'amélioration des taux de conversion des campagnes.

1.1.3. Modèle de classification

Dans le contexte de ce projet, un modèle de classification et une technique d'apprentissage supervisé sont mis en œuvre pour prédire la catégorie associée à une observation basée sur ses

[1] Hand, D.J. & Henley, W.E. (1997). *Statistical classification methods in consumer credit scoring: a review.* Journal of the Royal Statistical Society.

attributs. L'objectif du modèle est de décider si un client accepte ou refuse une proposition de crédit personnel. Parmi les techniques de classification des employés, en compte la régression logistique, les arbres décisionnels, la forêt aléatoire ainsi que les méthodes de renforcement par gradient. Ces modèles se distinguent les uns des autres par leur complexité, leur aptitude à apprendre et leur capacité d'interprétation.

1.1.4. Évaluation de la performance des modèles

L'appréciation d'un Le modèle de classification se base sur divers indicateurs de performance qui permettent d'évaluer son aptitude à réaliser des prédictions précises.

Parmi les plus pratiques employés nous avons :

- **L'exactitude (Accuracy)**, qui mesure la proportion de prédictions bonne prédictions,
- **La précision (Precision)** et le **rappel (Recall)**, qui évaluent respectivement la performance des prédictions positives et la capacité du modèle à ressortir les cas positifs,
- **Le score F1**, qui combine la précision et le rappel pour donner une mesure équilibrée de la performance,
- **L'AUC (Area Under the Curve)**, qui montre la capacité globale du modèle à ciblé les clients susceptibles d'accepter un prêt de ceux qui le refuseront,

1.2. Revue de littérature

1.2.1. Évolution des méthodes : des modèles statistiques aux algorithmes Machine Learning

Dans les années passé, la régression logistique a été la méthode de la plus utilisé en pratique bancaire à cause de sa simplicité, de son interprétabilité et de sa robustesse, Elle permet d'estimer directement la probabilité $P(Y = 1 \mid X)$ via la fonction logistique :

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp\left(-(\beta_0 + \sum_j \beta_j X_j)\right)},$$

Cependant, lors des années 2000, l'arrivé de grands base de données transactionnels et le développement d'algorithmes non linéaires ont conduit à l'adoption progressive de méthodes d'apprentissage automatique (arbres de décision, forêts aléatoires, gradient boosting, réseaux neuronaux), Ces méthodes captent des interactions compliqués et non-linéarités que des modèles linéaires simples peuvent ne pas ressortir, Plusieurs études empiriques montrent que les méthodes d'ensemble (Random Forest, Gradient Boosting) dépassent habituellement la régression logistique en termes de performances prédictives sur des données de credit scoring[2].

[2] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L.C. (2015). *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.* European Journal of Operational Research.

1.2.2. Travaux empiriques majeurs et enseignements

Revue classique : Hand et Henley (1997) ont Synthétisé les avantages et limites des méthodes statistiques appliquées au scoring, souligne les problèmes pratiques (déséquilibre des classes, coût asymétrique des erreurs) et insiste sur la nécessité d'évaluer un modèle au-delà de l'accuracy (métriques économiques, coût, etc,),

Application de méthodes Machine Learning au risque de crédit : montre que les techniques d'apprentissage machinent non paramétriques améliorent de façon substantielle la capacité de prédiction des défaillances par rapport aux approches linéaires classiques, surtout quand on combine données transactionnelles et données de bureau de crédit, Leurs résultats soulignent l'intérêt d'utiliser des séries temporelles comportementales et des features riches pour le scoring, Khandani, Kim et Lo (2010)[3]

Benchmarking moderne : étude comparative sur de nombreux classifieurs (41 algorithmes) appliqués au credit scoring ; conclusion principale : les méthodes d'ensemble (notamment Random Forest et boosting) offrent des gains de performance notables et stables par rapport aux approches traditionnelles, particulièrement lorsque l'évaluation est faite à l'aide de métriques robustes (AUC, etc,), Ils insistent aussi sur l'importance d'un protocole d'évaluation rigoureux (validation croisée, prise en compte du coût des erreurs) Lessmann et al, (2015),

XGBoost (méthode de boosting efficace) : XGBoost est devenu un standard pratique pour les données tabulaires (performant et optimisé pour la vitesse/mémoire), De nombreuses compétitions et applications industrielles (dont le scoring) l'adoptent pour son aptitude à capter rapidement des patterns complexes et à gérer des données clairsemées, Chen et Guestrin (2016)[4],

« Les études récentes convergent vers l'idée que les méthodes d'ensemble (Random Forest, Gradient Boosting / XGBoost) offrent des améliorations significatives de performance par rapport aux modèles linéaires traditionnels pour les tâches de scoring et de prédiction d'acceptation de prêts, surtout lorsqu'elles sont enrichies de variables comportementales et de features alternatives, Néanmoins, la littérature met en garde contre les risques d'opacité décisionnelle : l'implémentation opérationnelle exige une attention particulière à l'interprétabilité (SHAP, LIME) et à l'évaluation économique des erreurs (matrices de coût), Pour le cas de la banque, il apparaît donc pertinent de comparer des modèles simples et des modèles d'ensemble, de combiner une validation robuste (cross-validation stratifiée) et d'intégrer des méthodes d'explicabilité afin de rendre les recommandations actionnables pour le département marketing, »[5]

---

[3] Khandani, A.E., Kim, A.J., & Lo, A.W. (2010). *Consumer credit-risk models via machine-learning algorithms.* Journal of Banking & Finance.

[4] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* KDD.

[5] (2023) Machine Learning for Credit Risk Prediction : A Systematic Literature Review, Uddin N. (2023) An ensemble machine learning based bank loan approval predictions system with a smart application, Davis R. (2022) Explainable Machine Learning Models of Consumer Credit Risk

1.2.3. Problèmes récurrents et bonnes pratiques identifiées dans la littérature

- **Déséquilibre des classes** : en scoring, la classe d'intérêt (défaillance, ou ici acceptation) est souvent minoritaire, Les auteurs recommandent d'utiliser des techniques comme le rééchantillonnage (undersampling/oversampling), SMOTE, ou des métriques robustes (AUC, F1, recall) plutôt que l'accuracy seule,
- **Coût asymétrique des erreurs** : une fausse prédiction (faux négatif ou faux positif) n'a pas le même coût pour la banque ; l'évaluation doit intégrer cette asymétrie (ex, matrices de coût, profits),
- **Interprétabilité** : malgré la supériorité prédictive des méthodes d'ensemble, la régression logistique et les arbres simples restent appréciés pour leur explicabilité ; des méthodes d'interprétation (SHAP, importance de features) permettent d'obtenir de l'intelligibilité même pour des modèles complexes,
- **Validation rigoureuse** : cross-validation, séparation train/test temporelle si données longitudinales, et sélection d'hyperparamètres via recherche systématique sont des étapes indispensables pour éviter l'optimisme sur les performances,

Chapitre 2 : Méthodologie

2.1. Source et nature des données

L'étude repose sur l'utilisation d'une base de données réelle de la plateforme kaggle, dont la taille fait 5000 clients de la banque, chaque client se décrit par plusieurs variables socio-économique, financière et du comportement dans les dépenses

L'objectif est de prédire la probabilité qu'un client accepte l'offre de prêt personnel à partir de ses différentes caractéristiques.

Les variables clés utilisées sont les suivantes :

- Age : Âge du client (en années)
- Expérience : Années d'expérience professionnelle
- Incombe : Revenu annuel (en milliers de dollars)
- Family : Taille de la famille
- CCAvg : Dépense moyenne mensuelle sur carte de crédit (en milliers de dollars)
- Education : Niveau d'éducation (1 = Undergraduate, 2 = Graduate, 3 = Advanced/Professional)
- Mort-gage : Valeur du prêt immobilier (en milliers de dollars)
- Securities **Account**, **CD Account**, **Online**, **CreditCard** : variables binaires indiquant la possession de certains services bancaires
- Personal **Loan** : variable cible (1 = prêt accepté, 0 = prêt refusé)

2.2. Étapes méthodologiques

2.2.1. Préparation et nettoyage des données

La première étape consiste à préparer les données pour les rendre exploitables, Elle comprend :

- **Vérification et traitement des valeurs manquantes**, Dans ce jeu de données, aucune valeur manquante significative n'est reportée, mais une imputation par la médiane pourrait être envisagée en cas de données réelles,
- **Normalisation ou standardisation** des variables continues afin de garantir une meilleure performance des algorithmes sensibles aux échelles de mesure (comme la régression logistique),

  La standardisation suit la formule :

$$Z_i = \frac{X_i - \bar{X}}{s}$$

où :

- Z_i est la valeur standardisée,
- X_i est la valeur observée,

- $\bar{X}$ est la moyenne de la variable,
- s est l'écart-type,

2.2.2. Analyse descriptive des données

Cette étape vise à explorer les caractéristiques des clients et à mieux comprendre les profils associés à l'acceptation ou non du prêt,

Les outils utilisés incluent :

- **Statistiques descriptives** (moyenne, médiane, variance, etc,) ;
- **Visualisations** (histogrammes, boxplots, heatmap de corrélations) pour détecter les tendances et relations entre variables,

L'objectif est de décrire le profil type des clients ayant accepté le prêt, et d'identifier les différences significatives avec ceux qui l'ont refusé,

2.2.3. Modélisation prédictive

Plusieurs modèles de classification supervisée sont construits et comparés :

2.2.3.1. Régression logistique

La régression logistique est une technique statistique utilisée pour évaluer la probabilité qu'un événement se réalise, comme dans le cas où un client accepte l'offre d'un prêt personnel.

Cela se base sur une fonction logistique qui convertit une combinaison linéaire de variables explicatives en une probabilité se situant entre 0 et 1 :

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

où :

- $P(Y = 1 \mid X)$ représente la probabilité qu'un client accepte le prêt,
- X_i sont les variables explicatives (revenu, niveau d'éducation, dépenses par carte, etc,),
- β_i sont les coefficients estimés du modèle.

Les interprétations des résultats (coefficients) se font en termes d'odds ratio, c'est-à-dire le rapport entre la probabilité d'accepter le prêt et la probabilité de le refuser.

La régression logistique est un modèle simple mais capable d'identifier les facteurs qui détermine le comportement des clients (acceptation et refus).

2.2.3.2. Arbre de décision (Decision Tree)

L'arbre de décision est un modèle non paramétrique qui segmente la population en sous-groupes homogènes en fonction des variables explicatives.

À chaque nœud, une variable est sélectionnée afin de maximiser la pureté des classes selon un critère donné, tel que l'indice de Gini ou l'entropie.

Indice de Gini :

$$Gini(t) = 1 - \sum_{i=1}^{C} p(i \mid t)^2$$

Entropie :

$$Entropy(t) = -\sum_{i=1}^{C} p(i \mid t)\log_2(p(i \mid t))$$

où $p(i \mid t)$est la proportion d'observations appartenant à la classe $i$dans le nœud t,

L'arbre de décision présente l'avantage d'être facilement interprétable et de pouvoir représenter visuellement le processus de décision des clients de la banque, Cependant, il peut être sensible au surapprentissage (overfitting).

