Rapport de projet :

Scoring

Membres du groupe :

Cheddadi Radja
Abichou Nour Elhouda
Poladian Iskouhie
Si Youcef Tiziri
WU Zhuolin

2022/2023

Table de matière :

Introduction

L'octroi de prêts est un processus important pour les institutions financières, néanmoins celui-ci comporte également des risques financiers importants. Les prêteurs doivent donc prendre des décisions éclairées sur l'approbation ou le refus d'un prêt. Les données personnelles des clients, telles que le revenu, l'historique des crédits et la profession, sont des éléments clés à considérer lors de l'analyse de la solvabilité d'un client potentiel.

Dans ce contexte, nous allons donc nous concentrer sur la construction d'un modèle de prédiction de défauts de prêts, en utilisant l'ensemble de données HMEQ contenant des informations sur 5960 demandeurs de prêts.

Nous allons commencer par une analyse exploratoire des données pour mieux comprendre la structure des données et identifier d'éventuelles relations entre les variables. Nous effectuons ensuite un prétraitement des données en traitant les valeurs manquantes, en discrétisant les variables et en gérant les échantillons non équilibrés. Enfin, nous construisons deux modèles de classification : une régression logistique et une forêt aléatoire, et comparons leurs performances en utilisant différents critères d'évaluation. Nous allons conclure en discutant des résultats de l'analyse et des perspectives afin d'améliorer la qualité des prévisions dans ce domaine.

Problématique : **Quelles sont les caractéristiques clés d'un demandeur qui pourraient aider les institutions financières à prendre la décision d'octroyer un crédit avec un risque de défaillance le plus faible possible ?**

I. Analyse exploratoire des données

L'analyse exploratoire des données (AED) est une étape importante du processus d'analyse des données, car elle nous aide à mieux comprendre nos données et à identifier les modèles, les tendances ou les relations qui existent au sein de celles-ci. En comprenant les définitions des variables de notre ensemble de données, nous serons mieux armés pour effectuer une AED efficace et obtenir des informations qui pourront guider notre analyse et notre modélisation des données.

Le dataset HMEQ fournit des informations sur 5960 prêts sur valeur domiciliaire, y compris les caractéristiques des demandeurs et des informations sur leur historique de crédit **(voir annexe 01)**. La variable BAD indique si le demandeur n'a pas remboursé le prêt ou s'est mis en défaut de paiement grave, tandis que la variable LOAN indique le montant de la demande de prêt. La variable MORTDUE indique le montant dû sur l'hypothèque existante, et la variable VALUE indique la valeur de la propriété actuelle. La variable REASON fournit des informations sur le motif de la demande de prêt, DebtCon indiquant la consolidation de dettes et HomeImp l'amélioration de l'habitat. La variable JOB fournit des informations sur la profession du demandeur, tandis que la variable YOJ indique le nombre d'années pendant lesquelles le demandeur a occupé son emploi actuel. La variable DEROG indique le nombre de rapports dérogatoires majeurs, et la variable DELINQ indique le nombre de lignes de crédit en souffrance. La variable CLAGE indique l'âge de la ligne de crédit la plus ancienne en mois, et la variable NINQ indique le nombre de demandes de crédit récentes. Enfin, la variable CLNO indique le nombre de lignes de crédit, et la variable DEBTINC indique le ratio dette/revenu.

1. Analyse des statistique des variables numériques

Nous avons utilisé la fonction describe afin d'obtenir les statistiques de nos variables numériques **(voir tableau 1)**. La moyenne de la variable BAD est de 0,20, ce qui signifie qu'en moyenne, 20 % des demandeurs de l'ensemble de données n'ont pas remboursé leur prêt ou sont en retard de paiement **(voir annexe 02)**. Cela donne un aperçu important de la solvabilité des demandeurs dans l'ensemble de données. L'histogramme de la variable cible BAD démontre visuellement ce constat. Cela veut aussi dire que ce dataset n'est pas équilibré, et que les colonnes d'étiquettes sont réparties de manière inégale.

Tableau 1 : statistiques des variables numériques

Column Name	Count	Mean	Standard Deviation	Minimum	25th Percentile	50th Percentile	75th Percentile	Maximum
BAD	5960	0.20	0.40	0.00	0.00	0.00	0.00	1.00
LOAN	5960	18608	11207	1100	11100	16300	23300	89900
MORTDUE	5442	73761	44457	2063	46276	65019	91488	399550
VALUE	5848	101776	57385	8000	66075.5	89235.5	119824.25	855909
YOJ	5445	8.92	7.57	0.00	3.00	7.00	13.00	41.00
DEROG	5252	0.25	0.85	0.00	0.00	0.00	0.00	10.00
DELINQ	5380	0.45	1.13	0.00	0.00	0.00	0.00	15.00
CLAGE	5652	179.77	85.81	0.00	115.12	173.47	231.56	1168.23
NINQ	5450	1.19	1.73	0.00	0.00	1.00	2.00	17.00
CLNO	5738	21.30	10.14	0.00	15.00	20.00	26.00	71.00
DEBTINC	4693	33.78	8.60	0.52	29.14	34.82	39.00	203.31

Le montant moyen des **prêts (LOAN)** dans l'ensemble des données est de 18 608 $, avec un écart type de 11 207 $. Cela suggère que les montants des prêts dans l'ensemble de données sont relativement similaires, la plupart des demandeurs demandant des prêts dans une fourchette d'environ 7 401 $. Cependant, certains demandeurs ont demandé des prêts nettement supérieurs ou inférieurs à cette fourchette, comme l'indique l'écart type relativement important.

Le montant moyen du **prêt hypothécaire (MORTDUE)** est de 73 761 $, avec un écart type de 44 457 $. Cela suggère que les montants dus de l'ensemble des données sont relativement similaires, les personnes étudiées se situent dans la même tranche de revenus. La plupart des demandeurs ayant des montants dus compris dans une fourchette d'environ 22 304 $. Cependant, certains demandeurs ont des montants dus nettement supérieurs ou inférieurs à cette fourchette, ce qui rend l'écart-type relativement important.

La valeur moyenne de la **propriété actuelle (VALUE)** est de 101 776 $, avec un écart type de 57 386 $. Les valeurs minimale et maximale sont respectivement de 8 000 $ et 855 909 $. Cela indique que les valeurs immobilières des demandeurs se situent largement dans la moyenne, la plupart des propriétés ayant des valeurs comprises dans une fourchette d'environ 101 776 $. Cependant, certaines propriétés ont des valeurs nettement supérieures ou inférieures à cette fourchette, comme l'indique l'écart type relativement important.

Le nombre moyen **d'années dans l'emploi actuel (YOJ)** est de 8,9, et l'écart type est de 7,57. Les valeurs minimale et maximale sont respectivement de 0 et 41. Cela suggère qu'il y a un écart considérable dans les années d'expérience professionnelle, certains candidats ayant beaucoup plus d'expérience que d'autres. Le nombre minimal d'années d'expérience professionnelle est de 0, et le nombre maximal d'années d'expérience professionnelle est de 41.

La moyenne du **nombre de rapports dérogatoires (DEROG)** est de 0,25, ce qui indique qu'en moyenne chaque prêt a environ 0,25 rapport dérogatoire. La valeur minimale est de 0 et la valeur maximale de 10, ce qui montre un écart important entre les valeurs. Les 25e, 50e et 75e percentiles sont égaux à 0, ce qui signifie que 25 % des prêts de l'ensemble de données n'ont aucun rapport dérogatoire.

La valeur moyenne des nombre de **lignes de crédit en souffrance (DELINQ)** est de 0,45, ce qui indique qu'en moyenne chaque prêt a environ 0,45 ligne de crédit en souffrance. L'écart-type est de 1,13, ce qui signifie que les valeurs des lignes de crédit en souffrance sont en moyenne écartées de 1,13 de la valeur moyenne. La valeur minimale est de 0 et la valeur maximale de 15, ce qui montre que certains groupes de personnes ont un nombre élevé de lignes de crédit en souffrance.

La moyenne **d'âge de la ligne de crédit (CLAGE)** est de 179,77, ce qui signifie que l'âge moyen de la ligne de crédit la plus ancienne dans l'ensemble des données est d'environ 179,77 mois. L'écart type est de 85,81, ce qui signifie que les valeurs de l'âge de la ligne de crédit la plus ancienne s'écartent de la valeur moyenne de 85,81 mois en moyenne. La valeur minimale est de 0 et la valeur maximale de 1168,23, ce qui signifie que certains emprunteurs ont des lignes de crédit très anciennes, tandis que d'autres sont très récents.

