Projet n°8 : Communiquez vos résultats.

Analyse des déterminants à l'octroi de crédit bancaire destiné à des particuliers.

Sommaire

La finance subit une révolution.

Big Data, intelligence artificielle ou encore données alternatives: autant de concepts qui sont en train de profondément transformer la manière d'analyser les données, de gérer les actifs et d'interagir avec son gérant. La révolution en marche dans la finance est stimulée par l'explosion du volume d'informations numérisées: 90% des données dans le monde ont été créées au cours des deux dernières années et seulement moins de 1% serait activement analysé aujourd'hui. La marge de progression est considérable. Pour l'heure, ces données sont parfois accumulées dans ce que l'on appelle des «lacs de données» (data lakes). Mais ces derniers ne sont pas exploitables tant que l'on ne crée pas de systèmes capables de les analyser de manière autonome et évolutive.

C'est là qu'entrent en jeu l'apprentissage automatique (machine learning) et l'intelligence artificielle, qui annoncent un bouleversement profond de l'activité bancaire, qu'il s'agisse d'améliorer le conseil aux clients en proposant des produits adaptés à leurs habitudes de consommation, d'analyser l'activité des comptes pour prédire un besoin, ou d'utiliser des données alternatives pour améliorer la gestion. Les institutions financières attendent beaucoup de l'autonomisation des processus, qui sont pour l'heure très longs, complexes et coûteux, surtout pour les institutions actives dans plusieurs dizaines de pays. Les solutions disponibles aujourd'hui sont très pointues, mais génèrent encore trop de résultats qu'ont pourraient qualifier de « faux positifs », c'est-à-dire que pour des cas complexes ou évolutifs, l'opération peut être validée par le système alors qu'elle ne devrait pas être effectuée. L'intelligence artificielle et le machine learning apportent des solutions dans ce domaine.

Nous nous intéressons ici principalement aux banques commerciales. Une banque commerciale est un établissement financier dont les activités, basiques, sont majoritairement tournées vers les particuliers (dépôts, placements, solutions d'épargne, crédit), les entreprises ou les collectivités publiques. Bien qu'il n'existe pas de segmentation officielle, on distingue généralement les banques commerciales des banques coopératives (dont les usagers sont également les sociétaires), des banques d'affaires (majoritairement tournées vers les activités de marché ou vers le conseil), ou des banques privées (majoritairement tournées vers les clients fortunés). Dans la pratique, la frontière entre ces divers types d'établissements est relativement ténue.

Les banques commerciales ont donc pour but d'apporter un appui financier aux entreprises, aux particuliers et à d'autres formes d'organisations pour promouvoir les investissements et relancer la production dans la société. Bien que l'octroi de crédit constitue une source importante des revenus pour les banques, la rentabilité de ce mécanisme dépend d'une part des modalités définies par ces banques et d'autre part, de l'appréhension de ces modalités par les clients. Pour accorder un crédit à la consommation, un prêteur s'aide de la notation crédit, c'est-à-dire d'un ensemble de modèles de décision et de techniques qui s'y rattachent. Ces techniques permettent de déterminer les futurs emprunteurs, les montants accordés et les stratégies opérationnelles qui amélioreront la rentabilité des emprunteurs pour les prêteurs, ainsi qu'évaluer le risque lié aux emprunteurs, la notion de risque s'évaluant sur la capacité à rembourser mais également la loyauté face à une obligation de paiement (rappelons simplement que crédit vient du latin « credere » qui signifie croire).

« On ne prête qu'aux riches » affirme cette citation. Pour une banque, la gestion du risque que représente le crédit est un aspect fondamental de leur activité. Les garanties présentées par les demandeurs ne sont souvent pas suffisantes, la banque a besoin de plus de données pour pouvoir se décider à prêter de l'argent, d'où le besoin de faire une notation crédit. Au final, se basant sur des données réelles, la notation crédit se veut une évaluation fiable de la solvabilité d'un client.

Il existe de nombreuses variantes et complexités concernant la manière dont le crédit est accordé aux particuliers, aux entreprises et à d'autres organisations à diverses fins (achat d'équipement, immobilier, biens de consommation, etc) et en utilisant diverses méthodes de crédit (carte de crédit, prêt, plan de paiement retardé). Dans tous les cas et quelles que soient les techniques utilisées, il est essentiel qu'il existe un large échantillon de clients précédents avec les détails de leur application, leurs modèles de comportement et leurs antécédents de crédit disponibles.