2.2.3.3. Random Forest

La Random Forest est une méthode d'ensemble (ensemble learning) qui combine plusieurs arbres de décision construits sur des échantillons aléatoires des données.

Chaque arbre vote pour une classe, et la prédiction finale correspond à la majorité des votes.

Formellement, si l'on construit B arbres de décision $h_1(x), h_2(x), \dots, h_B(x)$, la prédiction finale est donnée par :

$$\hat{y} = \text{mode}\,\{h_1(x), h_2(x), \dots, h_B(x)\}$$

Le principe de la Random Forest repose sur deux sources de l'aléa :

- le bootstrap sampling (sélection aléatoire d'observations) ;
- le choix aléatoire des sous-ensembles des variables à chaque nœud.

Ce double mécanisme réduit la corrélation entre arbres et améliore la performance générale du modèle.

La Random Forest est souvent utilisée dans les études bancaires ppour de raison comme son excellente capacité à prédire et sa robustesse face au bruit dans les données.

2.2.3.4. XGBoost (Extreme Gradient Boosting)

L'algorithme XGBoost est une version améliorée du Gradient Boosting, une méthode d'ensemble qui regroupe plusieurs modèles de faible performance (généralement des arbres peu profonds) afin de créer un modèle performant.

Chaque nouvel arbre est ajusté sur les résidus (erreurs) du modèle précéden.

Le modèle final est obtenu par sommation pondérée des arbres :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

où $\mathcal{F}$désigne l'ensemble des fonctions d'arbres de décision,

L'objectif est de minimiser une fonction de coût composée d'un terme d'erreur et d'un terme de régularisation :

Le modèle minimise la fonction de coût suivante :

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

avec

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \mid\mid w \mid\mid^2$$

où :

- $l(y_i, \hat{y}_i)$: est la fonction de perte (log loss pour une classification binaire) ;
- T : est le nombre de feuilles ;
- w : représente les poids associés aux feuilles ;
- γ et λ : sont les paramètres de régularisation.

La rapidité d'exécution, la capacité à éviter trop apprentissage par la régularisation, et l'excellente performance sur les données en format table comme celles des banques sont autant de points forts qui font l'intérêt du XGBoost.

2.3. Évaluation des modèles de classification

Après la construction des différents modèles de classification (Régression logistique, Arbre de décision, Random Forest et XGBoost), il est essentiel d'évaluer leurs performances afin d'identifier celui offrant les meilleures capacités prédictives. L'évaluation repose sur un échantillon de test distinct des données d'apprentissage, afin de garantir une mesure objective de la qualité de généralisation du modèle.

Les métriques utilisées pour cette comparaison sont : **l'Accuracy**, la **Précision**, le **Rappel (Recall)**, le **F1-score**, et la **Surface sous la courbe ROC (AUC).**

2.3.1. Matrice de confusion

La matrice de confusion est l'outil de base pour évaluer un modèle de classification binaire, Elle permet de comparer les prédictions du modèle aux valeurs réelles et se présente comme suit :

Tableau 1 : Structure de la matrice de confusion

	Prédit : Oui (1)	**Prédit : Non (0)**
Réel : Oui (1)	VP (Vrai Positif)	FN (Faux Négatif)
Réel : Non (0)	FP (Faux Positif)	VN (Vrai Négatif)

- **VP** : le modèle a correctement prédit qu'un client acceptera le prêt,
- **VN** : le modèle a correctement prédit qu'un client refusera le prêt,
- **FP** : le modèle a prédit à tort qu'un client accepterait un prêt,
- **FN** : le modèle a prédit à tort qu'un client refuserait un prêt,

À partir de cette matrice, on calcule les principaux indicateurs de performance,

2.3.2. Accuracy (Taux de précision globale)

L'« Accuracy » mesure la proportion totale de prédictions correctes effectuées par le modèle, Elle est donnée par la formule :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Bien qu'elle soit simple à interpréter, l'Accuracy peut être trompeuse en présence de classes déséquilibrées (par exemple, si très peu de clients acceptent le prêt).

Dans ce cas, il convient d'examiner également d'autres métriques plus fines.

2.3.3. Précision (Precision)

La « Précision » évalue la proportion de clients prédits comme « acceptant le prêt » qui l'acceptent réellement :

$$Precision = \frac{VP}{VP + FP}$$

Une précision élevée signifie que la banque cible correctement les clients les plus susceptibles d'accepter le prêt, réduisant ainsi les coûts de campagnes marketing inutiles.

2.3.4. Rappel (Recall ou Sensibilité)

Le « Recall » mesure la capacité du modèle à identifier tous les clients qui ont effectivement accepté un prêt :

$$Recall = \frac{VP}{VP + FN}$$

Un rappel élevé est essentiel pour la banque si l'objectif est de ne pas manquer les clients intéressés par une offre de prêt.

2.3.5. F1-score

Le « F1-score » représente la moyenne harmonique entre la Précision et le Rappel, Il fournit une mesure équilibrée lorsque les deux critères sont importants :

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Cette métrique est particulièrement utile lorsque la base de données est déséquilibrée, c'est-à-dire lorsque le nombre de clients acceptant le prêt est faible par rapport à ceux qui refuse.

2.3.6. AUC (Area Under the ROC Curve)

La courbe ROC (Receiver Operating Characteristic) trace le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour différents seuils de probabilité.

La surface sous la courbe (AUC) mesure la capacité du modèle à distinguer les deux classes (acceptation ou refus du prêt).

$$TPR = \frac{VP}{VP + FN}; FPR = \frac{FP}{FP + VN}$$

Une AUC proche de 1 indique un excellent pouvoir de discrimination, tandis qu'une valeur proche de 0,5 traduit une performance équivalente à un tirage aléatoire.

2.3.7. Sélection du meilleur modèle

Le meilleur modèle est sélectionné sur la base d'une analyse combinée des métriques précédentes.
Dans le contexte de la banque :

- un modèle avec un F1-score élevé sera privilégié pour équilibrer précision et rappel ;
- un AUC supérieur à 0,85 sera considéré comme performant ;
- la matrice de confusion servira à identifier les cas d'erreurs critiques,

Cette approche permet d'assurer une sélection de modèle non seulement performante sur le plan statistique, mais également pertinente pour la prise de décision marketing de la banque,

2.4. Outils et environnement de travail

L'ensemble des analyses et modélisations de ce projet a été réalisé dans un environnement Python, choix motivé par sa puissance pour le traitement de données, sa flexibilité pour la modélisation statistique et le machine learning, ainsi que la richesse de ses bibliothèques dédiées à la visualisation et à l'analyse prédictive.

- **Langage et environnement**

- **Python** : utilisé pour le traitement des données, la construction des modèles de classification et la génération des visualisations ;
- **IDE** : Jupyter Notebook, permettant d'intégrer le code, les visualisations et les commentaires dans un seul document interactif, facilitant la reproductibilité et la présentation des résultats.

Chapitre 3 : Présentation des Résultats

Dans cette section, nous présentons les principales observations issues de l'analyse des données recueillies. Les résultats sont organisés de manière à répondre aux objectifs de l'étude et à mettre en évidence les tendances majeures.

3.1. Présentation et définition des variables

Tableau 2 : Présentation des variables

Variable	Définition
ID	Identifiant unique attribué à chaque client de la banque
Age	Âge du client exprimé en années complètes
Experience	Nombre d'années d'expérience professionnelle du client
Income	Revenu annuel du client, exprimé en milliers de dollars américains
ZIP Code	Code postal du lieu de résidence du client
Family	Taille du ménage du client (nombre de personnes dans la famille)
CCAvg	Dépenses moyennes mensuelles du client par carte de crédit, en milliers de dollars
Education	Niveau d'éducation du client : 1 = Undergraduate, 2 = Graduate, 3 = Advanced/Professional
Mortgage	Montant du prêt immobilier du client, en milliers de dollars
Personal Loan	Variable cible indiquant si le client a accepté un prêt personnel (1 = Oui, 0 = Non)
Securities Account	Indique si le client détient un compte de titres à la banque (1 = Oui, 0 = Non)
CD Account	Indique si le client possède un compte à terme (Certificate of Deposit) à la banque (1 = Oui, 0 = Non)
Online	Indique si le client utilise les services bancaires en ligne (1 = Oui, 0 = Non)
CreditCard	Indique si le client possède une carte bancaire émise par la banque (1 = Oui, 0 = Non)

Source : Auteur, base Kaggle

3.2. Analyse univarié

Tableau 3 : statistique describtive des variables quantitatives

	Age	Experience	Log_Income	Family	Log_CCAvg	Log_Mortgage
mean	45,338400	20,119600	4,085451	2,396400	0,929345	1,566982
std	11,463166	11,440484	0,696455	1,147663	0,533277	2,366428
min	23,000000	0,000000	2,079442	1,000000	0,000000	0,000000
25%	35,000000	10,000000	3,663562	1,000000	0,530628	0,000000
50%	45,000000	20,000000	4,158883	2,000000	0,916291	0,000000
75%	55,000000	30,000000	4,584967	3,000000	1,252763	4,624973
max	67,000000	43,000000	5,411646	4,000000	2,397895	6,455199

Source : Calcul de l'auteur, Base Kaggle

- **Âge**

L'âge présente une moyenne de 45,3 ans avec un écart-type de 11,46 ans, ce qui indique une population adulte relativement hétérogène. Les âges s'étendent de 23 à 67 ans, ce qui montre que les clients de la banque couvrent à la fois de jeunes adultes en début de carrière et des individus plus âgés, probablement en situation financière plus stable et ceux en fin de carrière.

- **Expérience Professionnelle**

L'expérience varie entre **0 et 43 ans**, avec une moyenne de **20,1 ans**. Cette distribution montre une présence significative de travailleurs expérimentés. Cependant, la valeur minimale à **0 an** révèle l'existence de clients nouvellement entrés sur le marché du travail, ce qui peut influencer leur capacité d'engagement financier.

- **Income (Revenu transformé)**

Le revenu présente une moyenne de **4,09**, avec une dispersion modérée (écart-type = 0,70). La fourchette va de **2,08 à 5,41**. Cela montre des disparités de revenu importantes entre les clients, ce qui est typique dans un portefeuille bancaire. Après transformation logarithmique, la distribution est mieux centrée et moins influencée par les valeurs extrêmes.

- **Family (Taille du ménage)**

La taille du ménage va de **1 à 4 personnes**, avec une moyenne de **2,40**. La plupart des clients ont donc un foyer de petite taille, ce qui peut affecter leurs dépenses, leur épargne et leur capacité d'emprunt.