La moyenne du **nombre de demandes de crédit récentes (NINQ)** est de 1,19, ce qui signifie qu'en moyenne chaque emprunteur a environ 1,19 demandes de crédit récentes. L'écart type est de 1,73, ce qui signifie que les valeurs du nombre de demandes de crédit récentes sont en moyenne écartées de 1,73 de la valeur moyenne. La valeur minimale est de 0 et la valeur maximale de 17, cela signifie que certains emprunteurs ont 17 enquêtes récentes, alors que la plupart des emprunteurs en ont 0 ou 1. La plupart ont un crédit raisonnable, mais il existe un petit nombre de personnes avec de nombreuses demandes de crédit risquent des créances irrécouvrables sur une longue période.

La valeur moyenne du **nombre de lignes de crédit par emprunteur (CLNO)** est de 21,30, ce qui indique qu'en moyenne, chaque prêt a environ 21,30 de lignes de crédit. L'écart-type est de 10,14. Enfin, La valeur minimale est 0 et la valeur maximale est 71, ce qui signifie que certains emprunteurs ont un grand nombre de lignes de crédit, alors que la plupart des emprunteurs en ont un nombre modéré .

Enfin la moyenne du **ratio dette/revenu (DEBTINC)** est de 33,779915, ce qui indique qu'en moyenne, le ratio de chaque emprunteur est d'environ 33,78. L'écart type est de 8,601746. La valeur minimale est de 0,524499, et la valeur maximale de 203,312149, ce qui signifie que certains emprunteurs ont un ratio dette/revenu très faible, tandis que d'autres ont un ratio dette/revenu très élevé .

2. Visualisation des données non numériques

Les deux seules variables non numériques du dataset sont les variables REASON et JOB; nous utilisons donc de la data visualisation afin de nous permettre de mieux comprendre la distribution des données au sein de leur catégorie et de voir la fréquence d'apparition de chaque catégorie.

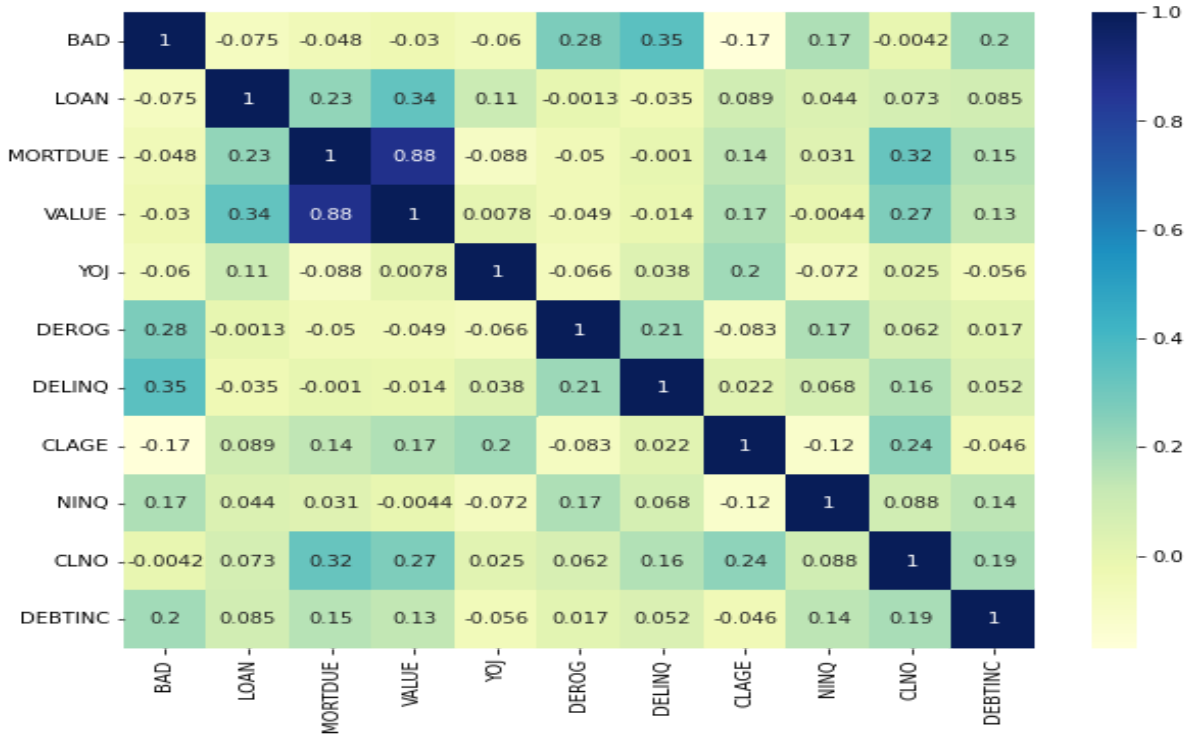
La colonne REASON possède deux catégories - "HomEmp" et "DebtCon". Sur le total des observations, 1780 observations sont de la catégorie "HomEmp" et 3928 observations sont de la catégorie "DebtCon". Ces informations peuvent aider à comprendre la distribution des raisons des demandes de prêt et peuvent également donner un aperçu de l'importance relative de chaque raison dans l'explication de la variable BAD **(voir annexe 03)**.

La colonne EMPLOI du dataset comporte 2388 observations pour la catégorie "Autre", 1276 pour "Professionnel et cadre", 948 pour "Bureau", 767 pour "Directeur", 109 pour "Ventes" et moins de 100 pour "Ventes". Ceci indique le type d'emploi occupé par les individus ayants souscrit un prêt. Cela suggère qu'un éventail diversifié de personnes, y compris des cols blancs et des cols bleus, demandent des prêts **(voir annexe 04)**.

3. Matrice de corrélation

La matrice de corrélation du dataset HMEQ nous montre la relation entre les différentes variables du jeu de données et nous fournit une mesure quantitative de la force et de la direction de la corrélation entre les variables.

Graphique 1 : Matrice de corrélations



Les corrélations positives les plus fortes sont entre BAD et DEROG (0,276081) et BAD et DELINQ (0,354107), ce qui signifie que le fait d'avoir un nombre plus élevé d'événements dérogatoires et de défaillances est positivement corrélé avec un prêt en défaut. La colonne BAD présente une corrélation positive avec les colonnes DELINQ, DEROG et CLNO. Cela indique qu'une augmentation des valeurs de ces colonnes est associée à une augmentation du nombre de créances douteuses. D'autre part, la colonne BAD présente une corrélation négative avec les colonnes VALUE, YOJ et LOAN, ce qui suggère qu'une augmentation de ces valeurs est associée à une diminution du nombre de créances douteuses **(voir Graphique 1).**

La variable LOAN a une corrélation positive avec la variable CLNO (0.07) qui est le nombre de lignes de crédit. D'autre part, la corrélation négative la plus forte est entre LOAN et YOJ (-0.105728), ce qui signifie que le fait d'avoir un nombre inférieur d'années d'emploi est négativement corrélé avec le montant du prêt. La variable MORTDUE présente une forte corrélation positive avec la variable VALUE (0,87), ce qui signifie que lorsque la valeur du bien immobilier augmente, l'hypothèque due augmente également, étant donné que la valeur de l'hypothèque est calculée sur la base de la valeur du bien immobilier. Enfin, la variable YOJ (Years of Job) présente une corrélation négative avec DEBTINC (Debt to Income ratio) (-0.05), ce qui signifie que plus le nombre d'années d'emploi augmente, plus le ratio dette/revenu diminue **(voir Graphique 1)**.

4. Analyse ANOVA (Analysis of Variance)

L'ANOVA est un outil statistique que nous utilisons pour déterminer si les moyennes de deux groupes de données sont significativement différentes. Si les moyennes sont égales, cela signifie que les groupes ont la même distribution sous-jacente et que toute différence observée peut être due au hasard. Si les moyennes ne sont pas égales, cela suggère que les groupes sont différents et que la différence entre les moyennes n'est pas due au hasard.

Notre code calcule les résultats de l'analyse ANOVA pour chaque variable numérique du dataset HMEQ sélectionnée sous forme d'un t-stat et d'un P-value, ainsi qu'une conclusion sur l'hypothèse nulle, qui peut être rejetée ou non en fonction de la valeur de P. Pour chaque variable, nous réalisons des tests t pour déterminer si une différence significative existe entre les moyennes de deux groupes de données. La valeur p indique la probabilité d'observer la statistique t en supposant que l'hypothèse nulle est vraie. Si la valeur p est inférieure à un certain niveau de signification (que nous allons fixer à 0,05), l'hypothèse nulle peut être rejetée et il y a des preuves d'une différence significative entre les moyennes. Si la valeur p est supérieure au niveau de signification, il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle et les moyennes des groupes sont considérées comme égales.

Les résultats montrent que pour les variables LOAN, MORTDUE, VALUE, YOJ, DEROG, DELINQ, CLAGE, NINQ, et DEBTINC, l'hypothèse nulle est rejetée car les valeurs p sont inférieures au niveau de signification. Pour la variable CLNO, l'hypothèse nulle ne peut être rejetée car la valeur p est supérieure au niveau de signification, ce qui suggère que les moyennes des deux groupes sont égales et qu'il n'y a pas suffisamment de preuves pour montrer une différence significative entre les moyennes des deux ensembles de données pour la CLNO **(voir annexe 05)**. Cependant, cela ne signifie pas nécessairement que la variable CLNO doit être éliminée de notre modèle. Elle peut encore être pertinente pour d'autres aspects de notre analyse.

5. Analyse chi deux

La statistique du test du chi deux nous aide à mesurer la différence entre les fréquences observées et attendues dans le tableau de contingence. Plus la valeur p est faible, plus la preuve contre l'hypothèse nulle qu'il n'y a pas d'association entre les deux variables est forte. Le tableau résume les résultats qui indiquent l'association entre les variables catégorielles JOB et REASON avec la variable cible BAD

La première ligne du tableau montre les résultats d'un test du chi carré comparant les variables JOB et BAD. La statistique du test est de 81,93, et la valeur p est de 0,00000. Cela suggère qu'il existe une association significative entre JOB et BAD. Avec une valeur p très faible (inférieure à 0,10), nous rejetons l'hypothèse nulle et concluons qu'il existe une relation entre les deux variables **(voir annexe 06)**.

La deuxième ligne du tableau montre les résultats d'un test du chi carré comparant les variables REASON et BAD. La statistique du test est de 8,24, et la valeur p est de 0,08305. Cela suggère qu'il existe une association plus faible entre REASON et BAD qu'entre JOB et BAD. Avec une valeur p supérieure à 0,10, nous rejetons l'hypothèse nulle et concluons qu'il n'y a suffisamment de preuves pour affirmer qu'il existe une relation entre les deux variables **(voir annexe 06)**.

II. Data processing

Avant de mettre en œuvre un modèle d'apprentissage automatique, il est indispensable de traiter les données pour garantir sa qualité et sa plus value décisionnelle. Le schéma en annexe représente les différentes étapes de notre démarche de traitement des données **(voir annexe 07)**.

1. Traitement des valeurs manquantes

Dans notre travail avec des données, il est fréquent de rencontrer des valeurs manquantes. Cependant, ces valeurs peuvent affecter les résultats de l'analyse, il est donc important de les traiter correctement. Dans notre cas, la quantité de valeurs manquantes varie considérablement entre les colonnes **(voir annexe 08)**.

Si nous choisissons de supprimer simplement les lignes avec des valeurs manquantes, nous perdrons environ 43% de nos données. Il est important de noter que les variables clés BAD et LOAN ne contiennent pas de valeurs manquantes, contrairement à la variable DEBTINC qui a le plus grand nombre de valeurs manquantes (1267, soit plus de 25% du total des observations) **(voir annexe 08)**. Cela signifie qu'une partie importante de l'information sur le ratio dette/revenu est manquante et que toute analyse impliquant cette variable peut ne pas être précise. De même, les variables MORTDUE, YOJ, DERO, DELINQ, NINQ et CLNO pourraient contenir des informations importantes sur le prêt et leurs valeurs manquantes doivent également être traitées.

Les deux méthodes les plus courantes pour traiter les valeurs manquantes consistent à soit les supprimer, soit à les imputer avec des valeurs de remplacement en utilisant des mesures statistiques telles que la moyenne, la médiane, le minimum et le maximum, ou des méthodes plus avancées telles que la régression linéaire ou les modèles d'apprentissage automatique. Dans notre cas, compte tenu du nombre important de valeurs manquantes, nous avons choisi d'implémenter l'imputation pour conserver un nombre suffisant d'observations dans notre ensemble de données.

Nous avons décidé de ne pas utiliser l'imputation par la moyenne en raison de l'existence d'un écart entre la moyenne et la médiane, ce qui peut sensiblement modifier la distribution initiale des variables et le jeu de données. À la place, nous avons choisi la méthode KNN Imputer qui conserve la valeur et la variabilité des données. Nous avons donc choisi d'utiliser la méthode KNN pour imputer les valeurs manquantes. Ce choix est fondé sur la façon dont les données sont distribuées et prend en compte la présence de valeurs aberrantes grâce à la normalisation des données **(voir annexe 09)**.

Lorsque nous imputons simultanément plusieurs variables, comme dans notre cas où nous pouvons rencontrer des séries de valeurs manquantes consécutives, l'un des principaux avantages du KNN est la préservation de la cohérence interne de chaque enregistrement. Cela

est possible en prenant en compte les autres variables ainsi que les caractéristiques communes des observations dans notre base de données.

D'une manière générale, la méthode d'imputation KNN se déroule en deux étapes. Supposons que nous avons n observations et p variables. Nous trouvons également un hyper paramètre k, qui est déterminé par l'utilisateur et qui permet de fixer le nombre de 'k' échantillons similaires ou proches dans l'espace de données grâce à un calcul de distance.Autrement dit la valeur k fixé conduit la performance de l'imputation c'est à dire un k grand va augmenter la distance moyenne entre les voisins, ce qui entraîne une perte de précision de la valeur imputée et une hausse du délai d'imputation.

Toutefois, le fait de choisir un faible k atténue les performances de KNN en ce sens que le processus d'imputation insiste un peu trop sur certaines observations prédominantes (les observations complètes qui servent à la base de notre modification) pour la modélisation de ces valeurs manquantes. Pour chaque échantillon avec k observations, les valeurs manquantes sont imputées en utilisant la moyenne des k valeurs les plus proches dans l'ensemble de données.

Pour cela, on a essayé manuellement plusieurs valeurs de k. En effet, dans le but de dégager le k optimal, nous testerons le modèle pour chaque k de 2 à 15, nous évaluerons l'erreur de test et nous visualiserons la performance en fonction de k. On a constaté qu'après k = 5, le surapprentissage apparaît dans les métriques d'évaluation de chaque modèle testé. Par exemple, pour k = 10 on a trouvé une classification de 100% ce qui déclenche le "overfitting" La première étape consiste a encoder les valeurs catégorielles en valeurs numériques étant donné que le KNN Imputer ne travaille que sur des variables numériques et à normaliser notre jeu de données afin de neutraliser l'effet des différentes échelle et de minimiser l'impact négatif des valeurs aberrantes.

Le KNN procède ensuite au calcul des distances entre les n-1 couple d'observations ou il existe notre observation qui contient des valeurs manquantes. La distance est calculée entre cette observation et des variables complètes dans le but de déterminer les K les plus proches à ce dernier. Dans notre cas, on calcule la distance euclidienne en ne tenant pas compte des valeurs manquantes et en augmentant le poids des coordonnées complètes.

Enfin, la valeur de variable à déterminer de l'individu **i** est calculée à partir des K plus proches voisins qui contient la donnée propre à la variable recherchée. En ce qui concerne les données quantitatives, il s'agit de calculer la moyenne des données des Ks voisins les plus proches. Quant aux données qualitatives, il s'agit d'un vote à la majorité sur les données K.

2. Discrétisation des données

Nous avons utilisé le processus de discrétisation afin de transformer des données continues en données discrètes, la discrétisation est fortement recommandée pour des modèles comme

celui de la régression logistique, également, cela nous permettra de faire face à notre grande quantité d'Outliers, sans avoir recours à la suppression.
Nous avons un large choix de méthodes quant à la discrétisation, nous avons retenu les deux modèles les plus adaptés à nos données, à savoir le K-means puis l'arbre de décision.
A- Le K-means:
On voit **(annexe 14)**, que le k-means discrétise bien nos données, en trois classe, nous avons choisi trois classe afin de nous faciliter la lecture, seulement on peut voir que pour la variable value et clage, on constate une observation isolée dans une classe, ce qui n'est pas nécessairement bon.
B- L'arbre de décision :
Maintenant, avec cette méthode **(annexe 13)**, on peut voir que la discrétisation sur trois classes donne de meilleurs résultats et cela sur l'ensemble de nos variables, nous choisirons donc ce modèle.
Nous allons donc ensuite renommer les trois intervalles en "Mid - Low -High", avec pour bornes les valeurs minimales et maximales de chaque groupe **(annexe 13)**.

3. Traitement des échantillons non équilibrés

Dans nos données originales, 19,9% étaient des clients en défaut et 80,1% des clients normaux. Cela indique une distribution déséquilibrée des colonnes d'étiquettes dans notre jeu de données. Ainsi, le modèle appris par le classificateur peut présenter de grandes déviations, entraînant une dégradation des performances de classification.