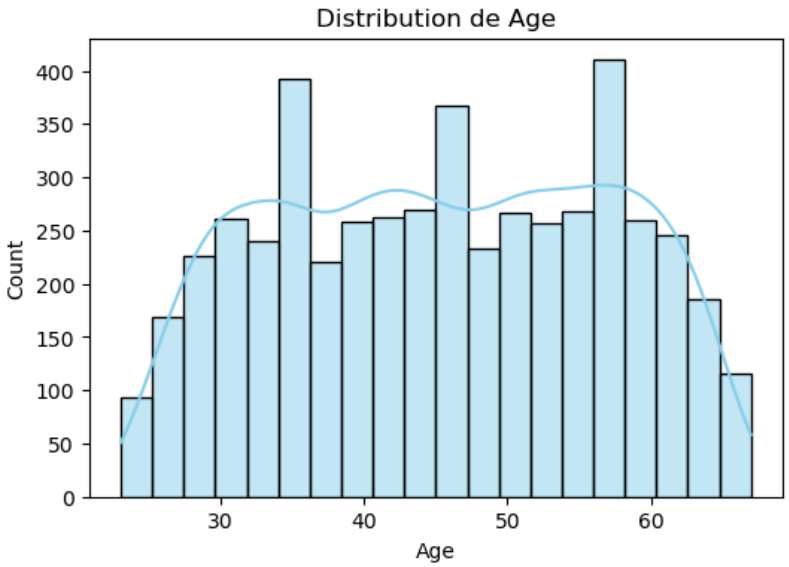
- **CCAvg (Dépenses moyennes carte de crédit, log-transformées)**

La moyenne des dépenses moyenne est de **0,93**, avec une variabilité notable (écart-type = 0,53). Certaines valeurs sont nulles, indiquant des clients n'utilisant pas ou très peu leur carte de crédit. Le maximum de **2,39** montre une catégorie de clients ayant une forte activité de crédit, probablement bancarisée et solvable.

- **Mortgage**

La valeur moyenne de **1,57** masque une forte hétérogénéité : beaucoup de clients n'ont **pas de crédit immobilier** (min = 0), tandis que d'autres présentent des montants relativement élevés (max = 6,45). L'écart-type de **2,37** confirme une répartition très dispersée, typique d'un portefeuille mélangeant accédants au logement et non-propriétaires.

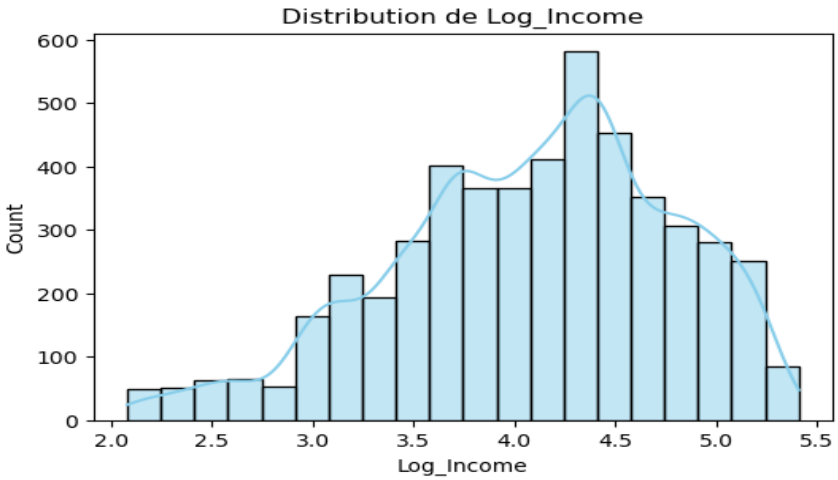
Figure 1 : Distribution de l'âge des clients de la banque



Source : Calcul de l'auteur, Base Kaggle

Le graphique des âges des clients présente une distribution multimodale très révélatrice, illustrant que la clientèle se concentre sur trois segments cruciaux du cycle de vie financier : les jeunes professionnels vers 35 ans, les clients d'âge moyen vers 45 ans, et le segment le plus important, les pré-retraités autour de 58-60 ans, un groupe clé pour les services de planification successorale et les produits de retraite.

Graphique 2 : Répartition de clients selon la revenue

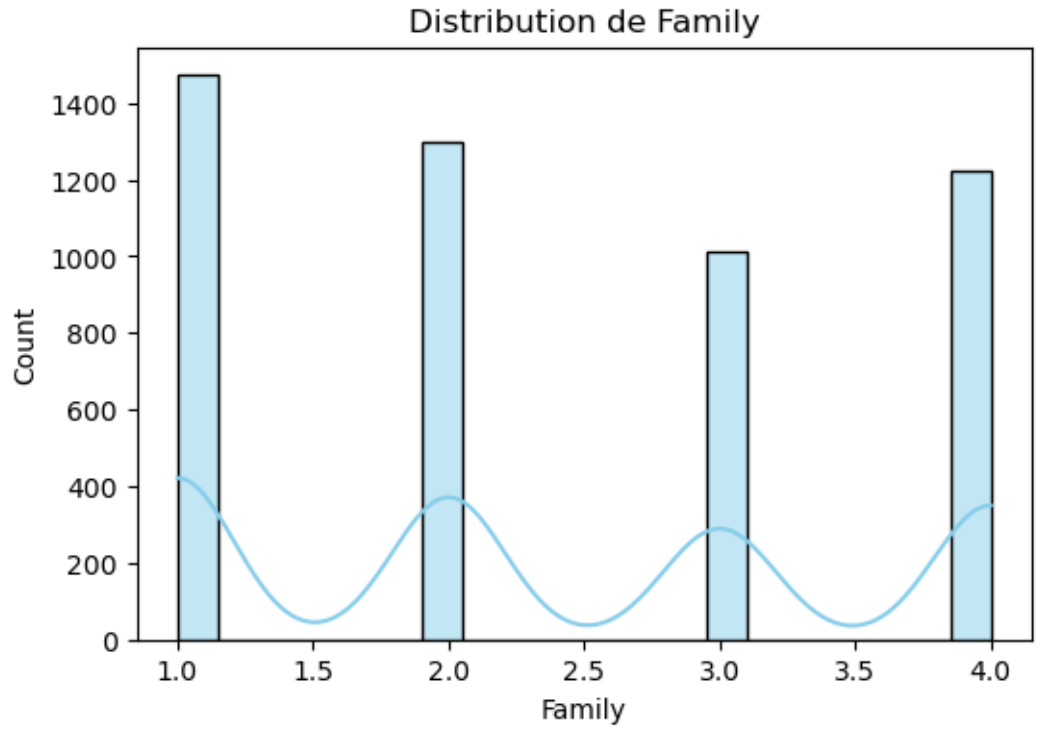


Source : Calcul de l'auteur, Base Kaggle

Le graphique représentant la distribution du *Log_Income* montre une répartition relativement étalée, avec une concentration importante des observations autour des valeurs comprises entre 3,5 et 5,0. La densité observée atteint son maximum entre 4,2 et 4,6, montrant que c'est dans cet intervalle que se situe la plus grande proportion de clients.

Les valeurs les plus faibles du *Log_Income* est autour de 2,0 et 3,0, qui correspondent aux revenus annuels les plus modestes, sont nettement moins fréquentes, ce qui confirme que peu de clients disposent de faibles revenus. À l'inverse, les valeurs très élevées (supérieures à 5,2) sont également moins représentées, traduisant une proportion limitée de clients ayant des revenus très élevés. La distribution prend ainsi une forme globalement asymétrique, légèrement étirée vers la droite, indiquant la présence de quelques clients aux revenus particulièrement importants.

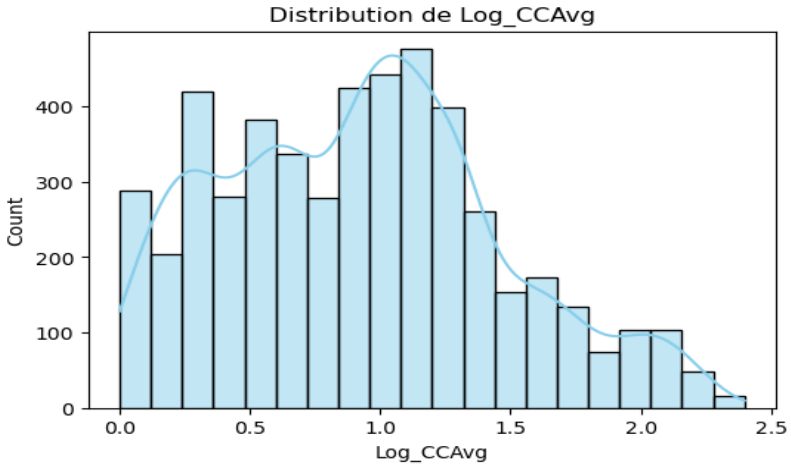
Figure 3 : repartions de clients selon la taille du ménage



Source : Calcul de l'auteur, Base Kaggle

La distribution des effectifs montre que les clients de la banque présentent des tailles familiales relativement variées. Le groupe le plus représenté est celui des clients ayant une seule personne dans le foyer (1 472 clients), suivi des ménages composés de quatre personnes (1 222 clients). Les familles de deux et trois personnes regroupent respectivement 1 296 et 1 010 clients. Cette structure suggère une clientèle équilibrée entre petits et moyens ménages, sans forte concentration sur une seule catégorie. Elle pourrait refléter une diversité de profils socio-économiques, ce qui est pertinent dans l'analyse du comportement d'acceptation de prêts, car la taille du ménage peut influencer les besoins financiers et la propension à recourir à des produits de crédit.

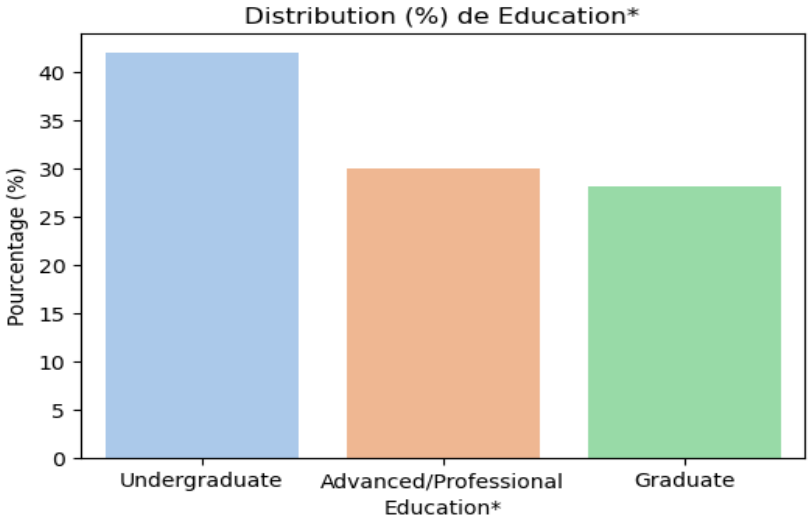
Graphique 4 : répartition des clients selon les dépenses mensuelles par cartes de crédits



Source : Calcul de l'auteur, Base Kaggle

La distribution de la variable **CCAvg**, représentant les dépenses mensuelles par carte de crédit, révèle une répartition hétérogène mais globalement concentrée autour des niveaux de dépenses modérés. Une forte densité d'observations apparaît entre **0,5 et 1,2**, indiquant que la majorité des clients réalisent des dépenses mensuelles moyennes comprises entre environ **1,6 et 3,3** le montant minimal observé. On observe également une décroissance progressive en direction des valeurs élevées (au-delà de 1,5), ce qui montre que les dépenses élevées sont moins fréquentes dans la clientèle. La présence d'un léger allongement de la queue droite (right-skewed) traduit que, même si une minorité de clients effectue des dépenses particulièrement importantes, ils sont peu représentés dans l'ensemble des données. Globalement, cette distribution suggère que la base de clients est principalement constituée d'individus ayant des dépenses modérées à moyennement élevées, tandis que les très gros dépensiers constituent un groupe marginal mais potentiellement stratégique pour la banque.