Par conséquent, lors du traitement d'un ensemble de données déséquilibré, il est nécessaire de procéder à un resampling pour équilibrer la distribution des catégories afin d'améliorer les performances et la précision du classificateur.

Le resampling peut être effectué de deux manières principales : l'undersampling et l'oversampling. L'undersampling consiste à réduire le nombre d'observations de la classe majoritaire, tandis que l'oversampling consiste à augmenter le nombre d'observations de la classe minoritaire.

Dans notre cas, la taille de l'échantillon des clients en défaut est beaucoup plus petite que celle de l'échantillon des clients normaux. Nous choisissons alors l'oversampling pour augmenter la taille de l'échantillon de clients en défaut. Cette méthode permet de mieux utiliser les données existantes et peut être plus facile à mettre en œuvre que l'undersampling, qui peut entraîner une perte d'informations et une réduction de la taille de l'ensemble de données.

Plus précisément, nous utilisons l'algorithme ADASYN (Adaptive Synthetic Sampling) pour implémenter l'oversampling, il vise à équilibrer les données en créant de nouveaux exemples synthétiques pour la classe minoritaire.

III. Modélisation

Dans cette partie de notre étude, nous allons explorer la modélisation de notre problème de classification binaire en utilisant deux algorithmes différents : la régression logistique, et les forêts aléatoires. Pour évaluer la performance de chaque modèle, nous utiliserons une variété de critères d'évaluation, tels que la matrice de confusion, l'accuracy, la précision, le recall, le F1 score et l'AUC. Après la construction des trois modèles, nous allons les comparer pour déterminer lequel d'entre eux donne les meilleures performances pour notre ensemble de données HMEQ.

1. Critères d'évaluation des modèles

Matrice de confusion : la matrice de confusion est une table qui montre les résultats de la classification en fonction des vraies classes et des prédictions du modèle. Elle est composée de quatre éléments : les vrais positifs (TP), les faux positifs (FP), les vrais négatifs (TN) et les faux négatifs (FN).

Accuracy : une mesure de la précision globale d'un modèle de classification. Elle représente le nombre de prédictions correctes (TP + TN) sur l'ensemble des prédictions (TP + FP + TN + FN).

Precision : une mesure de la précision des prédictions positives d'un modèle. Elle représente le nombre de vrais positifs (TP) parmi les prédictions positives (TP + FP).

Recall : une mesure de la capacité d'un modèle à trouver toutes les occurrences positives. Il représente le nombre de vrais positifs (TP) parmi tous les exemples positifs (TP + FN).

F1 score : une mesure de la précision d'un modèle de classification binaire qui prend en compte à la fois la précision (precision) et le rappel (recall).
F1 score = 2 / (1/precision + 1/recall)

ROC (Receiver Operating Characteristic) : la courbe ROC est un graphique représentant le taux de vrais positifs en fonction du taux de faux positifs pour un modèle de classification binaire.

AUC (Area Under the Curve) : L'AUC représente la surface sous la courbe ROC, qui représente la relation entre le taux de vrais positifs et le taux de faux positifs.

Notre objectif est de prédire les demandeurs à risque de défaillance, c'est-à-dire d'identifier correctement autant de demandeurs avec BAD = 1 que possible. Nous voulons donc augmenter le recall de notre modèle autant que possible tout en assurant son accuracy.

2. Constructions des modèles

a. Logistic Regression

Dans le cadre de la régression logistique, notre modèle est présenté sous la forme suivant :

$$\text{logit } p_\beta(x) = \beta 0 + \beta 1 x1 + \ldots + \beta p x p$$

Nous trallavions dans le cadre dans une loi binomial conditionnelle : Y| X = x, c'est le type de loi exponentielle qu'on choisit en se basant sur la nature binaire de notre variable Y. Nous admettons un transformeur g de l'espérance conditionnelle :

g(E[Y| X=x])= g(p(x))= **logit** p(x) = x`$\boldsymbol{\beta}$
avec g(u)=log(u/(1 − u)) : la fonction de lien de notre modèle.

Cette régression logistique est un GLM(Le modèle linéaire généralisé). La fonction glm va nous renvoyer les estimations du maximum de vraisemblance de $\beta 0$ à βn. Grâce à ces derniers, on peut ainsi avoir une estimation de la probabilité qu' un client ne rembourse pas son prêt ou en défaut de paiement grave (BAD= 1).

- **Modele complet :**

<u>Evaluation de modele par R2</u> :
Le modèle après sélection de variables prédit donc 22.08 % de la variance de la probabilité de qu'un demandeur soit en défaut de paiement ou bien en difficulté soit une traduction presque égale à la modèle complète avec une légère différence.

On peut faire l'évaluation aussi par rapport la **différence entre la déviance nulle et résiduelle** En calculant la statistique X2 du modèle :
X2 = Déviance nulle – Déviance résiduelle = 1232.4
Or il y a p = 26 de degres de liberte de variables preductives
On peut déterminer alors p value pour X2 = 12312.4 et on trouve qu'elle est égale a 0,0000.
⇒ Vu que la valeur de p est bien inférieure à 0,05, nous conclurons que le modèle est utile pour prédire la probabilité qu'un individu donné fasse défaut.

```
Call:
glm(formula = BAD ~ ., family = "binomial", data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3804  -0.5923  -0.4390  -0.2581   2.7778

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.11487    0.33475   3.330 0.000867 ***
LOANhigh      0.35054    0.23586   1.486 0.137226
LOANlow       0.34415    0.11340   3.035 0.002407 **
MORTDUEhigh  -0.13761    0.25799  -0.533 0.593758
MORTDUElow    0.09367    0.14887   0.629 0.529203
VALUEhigh     1.58797    0.56178   2.827 0.004704 **
VALUElow     -0.08184    0.15868  -0.516 0.606041
YOJhigh      -0.23651    0.09080  -2.605 0.009196 **
YOJlow       -0.23543    0.10081  -2.336 0.019517 *
DEROGhigh     0.44304    0.38134   1.162 0.245311
DEROGlow     -1.43574    0.15469  -9.281  < 2e-16 ***
DELINQhigh    3.83896    0.61339   6.259 3.89e-10 ***
DELINQlow    -1.66284    0.11709 -14.202  < 2e-16 ***
CLAGEhigh    -0.13534    0.16706  -0.810 0.417865
CLAGElow      0.99513    0.10559   9.425  < 2e-16 ***
NINQhigh      0.32650    0.20499   1.593 0.111221
NINQlow      -0.70393    0.11261  -6.251 4.08e-10 ***
CLNOhigh      0.54222    0.16696   3.248 0.001164 **
CLNOlow       0.39860    0.09345   4.265 2.00e-05 ***
DEBTINChigh  17.36572  268.76856   0.065 0.948483
DEBTINClow   -0.85273    0.09183  -9.286  < 2e-16 ***
REASON2       0.25086    0.08512   2.947 0.003208 **
JOB2          0.34731    0.34632   1.003 0.315924
JOB3         -0.32536    0.23853  -1.364 0.172553
JOB4          0.02885    0.22954   0.126 0.899977
JOB5         -0.91270    0.24760  -3.686 0.000228 ***
JOB6         -0.02468    0.24363  -0.101 0.919314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5597.6  on 5535  degrees of freedom
Residual deviance: 4351.5  on 5509  degrees of freedom
AIC: 4405.5

Number of Fisher Scoring iterations: 15
```

Nous avons donc un modèle initial, comprenant l’ensemble de nos variables binarisées (les modalités ayant été transformées en « variables binaires »).

*Nous avons décidé de changer la variable de référence, en définissant la variable MID comme variable de référence, et cela afin de nous faciliter l’interprétation.

- **Modele apres stepwise :**

Sur ce modèle complet nous avons fait un Stepwise sur R, Stepwise qui nous a permis de garder les variables les plus importantes (12), ce qui nous a également permis d’éviter le problème de multicolinéarité.

Le modèle que nous avons retenu après le Stepwise est le suivant :

BAD = DEROG_low + DELINQ_high + DELINQ_low + CLAGE_low + NINQ_low + CLNO_high + CLNO_low + DEBTINC_low + REASON_1 + JOB_3 + JOB_5 + YOJ_mid + LOAN_mid

```
Call:
glm(formula = BAD ~ DEROG_low + DELINQ_high + DELINQ_low + CLAGE_low +
    NINQ_low + CLNO_high + CLNO_low + DEBTINC_high + DEBTINC_low +
    REASON_1 + JOB_3 + JOB_5 + YOJ_mid + LOAN_mid, family = "binomial",
    data = dt2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4507  -0.6045  -0.4401  -0.2583   2.7962

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.64325    0.21786   7.543 4.60e-14 ***
DEROG_low    -1.53474    0.14154 -10.844  < 2e-16 ***
DELINQ_high   3.82700    0.61249   6.248 4.15e-10 ***
DELINQ_low   -1.64567    0.11577 -14.215  < 2e-16 ***
CLAGE_low     1.03107    0.08999  11.458  < 2e-16 ***
NINQ_low     -0.80464    0.09892  -8.134 4.14e-16 ***
CLNO_high     0.48848    0.16286   2.999 0.002706 **
CLNO_low      0.41793    0.09287   4.500 6.79e-06 ***
DEBTINC_high 17.39552  268.27963   0.065 0.948301
DEBTINC_low  -0.85789    0.09095  -9.432  < 2e-16 ***
REASON_1     -0.26099    0.08358  -3.123 0.001792 **
JOB_3        -0.34073    0.10167  -3.351 0.000804 ***
JOB_5        -0.95975    0.12138  -7.907 2.63e-15 ***
YOJ_mid       0.24543    0.07813   3.141 0.001682 **
LOAN_mid     -0.32659    0.10602  -3.080 0.002067 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5597.6  on 5535  degrees of freedom
Residual deviance: 4365.2  on 5521  degrees of freedom
AIC: 4395.2

Number of Fisher Scoring iterations: 15
```

1 er constat :

Dans un premier temps, nous voulons savoir si nos coefficients sont significativement différents de zéro, nous posons donc nos hypothèses :

H0 : Coeff = 0

Ha : Coeff ≠ 0

Nous allons ensuite utiliser le test de Wald, afin de tester la significativité de nos coefficients, nous fixons un seuil de 5% pour rejeter H0, on s'intéresse donc à la p-value, on peut constater que l'ensemble de nos coefficients sont différents de zéro au seuil de 5%, mis à part le coefficient de la variable DEBTIN_HIGH qu'on aurait retenu à un seuil de 10%, mais que nous rejetons dans notre cas. (Il est important de noter, que R facilite la tâche en mettant des étoiles devant nos coefficient afin de définir leur significativité).

2 -ème constat :

Maintenant on s'intéresse à la mesure de l'impact de chaque variable indépendante sur la variable dépendante, on appelle ça l'ODS RATIO.

Pour la calculer il suffit de calculer l'exponentiel de chaque coefficient du modèle :

	variable	odds_ratio
(Intercept)	(Intercept)	5.171971e+00
DEROG_low	DEROG_low	2.155107e-01
DELINQ_high	DELINQ_high	4.592440e+01
DELINQ_low	DELINQ_low	1.928830e-01
CLAGE_low	CLAGE_low	2.804075e+00
NINQ_low	NINQ_low	4.472497e-01
CLNO_high	CLNO_high	1.629829e+00
CLNO_low	CLNO_low	1.518817e+00
DEBTINC_high	DEBTINC_high	3.587375e+07
DEBTINC_low	DEBTINC_low	4.240568e-01
REASON_1	REASON_1	7.702871e-01
JOB_3	JOB_3	7.112495e-01
JOB_5	JOB_5	3.829885e-01
YOJ_mid	YOJ_mid	1.278167e+00
LOAN_mid	LOAN_mid	7.213775e-01

- ☐ **Pour la variable DEROG_low par exemple, qui représente les personnes ayant moins de 2 rapports dérogatoires, on a un odds ratio de 0,21, on peut interpréter cela comme ayant une probabilité plus élevée de 0,21 de rembourser leur prêt.**

<u>3 -ème constat</u> :

On constate, en comparant le premier modèle qui emboite le deuxième modèle, que l'AIC est différent.

L'AIC est une mesure de la qualité d'un modèle statistique, l'équivalent du coefficient de détermination R^2, qui démontre la robustesse du modèle, en effet, nous voulons toujours choisir l'AIC le plus petit possible en comparant deux modèles ou plus.

On remarque que l'AIC du modèle complet est de 4405, le second, qui représente un sous modèle du premier, est légèrement plus petit avec 4395.

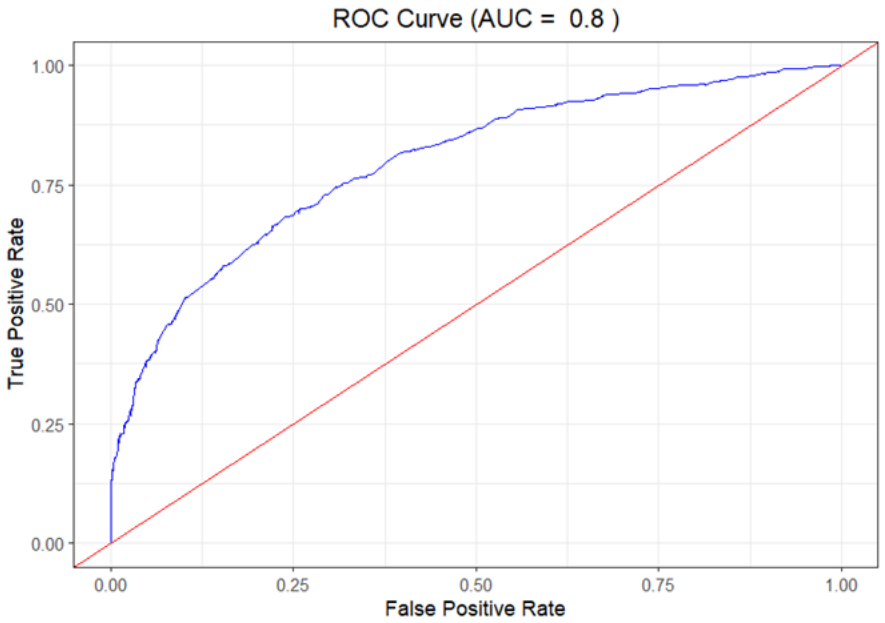
AIC M1 > AIC M2, nous choisirons donc le deuxième modèle.

<u>Evaluation de modele par R2</u> :
Le modèle après sélection de variables prédit donc 22.08 % de la variance de la probabilité de qu'un demandeur soit en défaut de paiement ou bien en difficulté soit une traduction presque égale à la modèle complète avec une légère différence.

On peut faire l'évaluation aussi par rapport la **différence entre la déviance nulle et résiduelle** En calculant la statistique X2 du modèle :
X2 = Déviance nulle – Déviance résiduelle = 1232,6
Or il y a p = 14 de degres de liberte de variables preductives
On peut déterminer alors p value pour X2 = 1232,6 et on trouve qu'elle est égale a 0,0000.
⇒ Vu que la valeur de p est bien inférieure à 0,05, nous conclurons que même le modèle après l'application de stewipe(la sélection de variables) est utile pour prédire la probabilité qu'un individu donné fasse défaut.

On constate aussi une légère augmentation de la probabilité log (-2LL) par rapport au modèle complet de -2175.733 a -2180.831 alors que notre but nous chercherons à réduire en ajoutant des variables prédictrices.

<u>Courbe ROC</u> :



L'avantage de la courbe ROC c'est qu'elle permet de mesurer la performance du modèle sans imposer un de seuil, et la première chose qu'on peut voir sur cette courbe, c'est qu'elle est au-dessus de la diagonale rouge (représentant une probabilité de 0,5 autrement dit, le pire des cas)

On doit calculer l'aire sous la courbe afin d'obtenir un AUC, celui-ci devrait idéalement être égal à zéro, on constate dans notre cas qu'il est égal à 0,8, ce qui démontre que nous avons là un bon modèle de prédiction**.**

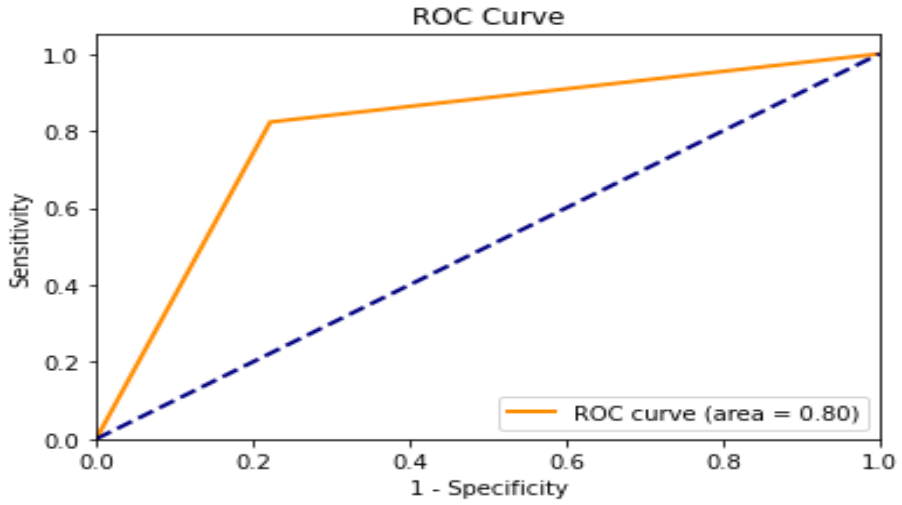
b. Random Forest

Random Forest est un algorithme d'apprentissage automatique qui combine plusieurs arbres de décision indépendants pour prédire les données en utilisant un vote majoritaire pour la classification et une moyenne pour la régression.