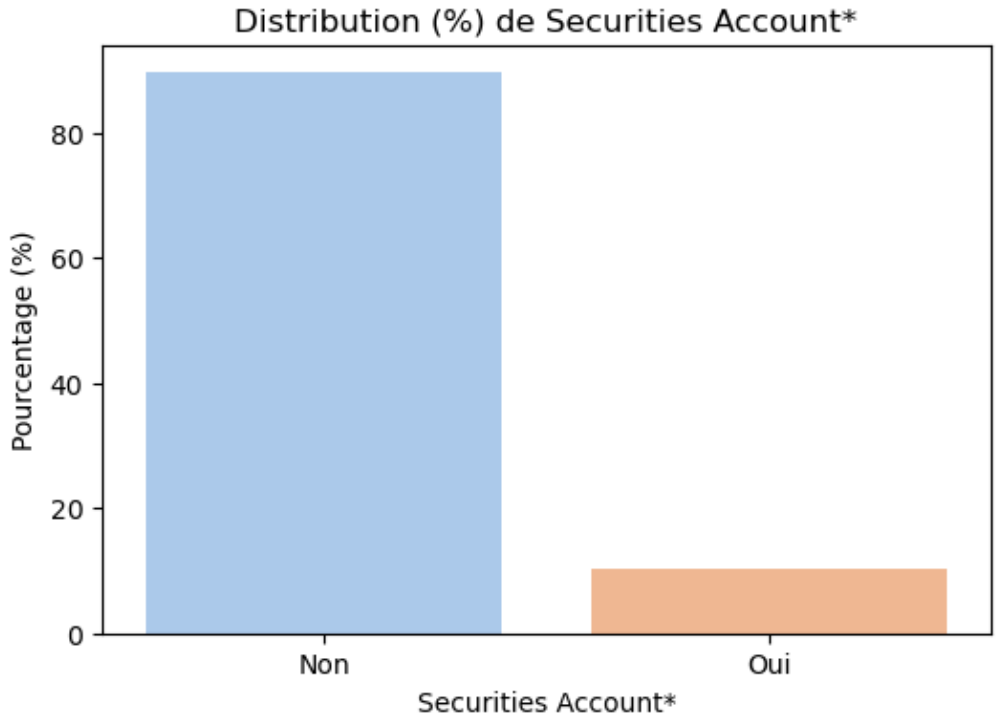
Figure 5 : Répartition des clients selon le niveau d'éducation



Source : Calcul de l'auteur, Base Kaggle

La répartition du niveau d'éducation des clients montre que les individus ayant un niveau *Undergraduate domine*, et représente 41,92 % des clients. Les diplômés de niveau *Graduate* constituent 28,06 % de la clientèle, traduisant une présence significative de clients ayant poursuivi des études supérieures intermédiaire. Les clients ayant un diplôme de *Advanced/Professional* représentent 30,02 % des clients, confirmant que près d'un tiers des clients ont un niveau d'éducation élevé. Dans l'ensemble, cette répartition montre une clientèle globalement instruite, avec une forte représentation des diplômés universitaires, ce qui peut favoriser leur compréhension et leur intérêt pour des produits financiers notamment les prêts personnels.

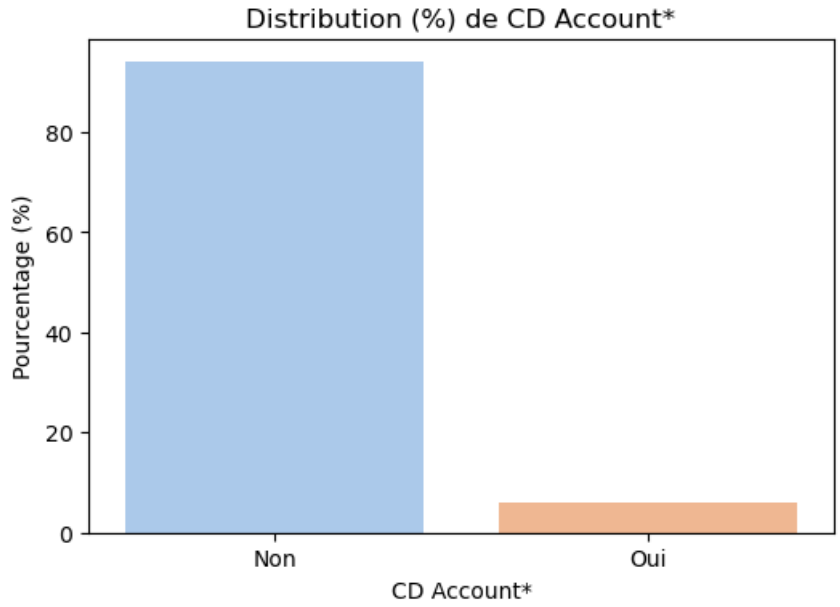
Figure 6 : répartition de clients détenant des compte-titres



Source : Calcul de l'auteur, Base Kaggle

La grande majorité des clients de la banque **ne détiennent pas de compte-titres**, représentant **89,56 %** de l'échantillon (soit 4 478 individus). Seuls **10,44 %** des clients déclarent posséder un compte-titres (522 individus). Cette répartition montre que l'investissement en titres reste peu répandu dans la clientèle, ce qui peut s'expliquer par un faible appétit pour les produits financiers à risque ou un manque d'information sur ces instruments. A partir de ce constat est important pour la banque dans l'optique d'éventuelles campagnes de promotion de produits d'investissement.

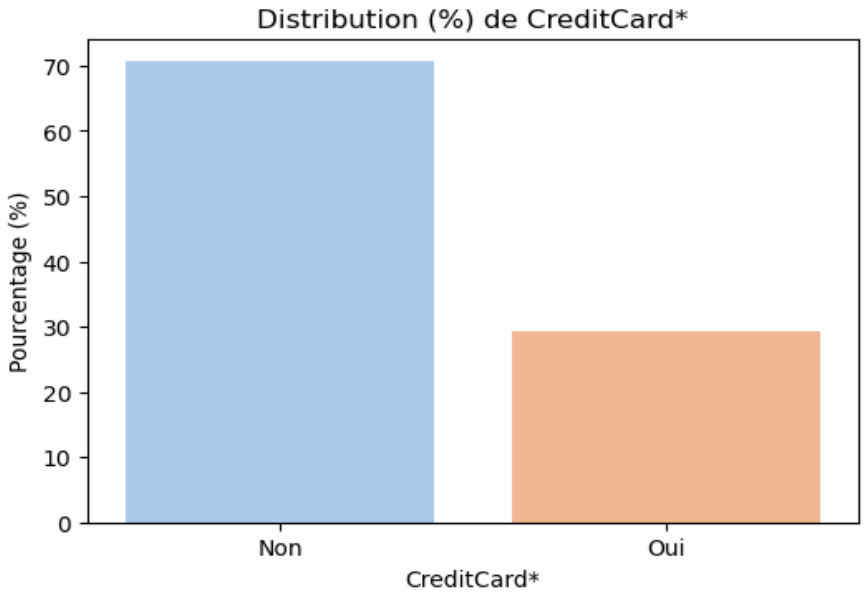
Figure 7 : Répartition des clients selon la détention des comptes à termes



Source : Calcul de l'auteur, Base Kaggle

La répartition des clients selon la détention d'un compte à terme (CD Account) montre une très forte dominance des non-détenteurs. En effet, 4 698 individus, soit 93,96 % de l'échantillon, ne possèdent pas de compte à terme, contre seulement 302 clients (6,04 %) qui en détiennent un. Cette distribution confirme que la variable cible est fortement déséquilibrée, avec une proportion de clients souscripteurs très faible. Cette situation est typique des produits financiers à engagement élevé et justifie l'utilisation de méthodes robustes pour traiter le déséquilibre lors des analyses prédictives.