Nous ajustons l'ensemble d'entraînement en utilisant le RandomForestClassifier de la librairie sklearn. Avec les paramètres par défaut, nous constatons que l'accuracy atteint 85,6% et le recall atteint 71,01%. Après avoir ajusté les paramètres en utilisant un GridSearchCV, nous avons déterminé la combinaison optimale de paramètres comme suit : { 'n_estimators': 200, 'max_depth': 10, 'min_samples_split': 5, 'min_samples_leaf': 1, 'class_weight': 'balanced'}.

Nous pouvons observer que, malgré une certaine baisse de l'accuracy, le recall a augmenté de 71,01% à 82,96%. Nous constatons aussi que la précision a chuté de 89,82% à 70,35%. Ce qui est normal étant donné que la précision et le recall sont des mesures contradictoires, augmenter l'un d'entre eux peut entraîner une baisse de l'autre **(voir annexe 15)**. Par ailleurs, La valeur de L'AUC de notre modèle est de 0,80, ce qui indique un meilleur classificateur.

En ce qui concerne la courbe ROC qui est est une représentation graphique de la performance d'un classificateur binaire à différents seuils de discrimination. La courbe ROC représente le taux de vrais positifs (sensibilité) par rapport au taux de faux positifs (1-spécificité) à différents seuils de classification. Notre courbe est proche du coin supérieur gauche du graphique, ce qui signifie que notre classificateur est performant. Ceci signifie qu'on a des taux de vrais positifs élevés et des taux de faux positifs faibles.



La matrice de confusion nous indique que la classe BAD a été prédite correctement 924 fois. La classe BAD a été prédite comme n'étant pas BAD 264 fois. La classe non BAD a été prédite comme étant la classe BAD 134 fois et la classe non BAD a été prédite correctement 626 fois **(voir annexe 17)**.

Enfin, après avoir calculé les coefficients de nos paramètres et confirmé la performance de notre modèle nous avons mis en place un grille de score basée sur la normalisation des paramètres (les coefficients) sur la base d'un score entre 0 et 100 tout en s'assurant que le meilleur profile client aura un score de 100 **(voir annexe 18)**.

Cette grille de score présente les sous-groupes (intervalles) de chaque variable de notre modèle de scoring, avec leur plage de valeurs et le score attribué à chacun d'entre eux. Le score est une mesure de la contribution de chaque sous-groupe à la prédiction du risque de défaut de paiement. Pour chaque variable, les sous-groupes sont classés par ordre décroissant de score, ce qui signifie que les sous-groupes ayant le score le plus élevé ont la plus grande influence sur la prédiction de risque de défaut de paiement.

En utilisant cette grille de score, les analystes en banques peuvent évaluer la contribution de chaque variable et de chaque sous-groupe à la prédiction du risque de défaut de paiement. Cela peut également aider à comprendre quelles variables et sous-groupes sont les plus importants pour prendre des décisions éclairées en matière de prêt ou d'investissement.

Grille de score Random Forest

Feature	Ranges	Score
DELINQ	0-1.6	27.68%
	2-4.2	12.34%
	5-15	11.46%
DEROG	0-1.4	16.90%
	1.6-4	11.21%
	5-10	2.70%
DEBTINC	0.52-32.62	9.67%
	32.64-56.39	7.47%
	56.96-203.31	11.59%
CLAGE	0-193.50	9.31%
	193.63-281.57	5.44%
	281.92-1168.23	2.72%
NINQ	0-2.4	8.22%
	2.6-5	3.72%
	6-17	3.20%
JOBS	Mgr	3.01%
	Office	3.22%
	Other	7.20%
	ProfExe	2.87%
	Sales	1.80%

	Self	1.55%
REASON	DebtCon	3.58%
	HomeImp	3.56%
CLNO	0-23.4	3.41%
	23.6-36	3.39%
	37-71	1.99%
YOJ	0-5	4.04%
	5.6-12.4	3.14%
	12.6-41	2.85%
MORTDUE	2063-93805	2.03%
	93868-166000	2.01%
	166244-399550	2.13%
VALUE	8000-136877	1.99%
	137000-364000	1.95%
	415000-855909	0.84%
	855910-8559090	0.02%
LOAN	1100-23900	3.95%
	24000-42700	3.51%
	42900-89800	1.57%
	89844-898440	0.04%
	898455-1785600	0,02%

Conclusion

En conclusion, la mise en place d'un modèle de scoring est un processus complexe qui nécessite une analyse minutieuse des données, un traitement adéquat des valeurs manquantes et une modélisation précise.

L'analyse exploratoire des données a permis de comprendre les relations entre les variables et d'identifier les variables importantes pour la modélisation. Le traitement des données a permis de rendre les données utilisables pour la modélisation en traitant les valeurs manquantes, en discrétisant les données et en équilibrant les échantillons.

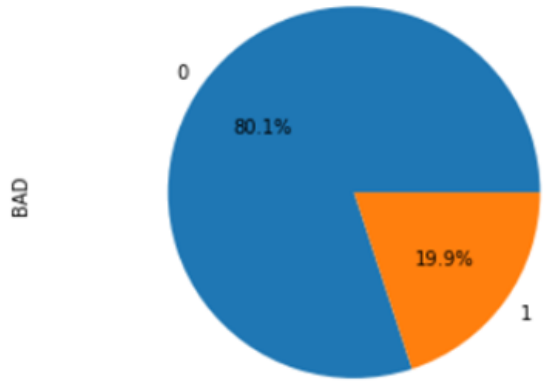
Enfin, la modélisation a été réalisée en utilisant des critères d'évaluation appropriés pour choisir les meilleurs modèles et en utilisant la régression logistique et la forêt aléatoire pour prédire le score. Ce modèle pourra être utilisé pour prendre des décisions éclairées dans le domaine de l'octroi de prêts.

Annexes :

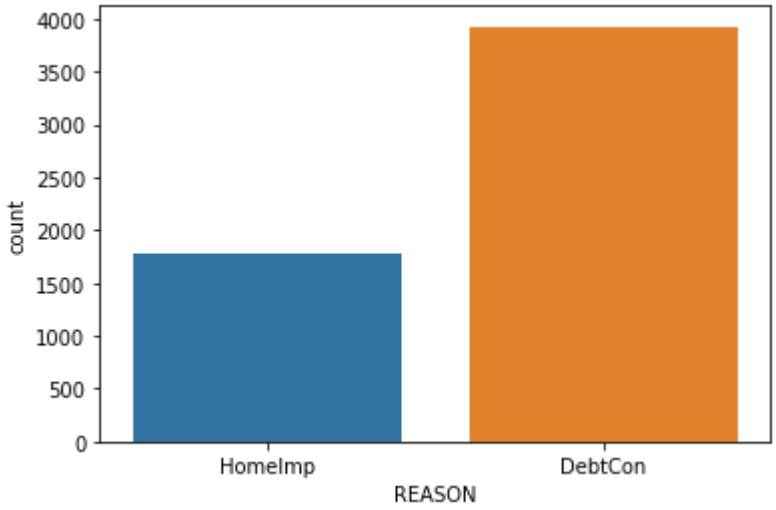
Annexe 01: description des variables

Variable	Description
BAD	Indique si le demandeur n'a pas remboursé le prêt ou s'est mis en défaut de paiement grave (0 = remboursement réussi, 1 = défaut de paiement)
LOAN	Montant de la demande de prêt (quantitative)
MORTDUE	Montant dû sur l'hypothèque existante (quantitative)
VALUE	Valeur actuelle de la propriété (quantitative)
REASON	Motif de la demande de prêt (qualitative)
JOB	Profession du demandeur (qualitative)
YOJ	Nombre d'années pendant lesquelles le demandeur a occupé son emploi actuel (quantitative)
DEROG	Nombre de rapports dérogatoires majeurs (quantitative)
DELINQ	Nombre de lignes de crédit en souffrance (quantitative)
CLAGE	Âge de la ligne de crédit la plus ancienne en mois (quantitative)
NINQ	Nombre de demandes de crédit récentes (quantitative)
CLNO	Nombre de lignes de crédit (quantitative)
DEBTINC	Ratio dette/revenu (quantitative)