Figure 8 : répartition des clients procédant une carte de crédits

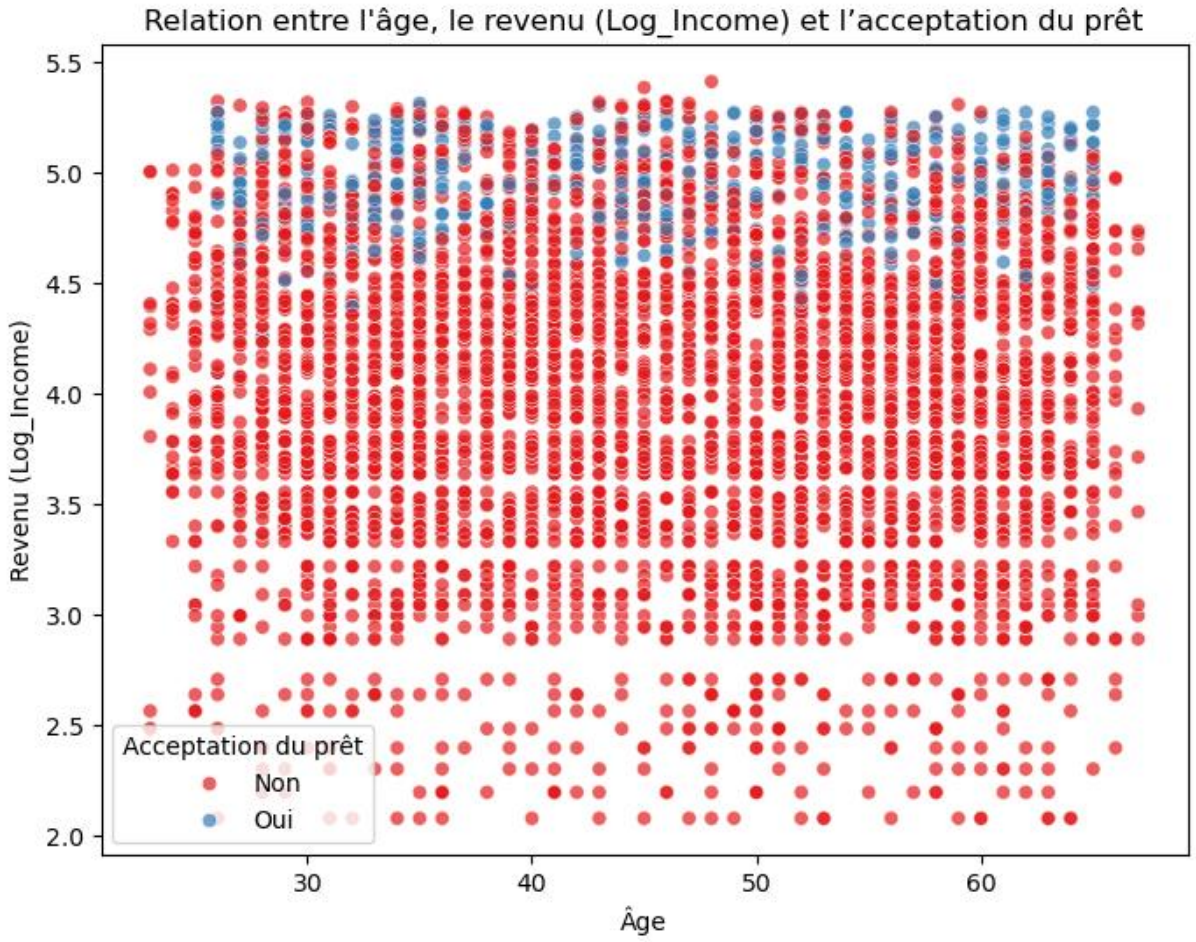


Source : Calcul de l'auteur, Base Kaggle

Les résultats montrent que **70,60 %** des clients ne possèdent pas de carte de crédit, contre **29,40 %** qui en possèdent une. Cette forte proportion de clients sans carte de crédit révèle une adoption relativement faible de ce service au sein de la clientèle. Cela peut indiquer un manque d'intérêt, un accès limité ou une stratégie commerciale peu orientée vers la promotion des cartes de crédit. À l'inverse, le tiers environ de clients qui disposent d'une carte de crédit représente un segment potentiellement plus engagé dans les services bancaires, susceptible d'avoir une activité financière plus diversifiée. Cette répartition est importante pour comprendre le profil financier des clients et orienter les actions marketing visant à accroître l'utilisation des produits de crédit.

3.3. Analyse bivarié

Graphique 9 : Répartition des clients ayants accepté les prêts ou non par âge et selon les revenus



Source : Calcul de l'auteur, Base Kaggle

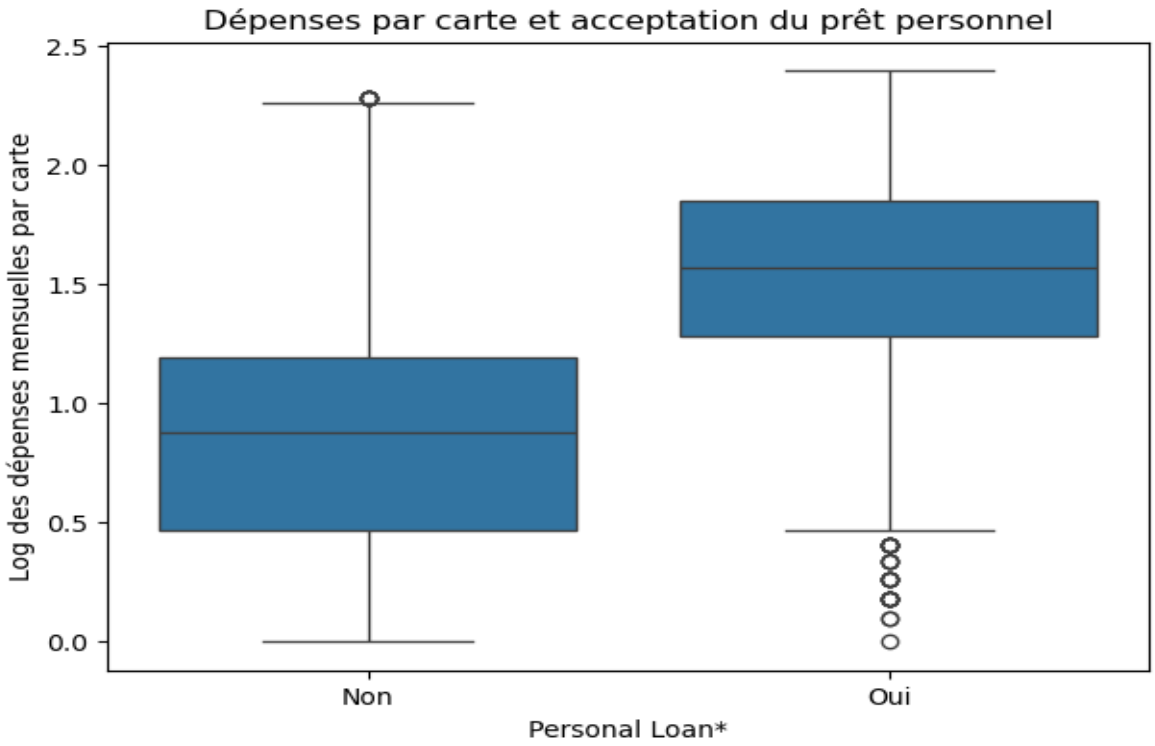
Le nuage de points ci-dessus met en relation l'âge des clients, leur niveau de revenu (approximé par le Log_Income) et leur décision d'accepter ou non un prêt personnel. Dans l'ensemble, le graphique révèle que la majorité des clients refusent le prêt (points rouges), quel que soit leur âge ou leur revenu, ce qui confirme le faible taux d'acceptation déjà observé dans la base.

On constate toutefois une tendance marquée : les clients qui acceptent le prêt se situent principalement dans les tranches de revenus les plus élevés, souvent au-dessus d'un Income de 4.7. Cela indique que le profil des clients intéressés par le prêt correspond généralement à des individus disposant d'un revenu plus conséquent.

En revanche, l'âge ne semble pas jouer un rôle discriminant notable : les points bleus sont dispersés entre 30 et 65 ans, sans concentration particulière.

Ce résultat suggère que **le revenu est un déterminant clé de l'acceptation du prêt**, tandis que l'âge exerce une influence beaucoup plus limitée.

Figure 10 : Répartition des dépenses par cartes de crédits selon l'acceptation des prêts personnels

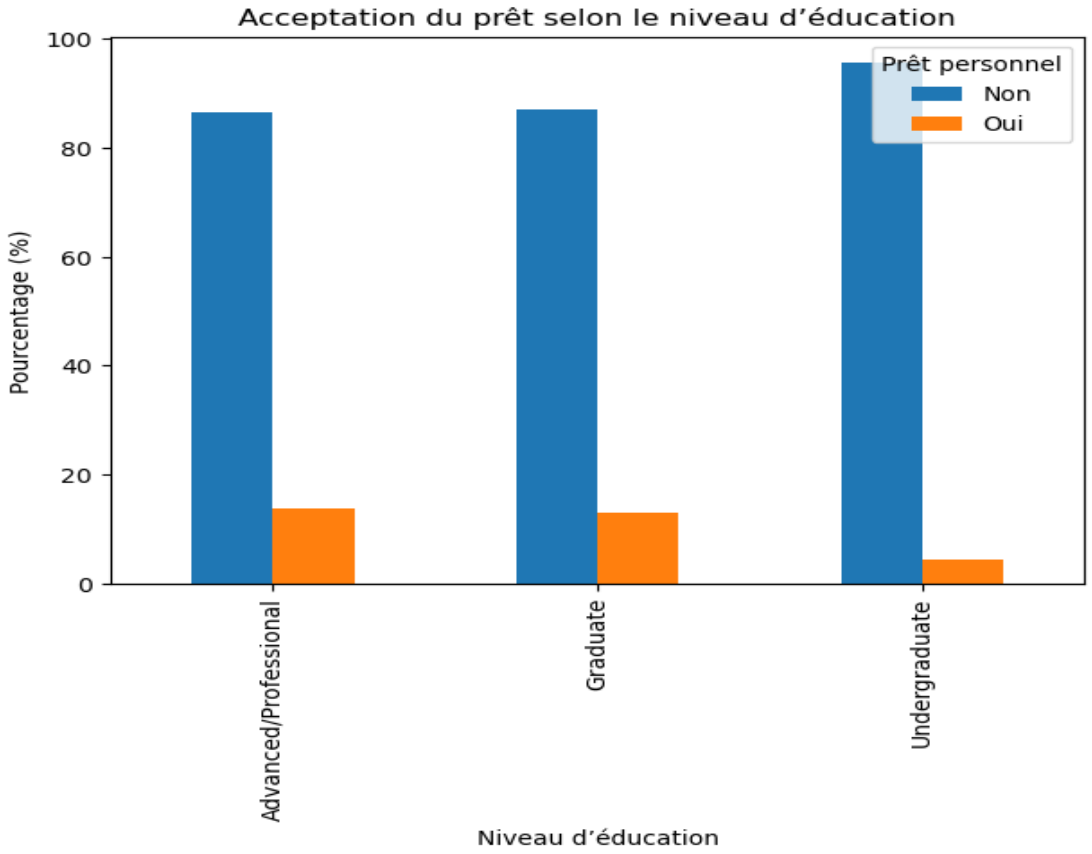


Source : Calcul de l'auteur, Base Kaggle

Le graphique met en évidence une différence nette entre les niveaux de dépenses mensuelles par carte bancaire (CCAvg) des clients ayant accepté un prêt personnel et ceux l'ayant refusé. Les individus n'ayant pas accepté le prêt présentent une médiane relativement faible de leurs dépenses mensuelles, traduisant un comportement de consommation plus modéré. À l'inverse, les clients ayant accepté le prêt affichent une médiane sensiblement plus élevée, témoignant d'un niveau de dépenses régulier et plus important.

La dispersion des valeurs est également plus marquée chez les accepteurs : leurs dépenses s'étendent à des niveaux beaucoup plus élevés, indiquant une activité financière plus dynamique. Cette différence structurelle suggère que les clients ayant des dépenses conséquentes, potentiellement indicative d'un style de vie plus coûteux ou d'une plus grande capacité financière, sont plus enclins à souscrire un prêt personnel.