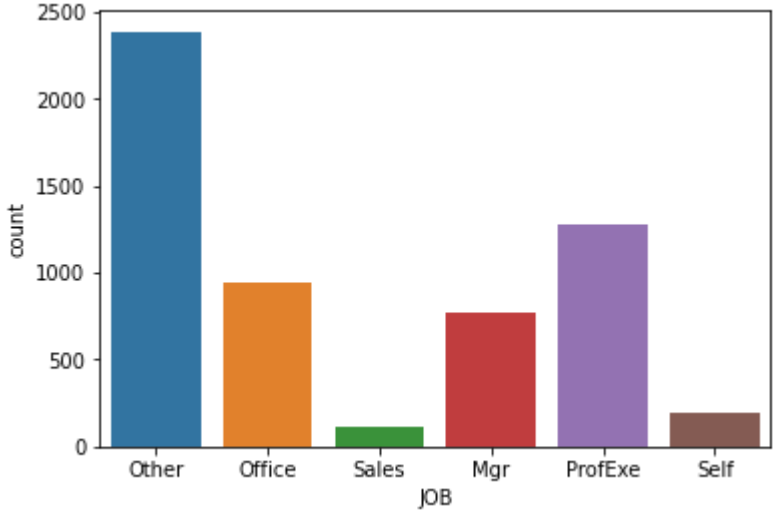
Annexe 02: Répartition des modalités de la variable cible BAD



Annexe 03: variable REASON par catégorie



Annexe 04: variable JOB par catégorie



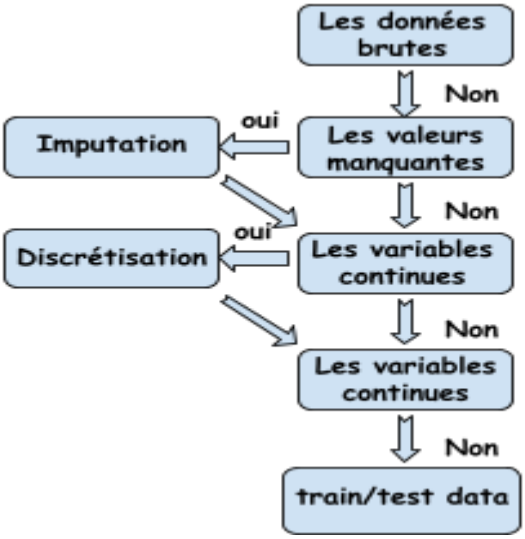
Annexe 05: analyse ANOVA

Variable	t-statistic	P-value	Result
LOAN	5.720	1.2455e-08	Reject the null hypothesis, variables have different means
MORTDUE	3.1999	0.00139974	Reject the null hypothesis, variables have different means
VALUE	2.2740	0.023107	Reject the null hypothesis, variables have different means
YOJ	5.753	1.0126e-08	Reject the null hypothesis, variables have different means
DEROG	-13.3536	3.7398e-38	Reject the null hypothesis, variables have different means
DELINQ	-18.1391	1.6169e-65	Reject the null hypothesis, variables have different means
CLAGE	12.7180	1.4956e-35	Reject the null hypothesis, variables have different means
NINQ	-11.7809	1.1294e-30	Reject the null hypothesis, variables have different means
CLNO	0.0265	0.978875	Fail to reject the null hypothesis, variables have equal means
DEBTINC	-14.1517	1.5264e-42	Reject the null hypothesis, variables have different means

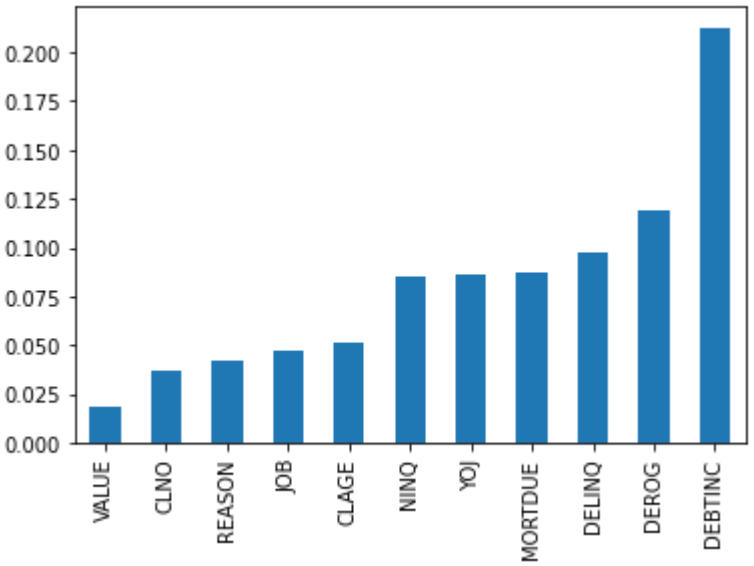
Annexe 06: analyse Khi 2

Variable	Test Statistic	p-value
JOB and BAD	81.9324895369	0.00000
REASON and BAD	8.24361	0.08305

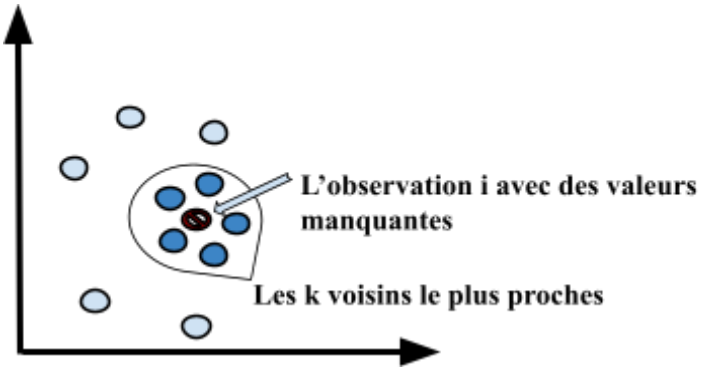
Annexe 07: Traitement des données



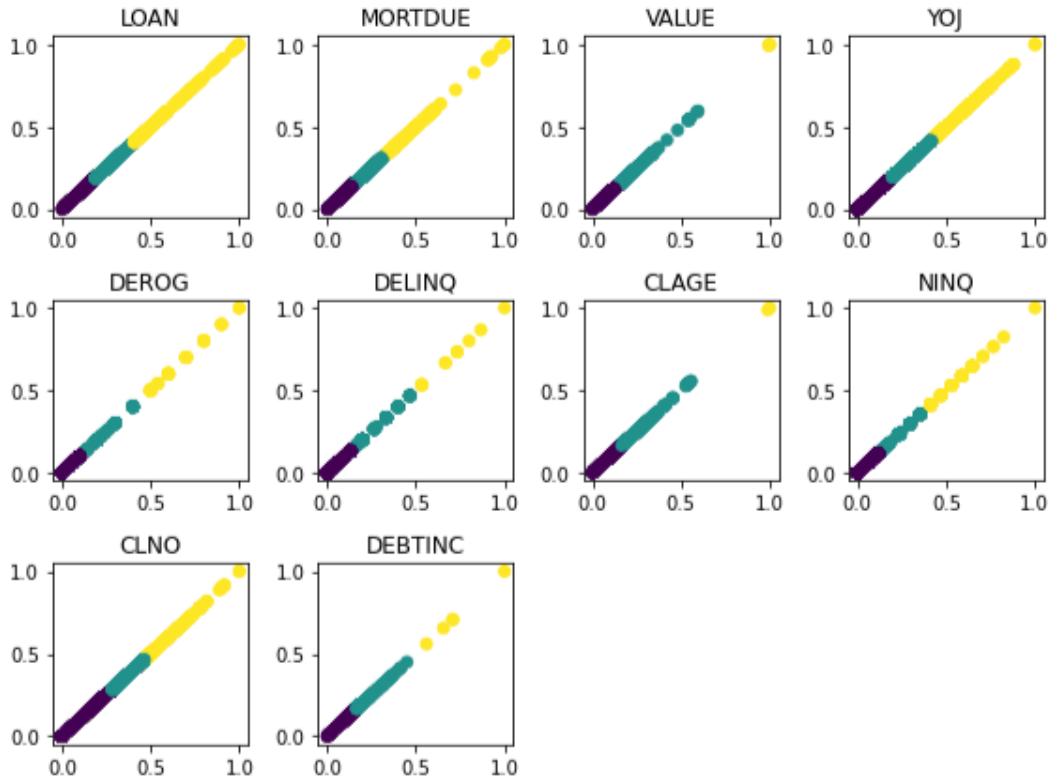
Annexe 8: visualiser le nombre des valeurs nuls