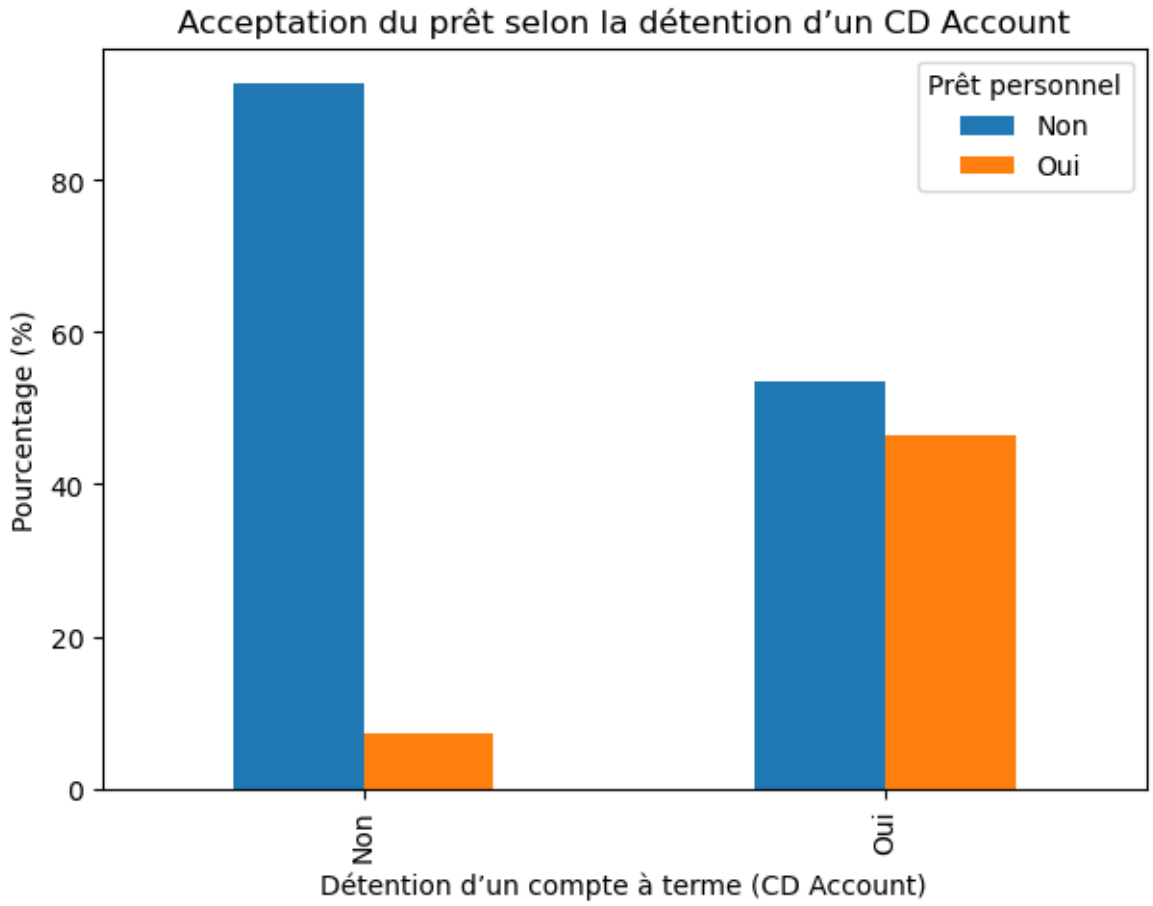
Figure 11 : Répartition de l'acception des prêts selon le niveau d'éducation



Source : Calcul de l'auteur, Base Kaggle

Le graphique montre une relation nette entre le niveau d'éducation des clients et leur propension à accepter un prêt personnel. Le diagramme illustre un lien clair entre le niveau d'instruction des clients et leur tendance à accepter un crédit personnel. Il est observé que les clients détenant un diplôme supérieur ou professionnel, tout comme ceux ayant atteint le niveau graduate, affichent des taux d'approbation nettement plus élevé, affichent des taux d'approbation nettement plus comparativement aux comparativement aux clients de niveau undergraduate. clients de niveau premier cycle. Pour les Deux premiers segments, le taux d'acceptation de l'offre de crédit par les clients varie entre 13 et 15 %, alors que celui des clients ayant un diplôme de premier cycle est de 5 %. Ce qui veut dire que le niveau d'instruction est un élément important dans la prise de décision en ce qui concerne le un prêt personnel.

Figure 12 : Répartition de l'acception de prêts selon la détention d'un compte à terme



Source : Calcul de l'auteur, Base Kaggle

Le graphique montre une différence très marquée dans l'acceptation du prêt personnel selon que le client détienne ou non un compte à terme (CD Account). Parmi les clients ne disposant pas d'un CD Account, la majorité refuse l'offre de prêt, tandis qu'une faible proportion seulement l'accepte. Cette tendance traduit une faible appétence au crédit parmi les clients ayant un profil bancaire plus simple ou moins engagé.

En revanche, chez les clients détenant un CD Account, le comportement est profondément différent : la proportion d'acceptation du prêt personnel est nettement plus élevée, approchant presque la moitié des détenteurs de ce type de produit bancaire. Cette dynamique suggère que les clients ayant déjà un CD Account entretiennent une relation financière plus diversifiée et sont davantage partant à souscrire de nouveaux produits qu'on leurs proposera, y compris des prêts personnels.

- **Test de mann-whitney**

Le test de Mann-Whitney U a permis d'évaluer si les distributions des variables diffèrent significativement entre les clients ayant accepté un prêt personnel et ceux l'ayant refusé. Les résultats obtenus sont les suivants :

Tableau 4 : test de mann-whitney

	p_value
Log_Income	1.035175e-203
Log_CCAvg	3.599450e-109
Experience	5.720973e-01

Source : Calcul de l'auteur, Base Kaggle

Income : La p-value est extrêmement faible, proche de zéro, ce qui indique une différence statistiquement très significative entre les revenus de ceux qui acceptent et de ceux qui refuse le prêt. Les clients qui acceptent le prêt ont un revenu nettement différent (et généralement plus élevée) que ceux qui refusent. Cela nous donne le rôle majeur du revenu comme déterminant clé de l'acceptation du prêt.

CCAvg : Comme pour le revenu, la p-value est aussi faible, indiquant également une différence très significative entre les deux groupes. Les clients qui dépensent davantage via leur carte bancaire ont un comportement distinct et sont plus susceptibles d'accepter un prêt. Cela montre aussi que le niveau d'activité financière mesuré par les dépenses sur carte est un facteur discriminant important.

Experience : La p-value est largement supérieure à 5 %, ce qui signifie absence de différence significative entre les deux groupes. Le nombre d'années d'expérience professionnelle n'a pas d'impact notable sur la probabilité d'accepter un prêt. La variable n'est pas discriminante dans la décision d'acceptation.

- **Test de chi2**

Le test du Chi-deux a été utilisé pour analyser l'existence association statistique entre l'acceptation du prêt personnel (variable cible binaire) et plusieurs caractéristiques qualitatives des clients. Les p-values obtenues sont les suivantes :

Tableau 5 : Tableau des tests de Chi 2

	p_value
Education	6.991474e-25
CreditCard	8.843861e-01
Online	6.928600e-01
Securities Account	1.405150e-01
CD Account	7.398298e-110

Source : Calcul de l'auteur, Base Kaggle

Niveau d'éducation : La p-value est très faible, cela indique l'existence d'une association très significative entre le niveau d'éducation et le fait d'accepter un prêt. Le niveau d'éducation a une forte influence sur la probabilité d'accepter de faire un prêt. Les clients ayant un niveau d'instruction élevé sont davantage enclins à accepter le produit bancaire.

- **Possession d'une carte de crédit**

La p-value est également très élevée cela indique qu'il n'y a pas **d'association significative** entre la possession d'une carte de crédit et le fait d'accepter un prêt. Qu'un client ai une carte de crédit ou pas, cela n'agit pas sur la décision d'accepter un prêt.

- **Usage des services en ligne (Online)**

La p-value élevée indique qu'il n'y a également pas d'association significative. Les clients utilisant les services bancaires en ligne n'a rien avoir avec le fait d'accepter un prêt.

- **Possession d'un compte titres (Securities Account)**

La p-value dépasse largement 5 %, indiquant absence d'association entre la détention d'un compte titres et l'acceptation du prêt. Les clients ayant un profil d'investissement (compte titres) ne sont ni plus ni moins susceptibles d'accepter le prêt.

- **Détention d'un CD Account**

La p-value est extrêmement faible, ce qui montre une extrême **association significative**. Les clients ayant un compte à terme (CD Account) sont beaucoup plus susceptibles d'accepter le prêt. Cela démontre un niveau d'engagement bancaire nettement plus fort et un profil plus favorable au crédit.

- **Corrélation de Spearman**

Le tableau ci-dessous présente les coefficients de corrélation de Spearman entre la variable cible «Personal Loan » et les variables explicatives quantitatives du modèle :

Tableau 6 : test de corrélation de Spearman

Variable	Spearman
Personal Loan	1.000000
Log_Income	0.430739
Log_CCAvg	0.313957
Log_Mortgage	0.063294
Family	0.061757
Experience	-0.007991
Age	-0.008151

Source : Calcul de l'auteur, Base Kaggle

L'analyse de corrélation de Spearman révèle que l'acceptation d'un prêt personnel est associée aux caractéristiques financières des clients. En effet, le revenu ($\rho = 0,43$) présente la corrélation positive la plus élevée avec la variable *Personal Loan*, confirmant que les clients à revenu élevé sont significativement plus susceptibles d'accepter faire un prêt. Les dépenses moyennes par carte bancaire ($\rho = 0,31$) montrent également une corrélation positive modérée, indiquant que les clients ayant une activité financière plus dynamique manifestent davantage d'intérêt pour les produits de crédit. En revanche, les variables telles que *Mortgage* et *Family* présentent des corrélations positives mais très faibles, traduisant un lien limité avec la décision d'acceptation. Enfin, l'âge et l'expérience professionnelle affichent des corrélations presque inexistantes et

montrant qu'ils n'influencent pratiquement pas le comportement d'acceptation d'un prêt. Dans l'ensemble, ces résultats confirment que la capacité financière et le niveau d'activité bancaire constituent les déterminants majeurs d'accepter l'offre de prêt.

3.4. Analyse multivarié / Modélisation

3.4.1. Régression légistique

La régression logistique est utilisée afin de modéliser la probabilité qu'un client accepte une offre de prêt personnel proposée par la banque.

Ce modèle est particulièrement adapté lorsque la variable dépendante est binaire, comme dans notre cas où la variable *Personal Loan* prend la valeur 1 si le client accepte le prêt et 0 sinon.

Tableau 7 : séparation de train / test

Métrique	Valeur
Taille train	(3500, 20)
Taille test	(1500, 20)
Taux d'acceptation (train)	0.096
Taux d'acceptation (test)	0.096

Source : Calcul de l'auteur, Base Kaggle

Afin de tenir compte du déséquilibre des classes (seulement 9,6 % des clients acceptent le prêt), une pondération des classes a été appliquée lors de l'estimation du modèle.

- **Évaluation des performances du modèle**
- **Matrice de confusion**

	Prédit : Refus (0)	**Prédit : Acceptation (1)**
Réel : Refus (0)	1211 (Vrais négatifs)	145 (Faux Positifs)
Réel : Acceptation (1)	10 (Faux négatifs)	134 (Vrais Positifs)

Source : Calcul de l'auteur, Base Kaggle

Le modèle identifie correctement **93 %** des clients ayant effectivement accepté un prêt, ce qui est essentiel dans une logique de ciblage marketing. Le nombre de clients intéressés non détectés (faux négatifs) reste très faible.

- **Indicateurs de performances**

Indicateurs	**Valeur**
Accuracy	90%
Precision (classe 1)	48%
Recall (classe 1)	**93%**
F1-score (classe 1)	63%
AUC	**0,975**

Source : Calcul de l'auteur, Base Kaggle

Le **Recall élevé** indique que le modèle capture efficacement la majorité des clients susceptibles d'accepter un prêt. Le **F1-score** traduit un bon équilibre entre précision et rappel. L'**AUC de 0,975** révèle une capacité de discrimination exceptionnelle entre clients acceptant et refusant le prêt.

- **Analyse des facteurs explicatifs (Odds Ratios)**

L'analyse des coefficients est effectuée à l'aide des **odds ratios**, qui mesurent l'effet marginal de chaque variable sur la probabilité d'acceptation du prêt.

- **Principaux déterminants positifs**

Variable	Odds Ratio	Interprétation
Log_Income	48,97	Le revenu est le facteur le plus déterminant
Education (niveau avancé)	3,43	Les clients plus instruits acceptent davantage
Education (niveau supérieur)	3,14	Effet positif significatif
CD Account	2,69	Relation bancaire renforcée
Family	1,91	Besoins financiers plus importants
CCAvg	1,72	Comportement de consommation actif

Source : Calcul de l'auteur, Base Kaggle

- Facteurs réduisant la probabilité d'acceptation

Variable	Odds Ratio	Interprétation
Age	0,87	Les clients plus âgés sont moins enclins
Online	0,71	Préférence pour des services non crédit
CreditCard	0,61	Déjà bien équipés en produits bancaires
Securities Account	0,7	Profil plus prudent et moins endetté

Source : Calcul de l'auteur, Base Kaggle

Validation des hypothèses de recherche

Les résultats obtenus permettent de confirmer empiriquement plusieurs hypothèses formulées au départ :

- H1 : Les clients à revenu élevé sont plus susceptibles d'accepter un prêt personnel : validée.
- H2 : Le niveau d'éducation influence positivement l'acceptation du prêt : validée.
- H3 : La relation bancaire influence l'acceptation du prêt : (effet positif du CD Account). Les dépenses par carte de crédit sont un indicateur significatif : validée.

Implications managériales pour la banque

Les résultats suggèrent que les campagnes de marketing doivent cibler en priorité :

- les clients à revenus élevés,
- les clients disposant d'un niveau d'éducation supérieur,
- les détenteurs de comptes à terme,
- les clients ayant des dépenses élevées par carte bancaire.

Cette stratégie permettrait d'améliorer le taux de conversion tout en réduisant les coûts liés à des campagnes peu ciblées.

La régression logistique fournit un modèle performant, interprétable et directement exploitable pour la prise de décision marketing de **la banque**.

Compte tenu de ses excellentes performances, ce modèle servira de référence pour la comparaison avec les modèles plus complexes présentés dans la suite de l'étude.

3.4.2.Arbre de décision (Decision Tree Classifier)

Bien que la régression logistique offre une excellente capacité prédictive, elle repose sur une relation linéaire entre les variables explicatives et le logit de la probabilité. Afin de capturer d'éventuelles relations non linéaires et interactions entre variables, le modèle d'arbre de décision est estimé dans la suite.

Présentation du modèle

Le modèle de classification non paramétrique qu'est l'arbre de décision qui divise la population de clients en sous-groupes homogènes basés sur leurs caractéristiques.
Dans chaque point de décision de l'arbre, une variable explicative est choisie pour optimiser la distinction entre les classes d'acceptation et de refus du prêt.
Dans cette recherche, le critère de séparation choisi est l'indice de Gini, qui est défini comme suit :

$$Gini(t) = 1 - \sum_{i=1}^{C} p(i \mid t)^2$$

où $p(\, i \mid t\,)$ représente la proportion d'observations de la classe *i* dans le nœud *t*.

Un L'arbre de décision supervisé a été ajouté afin de prédire la probabilité d'approbation d'une carte bancaire (variable binaire Personal_Loan).

L'apprentissage du Le modèle a été réalisé sur un ensemble de données d'entraînement, suivi d'une évaluation sur un ensemble de données de test.

Les paramètres principaux du modèle sont :

- criterion = « gini »

- max_depth = 5
- min_samples_leaf = 50
- class_weight = « balanced »
- random_state = 42

Ces choix permettent :

- d'éviter le sur-apprentissage,
- de prendre en compte le déséquilibre de classes,
- de garantir une reproductibilité du modèle.

- **Matrice de confusion du modèle Arbre de décision**

	Prédit : Refus (0)	**Prédit : Acceptation (1)**
Réel : Refus (0)	**1242** (Vrais négatifs)	**114** (Faux positifs)
Réel : Acceptation (1)	1 (Faux négatifs)	143 (Vrais positifs)

Source : Calcul de l'auteur, Base Kaggle

- Le modèle identifie correctement **1242 non-accepteurs** et **143 accepteurs**.
- Il commet **114 faux positifs** et seulement **1 faux négatif**, ce qui montre qu'il détecte quasi parfaitement les accepteurs.

- **Classification report**

Indicateur	**Valeur**
Accuracy	92%
Precision (classe 1)	56%
Recall (classe 1)	**99%**
F1-score (classe 1)	71%
AUC	0,991

Source : Calcul de l'auteur, Base Kaggle

Concernant la classe 0, les résultats sont remarquables (F1 = 0.96) tandis que pour la classe 1, le modèle identifie presque l'intégralité des accepteurs véritables (rappel = 0.99), mais cela entraîne plus de fausses alertes (précision=0.56).

Un une précision globale (Accuracy) de 92 % indique une performance supérieure.

Le modèle favorise l'identification des accepteurs (classe 1), une approche privilégiée dans le secteur bancaire où l'on privilégie l'analyse des profils à forte probabilité d'acceptation. Il détecte avec précision 99 % des clients qui ont réellement accepté l'offre (rappel = 0,99). Cela implique que le modèle détecte presque tous les clients qui manifestent un intérêt réel, ce qui est particulièrement intérêt bénéfique ce qui est particulièrement bénéfique dans une perspective de ciblage marketing où l'ambition est de ne laisser passer aucun prospect potentiel.

Dans une perspective de marketing ciblage où l'ambition est de ne laisser passer aucun prospect potentiel.

Le nombre de clients intéressés non détectés (faux négatifs) est extrêmement faible : 1 seul cas sur 144.

Ceci améliore la capacité du modèle à réduire les occasions manquées. Le modèle détecte également avec précision 92 % des individus qui ne vont pas acceptés, bien que le nombre de faux positifs (114) soit plus important que dans le modèle logistique. En d'autres termes, l'arbre décisionnel à une propension à identifiant plus de clients comme étant intéressés même s'ils ne le sont pas véritablement, ce qui peut faire grimper les frais de prospection.

- **Importance des variables**

Rang	Variable	Importance	Interpretation
1	Log_Income	0.682	Les clients avec un revenu plus élevé sont nettement succeptible d'accepter l'offre.
2	Education_2	0.151	Les diplômes plus élevés augmentent la propension à accepter.
3	Log_CCAvg	0.068	Les clients ayant une forte activité sur carte sont plus susceptibles d'être intéressés.
4	Education_3	0.054	Les diplômes plus élevés augmentent la propension à accepter.
5	Family	0.045	Les clients avec plus de charges familiales semblent plus intéressés.
	Autres	≈ 0	Aucune importance

Source : Calcul de l'auteur, Base Kaggle

L'arbre ne prend pas en compte certaines variables : Âge, Expérience, Prêt hypothécaire, Compte de titres, Compte de certificat de dépôt.

Elles ne figurent pas parmi les points de découpage, car elles n'améliorent pas l'information mutuelle.
L'arbre de La décision démontre des performances globales excellentes, affichant une précision de 92% et un AUC frôlant l'excellence (0,991). Il identifie presque tous les récepteurs, ce qui est essentiel pour une stratégie de marketing ciblé efficace.

Le modèle indique que les facteurs les plus influents dans l'acceptation bancaires d'un prêt bancaire sont liées aux caractéristiques socio-économiques des clients, telles que le revenu, le niveau d'études et les dépenses effectuées via la carte de crédit

3.4.3. Random Forest

La Random Forest est une technique d'apprentissage par agrégation qui regroupe plusieurs arbres de décision dans le but d'optimiser la stabilité et l'efficacité de la prédiction.
Elle introduit deux sources d'incertitude :

- Échantillonnage bootstrap : chaque arbre est élaboré à partir d'un échantillon sélectionné avec remise ;

- Choix aléatoire de variables à chaque nœud.
  Ces deux méthodes offrent la possibilité de :
- diminuer la variance ;
- éviter le surajustement ;
- renforcer la puissance prédictive, particulièrement avec des données sous forme de tableaux telles que celles des clients dela banque

Dans le cadre de cette recherche, une forêt aléatoire de 500 arbres a été formé en appliquant une pondération des classes afin de remédier au déséquilibre (environ 9,6 % d'approbation des prêts).

- **Matrice de confusion**

	Prédit : Refus (0)	**Prédit : Acceptation (1)**
Réel : Refus (0)	**1319** (Vrais négatifs)	**37** (Faux positifs)
Réel : Acceptation (1)	4 (Faux négatifs)	140 (Vrais positifs)

Source : Calcul de l'auteur, Base Kaggle

Le modèle a correctement identifié 1319 clients parmi les 1356 qui ont rejeté le prêt, ce qui représente un taux de classification correct de 97 %. Il a correctement identifié 140 clients sur 144 qui ont accepté le prêt, le taux de rappel (sensibilité) est de 97 %. Le nombre de faux négatifs est très bas (4), ce qui est crucial pour la banque pour éviter de passer à côté de clients intéressés.
Jusqu'à présent, la Random Forest est le modèle qui offre le meilleur équilibre entre les deux classes.

- **Classification report**

Indicateur	**Valeur**
Accuracy	97%
Precision (classe 1)	79%
Recall (classe 1)	**97%**
F1-score (classe 1)	87%
AUC	0,996

Source : Calcul de l'auteur, Base Kaggle

Le le modèle reconnaît avec précision que 79 % des clients prédits comme « acceptant » le sont réellement. Il détecte 97% des clients potentiellement intéressés par un crédit. Le F1- score de la classe 1 = 0,87, démontre une performance remarquable pour une classe minoritaire. L'excellente performance globale (AUC=0,996) démontre une aptitude remarquable à distinguer les clients qui acceptent de ceux qui refuser le prêt. C'est un modèle particulièrement performant pour des actions de marketing ciblées, car il réduit au minimum les erreurs graves.

- **Importance des variables**

Rang	Variable	Importance	Interprétation
1	Log_Income	0.507	Plus le revenu est élevé, plus la probabilité d'accepter un prêt augmente.
2	Log_CCAvg	0.225	Les dépenses par carte sont un excellent indicateur de solvabilité.
3	Family	0.052	Les clients ayant une famille plus grande ont une probabilité plus élevée d'accepter un prêt.
4	CD Account_1	0.048	Les détenteurs de dépôts à terme montrent un intérêt plus marqué pour les produits financiers.
5	Education_3	0.043	Le niveau d'éducation accroît l'ouverture aux produits bancaires.
6	Education_2	0.040	Le niveau d'éducation accroît l'ouverture aux produits bancaires.

Source : Calcul de l'auteur, Base Kaggle

Le modèle est solide, constant et particulièrement bien adapté aux données déséquilibrées. Actuellement, il est le plus performant avec une capacité prédictive exceptionnelle (AUC = 0,996) et une détection presque parfaite des clients potentiellement intéressés par un prêt. Elle représente donc l'instrument le plus sûr pour un ciblage marketing à la fois marketing efficace à la fois efficace et précis.