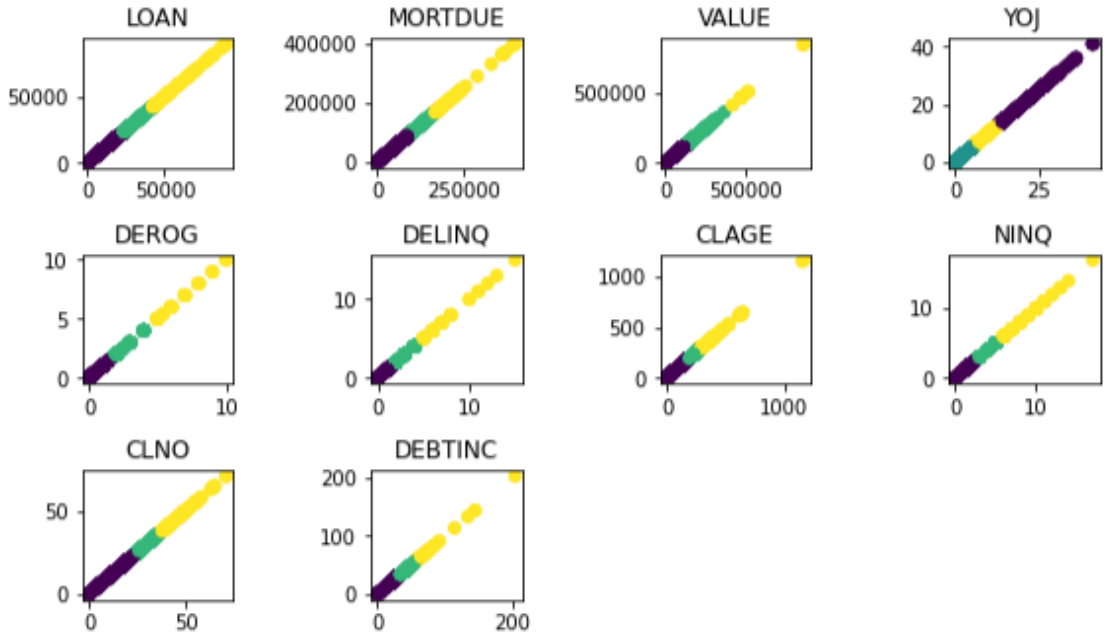
Annexe 09: KNN



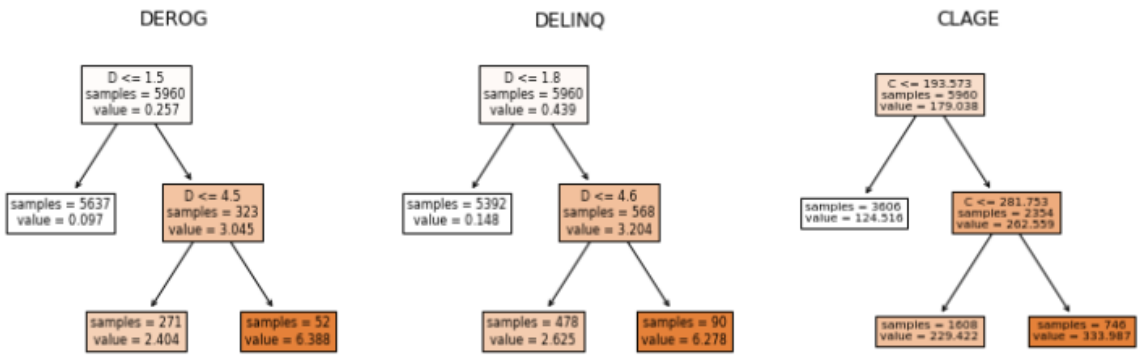
Annexe 10 : Discrétisation [K-means]



Annexe 11 : Discrétisation [Decision tree]



Annexe 12: Arbre de décision



Annexe 13: intervalles des variables

Variables	**Intervalle 01 (low)**	**Intervalle 02 (mid)**	**Intervalle 03 (high)**
LOAN	(1100-23900)	(24000-42700)	(42900-89800)
MORTDUE	(2063-93805)	(93868-166000)	(166244-399550)
VALUE	(8000-136877)	(137000-364000)	(415000-855909)
YOJ	(0-5)	(12.6-41)	(5.6-12.4)
DEROG	(0-1.4)	(1.6-4)	(5-10)
DELINQ	(0-1.6)	(2-4.2)	(5-15)
CLAGE	(0-193.50)	(193.63-281.57)	(281.92-1168.23)
NINQ	(0-2.4)	(2.6-5)	(6-17)
CLNO	(0-23.4)	(23.6-36)	(37-71)
DEBTINC	(0.52-32.62)	(32.64-56.39)	(56.96-203.31)

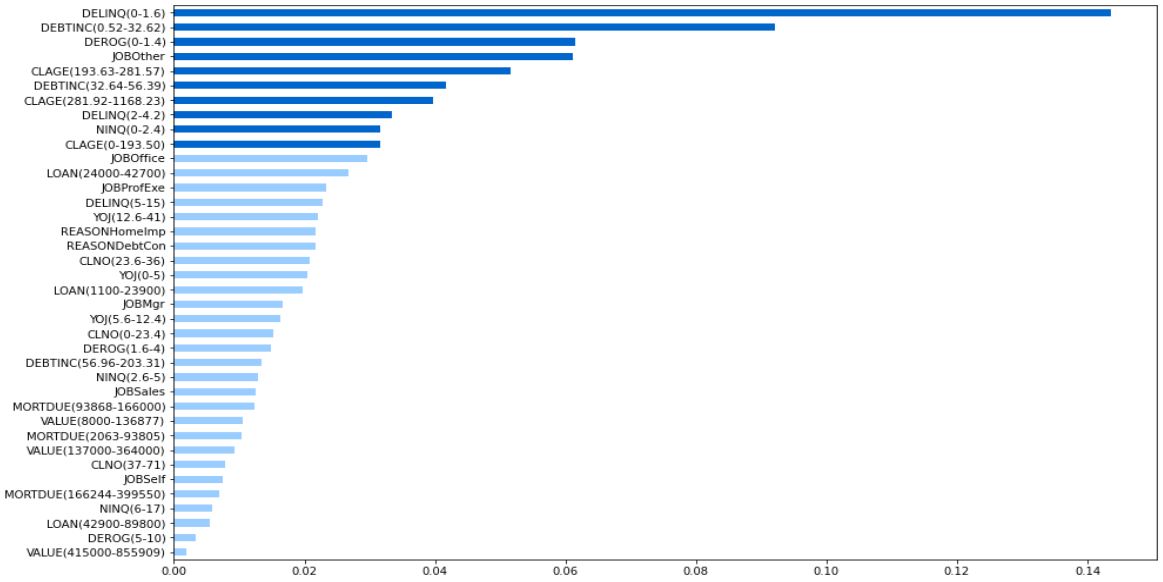
Annexe 14: résultats de la régression logistique

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.64325387070631	0.217855307219785	7.54286820769762	4.59745726660541e-14
DEROG_low	-1.53474461234313	0.14153570131219	-10.8435157922303	2.1408232762918e-27
DELINQ_high	3.82699654801995	0.612492979176521	6.24822925017889	4.15132065600614e-10
DELINQ_low	-1.64567130195402	0.115767399208775	-14.2153258447676	7.36080244162522e-46
CLAGE_low	1.03107378223212	0.0899911514715177	11.4575018251484	2.15645348305468e-30
NINQ_low	-0.804638169974194	0.0989168663575576	-8.13448908769154	4.13679234898185e-16
CLNO_high	0.488475237132323	0.162861484763693	2.99932938620254	0.0027057461635231
CLNO_low	0.417931941080738	0.0928726237848271	4.50005528054239	6.79357928949112e-06
DEBTINC_high	17.3955163022383	268.279632210116	0.0648409875879587	0.948300606721078
DEBTINC_low	-0.857887795051877	0.0909535941617648	-9.43214837146608	4.01774001452901e-21
REASON_1	-0.26099197446786	0.0835773133488892	-3.12276099829103	0.00179163172605731
JOB_3	-0.340732006054932	0.101670471228706	-3.35133694117008	0.000804223851953825
JOB_5	-0.959750225701307	0.121377003329339	-7.90718339863079	2.63277622284534e-15
YOJ_mid	0.245427329508114	0.0781278946560481	3.14135342554139	0.00168168961305889
LOAN_mid	-0.326592765034243	0.106020360343442	-3.08047212795995	0.00206672697198858

Annexe 15: Métriques d'évaluation de Random Forest

Metric	**Value**
Accuracy	0.7964
Precision	0.7035
Recall	0.8243
F1 Score	0.7591
AUC	0.8015

Annexe 16: feature importance de Random Forest



Annexe 17: Matrice de confusion de Random Forest