3.4.4. XGBoost (Extreme Gradient Boosting)

Une variante perfectionnée du Gradient Boosting (XGBoost), particulièrement efficace pour les données tabulaires telles que celles de la Banque FinaTrust.
Il combine plusieurs arbres peu profonds, et chaque nouvel arbre rectifie les erreurs de son précédent. En intégrant une régularisation, il permet d'éviter trop apprentissage tout en garantissant une grande précision.

- **Matrice de confusion**

	Prédit : Refus (0)	**Prédit : Acceptation (1)**
Réel : Refus (0)	**1350** (Vrais négatifs)	**6** (Faux positifs)
Réel : Acceptation (1)	**7** (Faux négatifs)	**137** (Vrais positifs)

Source : Calcul de l'auteur, Base Kaggle

Le modèle identifie 95 % des clients qui sont véritablement intéressés par le prêt (Rappel = 0,95). Les erreurs critiques (faux négatifs) sont particulièrement non signifiantes : Sur 144 clients, 7 n'ont pas été pris en compte ; les faux positifs sont presque inexistants, ce qui diminue les dépenses en actions marketing superflues.

Ce modèle fait preuve d'une excellente compréhension des deux classes, ce qui est assez inhabituel avec des données déséquilibrées telles que celles liées au prêt.

- **Rapport de classification**

Indicateur	Valeur
Accuracy	99%
Precision (classe 1)	96%
Recall (classe 1)	95%
F1-score (classe 1)	95%
AUC	0,998

Source : Calcul de l'auteur, Base Kaggle

Avec avec une précision globale de 99% et un F1- score considérable pour les clients qui acceptent, le modèle atteint un équilibre parfait entre précision et rappel. C'est précisément ce que la banque recherche pour viser les clients susceptibles d'accepter un prêt. Une AUC (0,998) se rapproche de 1 démontrer une aptitude presque parfaite à différencier les deux classes.

- **Importance des variables**

Variable	Importance	Interprétation
Log_Income	0.2729	Le revenu (Log_Income) reste le facteur prédictif le plus important.
Education_2	0.1780	Le niveau d'éducation joue un rôle majeur dans la décision d'accepter ou non un prêt.
Education_3	0.1649	Le niveau d'éducation joue un rôle majeur dans la décision d'accepter ou non un prêt.
CD Account_1	0.1018	Le fait de posséder un CD Account augmente fortement la probabilité d'acceptation.
Log_CCAvg	0.0747	Les clients ayant une forte activité sur carte sont plus susceptibles d'être intéressés.
Family	0.0742	Les clients avec plus de charges familiales semblent plus intéressés.
Log_Mortgage	0.0271	Mortgage influencent également le comportement client.

Source : Calcul de l'auteur, Base Kaggle

3.4.5. Choix final du meilleur modèle

Suite à l'évaluation rigoureuse de divers modèles prédictifs, notamment la régression logistique, l'Arbre de décision, le Random Forest et le XGBoost, il est désormais crucial d'effectuer une sélection finale après avoir procédé à une comparaison synthétique basée sur des indicateurs solides (Accuracy(exactitude), Précision, Rappel, AUC, Score F1). Cette la mise en parallèle permettra de déterminer le modèle offrant la performance optimale pour prédire l'acceptation d'un crédit personnel.

Tableau 8 : Comparaison finale des modèles

Modèle	Accuracy	Recall (classe 1)	F1-score (classe 1)	AUC	Commentaire
Logistic Regression	0,90	0,93	0,63	0,975	Bonne interprétation, mais faible précision pour la classe 1
Decision Tree	0,92	0,99	0,71	0,991	Très bon rappel, mais modèle instable
Random Forest	0,97	0,97	0,87	0,996	Excellent compromis stabilité + performance
XGBoost	0,99	0,95	0,95	0,999	Modèle le plus performant

Source : Calcul de l'auteur, Base Kaggle

Apres analyse a travers les indicateurs de performance de chaque modèles, le meilleur modèle est clairement : XGBoost

Justifications :

- Amélioration de la précision générale (Précision) (99 %)
- Parfait équilibre entre la Précision et le Racall pour la classe 1 : crucial pour repérer les clients concernés par l'offre de crédit.
- AUC presque parfaite (0,999).
- Modèle solide, stable et très efficace sur des données sous forme de tableau.
- Il y a très peu d'erreurs majeures (seulement 7 faux négatifs) et les faux positifs sont presque nuls en comparaison avec d'autres modèles : cela permet à la banque de diminuer les dépenses liées aux actions marketing superflues.

Le modèle XGBoost, qui présente la performance prédictive la plus optimale, est l'instrument le plus sûr pour peaufiner la stratégie marketing de la banque.
En identifiant avec précision les clients étant réceptifs à un emprunt, il contribue à optimiser l'efficacité des opérations commerciales tout en minimisant les dépenses liées aux ciblages non pertinents.

Cette partie a permis d'examiner l'efficacité de quatre modèles de classification, à savoir la Régression Logistique, l'Arbre de Décision, le Random Forest et le XGBoost ; dans l'objectif d'estimer la chance qu'un client accepte un prêt. Les résultats indiquant que les modèles fondés sur les arbres, notamment XGBoost, procurent les performances les plus optimales, avec une précision, un rappel et une AUC remarquablement élevée. Ces résultats révèlent leur aptitude à saisir des liens complexes entre les attributs des clients et leur comportement en matière d'acceptation de crédit.

Malgré les caracteristiques supérieure de la régression logistique, ses résultats restent moins performants que ceux des modèles non linéaires. En revanche, XGBoost, suivi de près par Random Forest, se révèle comme le modèle le plus solide et digne de confiance pour améliorer la stratégie marketing de la banque. Sa capacité de pourvoir identifier avec précision les clients susceptibles d'accepter un emprunt contribue non seulement à optimiser l'efficacité des

campagnes de prospection, mais également à minimiser les dépenses liées à une orientation inappropriée.

3.4.6. Recommandations

Recommandations stratégiques pour banque

- **Mettre en place un ciblage marketing intelligent basé sur XGBoost**

Puisque XGBoost est le modèle le plus performant :

- Utiliser ses prédictions pour établir une liste prioritaire des clients les plus susceptibles d'accepter la proposition d'un prêt.
- Allouer les efforts marketing (SMS, appels, visites, emails) en priorité vers les clients avec score élevé.
- Réduire les coûts de prospection en limitant les campagnes vers les clients classés comme « peu susceptibles » ou « pas susceptibles ».

- **Cibler les segments à plus forte propension**

Le modèle montre que certains profils ont une probabilité **nettement plus élevée** d'accepter un prêt :

- **Clients à revenu élevé (Income)**
- Proposer des prêts premium, des taux différenciés, des conditions flexibles.
- Créer des offres personnalisées (« High Income Loan Pack »).
- **Clients avec niveau d'éducation supérieur**
- Utiliser des approches marketing plus techniques (argumentaires chiffrés, simulations financières).
- **Clients actifs sur carte bancaire (CCAverg élevé)**
- Offrir des prêts conso rapides liés aux habitudes de dépense.
- **Clients détenteurs d'un CD Account**
- Proposer des prêts garantis sur dépôt, avec taux préférentiel.
- **Clients avec charges familiales élevées**
- Mettre en avant les prêts scolaires, prêts d'équipement ménager, etc.

- **Automatiser le scoring d'éligibilité pour accélérer l'octroi des prêts**

Mettre en place un système automatisé basé sur XGBoost pour :

- Scorer chaque nouvelle demande ou mise à jour client.
- Proposer une pré-qualification instantanée du type :« Vous êtes éligible à un prêt jusqu'à X FCFA »

- Réduire les délais opérationnels pour les conseillers bancaires.

    - **Optimiser la gestion des campagnes commerciales**

Utiliser les scores fournis par le modèle pour :

- **Segmenter** la base clients en 3 groupes :
    - **Groupe A : Forte probabilité (> 0,7)** : prioritaire
    - **Groupe B : Probabilité moyenne (0,4 - 0,7)** : campagnes ciblées conditionnelles
    - **Groupe C : Faible probabilité (< 0,4)** : ne pas cibler pour limiter les coûts
- Ajuster le budget marketing en fonction des retours prédits.

    - **Concevoir des offres adaptées aux groupes moins susceptibles**

Pour les segments ayant une faible probabilité d'accepter le prêt :

- **Clients âgés**
- Proposer des prêts adaptés : santé, retraite, etc.
- **Clients prudents (Securities Account)**
- Mettre en avant des garanties plus fortes et plus de transparence.
- **Clients très digitalisés (Online)**
- Utiliser un marketing 100 % digital :
    - chatbots intelligents
    - simulateurs de prêts en ligne
    - push notifications personnalisées

    - **Renforcer la relation bancaire en amont**

Puisque CD Account et l'activité carte sont déterminants :

- Encourager les clients à ouvrir un CD Account (compte à terme).
- Promouvoir l'utilisation de la carte pour augmenter l'éligibilité.
- Mettre en place des programmes de fidélisation liés aux prêts.

    - **Mettre en place un tableau de bord de suivi**

Créer un tableau de bord pour suivre :

- Score moyen des clients
- Conversion par segment
- Taux de faux négatifs (prospects perdus)

- Taux de faux positifs (actions marketing inutiles)
- Gains financiers générés grâce au modèle
    - **Réviser périodiquement le modèle**
- Rafraîchir le modèle tous les **6 à 12 mois**.
- Ajouter de nouvelles variables (ex : historique de prêts, flux de comptes, comportement digital).
- Tester systématiquement si XGBoost reste le meilleur ou si un autre modèle devient plus performant.

Conclusion générale

Cette etude était d'élaborer un modèle prédictif performant afin d'identifier les profils clients les plus enclins à accepter une offre de prêt, dans le but d'optimiser la stratégie marketing et commerciale de croia ecrire . À l'investigation, combinée aux techniques de préparation des données et à l'évaluation de plusieurs algorithmes, le modèle XGBoost a démontré la meilleure précision prédictive en raison de sa robustesse et de sa gestion performante des interactions non linéaires entre les variables. Les résultats montrent que des éléments tels que le revenu, le niveau d'éducation, la détention d'un compte d'épargne, l'usage de la carte de crédit et les dépenses du ménage influencent significativement la probabilité d'obtenir un prêt.

Ces informations constituent un socle pertinent pour optimiser la segmentation de la clientèle, peaufiner les campagnes de marketing et approfondir la prise de décision basée sur les données.

La mise en œuvre d'un système de scoring automatisé, associé à une stratégie marketing différenciée, permettra à la banque de mieux cibler les clients à forte propension, de réduire les coûts opérationnels et d'accroître le taux de conversion des offres de prêt. Enfin, la dynamique d'actualisation régulière du modèle garantit son adaptation continue aux évolutions du comportement client et du marché bancaire.

En somme, cette étude démontre la pertinence de l'approche basée sur la data science pour renforcer la performance commerciale de la banque, tout en proposant un cadre analytique pouvant être élargi et amélioré dans les travaux futurs.