Nous tenterons d'identifier les critères de sélection qui ont donc été déterminant à cet octroi. Puis, par le biais d'une modélisation prédictive basée sur ces clients en particulier (leur particularité est soulignée car le modèle évoluerait forcément avec l'ajout de données de clients différents, d'où l'intérêt de l'apprentissage profond), nous le testerons sur un jeu de données-clients afin de simuler les accords possibles de prêt par la banque et vérifier la fiabilité du modèle.

Nous nous donnons donc pour cadre de travail celui d'un site bancaire proposant une première simulation d'accord de prêt à de potentiels futurs clients, avant une expertise plus minutieuse des risques liés à l'emprunt.

Cette simulation aura pour vocation de laisser une certaine souplesse dans la réponse, afin d'inciter une personne, dans le cas d'une réponse positive, à se rapprocher d'un conseiller, qui ne manquera pas de lui présenter une gamme de services bancaires adaptée à sa demande, et l'amener à se fidéliser.

I Description des données.

Nous étudierons donc les données d'un jeu récupéré sur la plateforme Kaggle, base de données dans de nombreux domaines, et accessible en libre service. Ces données sont consultables dans le domaine public depuis décembre 2018. La véritable datation de ces données n'est en revanche pas connue.

Elles se composent d'un échantillon d'entraînement de 614 clients et d'un échantillon-test de 50 clients, qui sera mis de côté pour le moment.

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0	NA	360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N

Figure 1 – Extrait de l'échantillon d'entraînement

Les colonnes associées à ces clients sont divisées en deux catégories : des données qualitatives (nominales et ordinales) et quantitatives.
La notation crédit (en rouge) a été déterminé en fonction de l'appartenance ou non des clients à certaines catégories sociales et professionnelles (identifiant, genre, statut marital, nombre de personnes à charge, diplôme, antécédents de crédit,etc) mais également en fonction des revenus du ménage, du prêt demandé, de la durée de remboursement demandé, et des antécédents de crédit.

Remarque : Ces antécédents de crédits sont ici des valeurs nominales, contrairement à ce qu'elles laissent au départ sous-entendre.

1. Nettoyage des données.

a) Valeurs aberrantes.

Nous remarquons très vite des valeurs aberrantes concernant les revenus propres à chacun des conjoints.

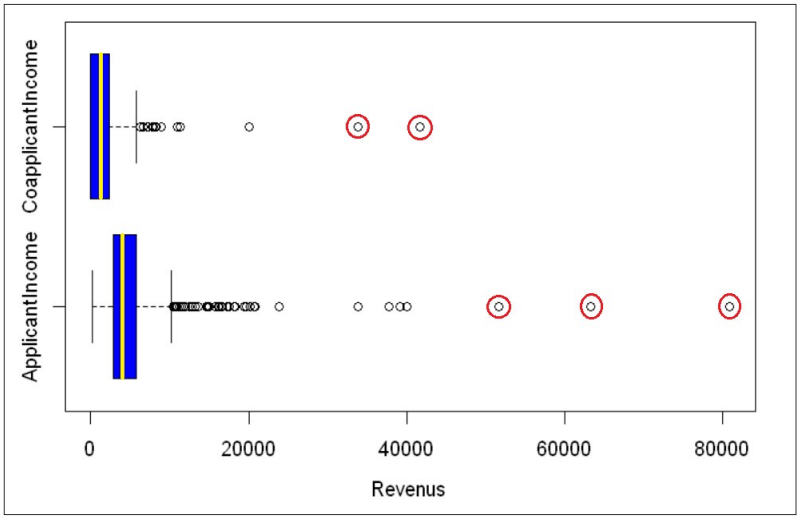


Figure 2 – Répartition des revenus du demandeur et du codemandeur.

Bien que ces valeurs soient extrêmement grandes, il n'est pas impossible que la banque aient au sein de sa base de données des clients possédant de tels revenus. Nous ne considérons alors pas ces valeurs comme aberrantes.

On regarde également les données clients relatives aux deux valeurs les plus extrêmes des revenus :

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Male	Yes	3+	Graduate	No	81000	0	360	360	0	Rural	N

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Female	No	3+	Graduate		416	41667	350	180	NA	Urban	N

Figure 3 – données clients de deux valeurs extrêmes.

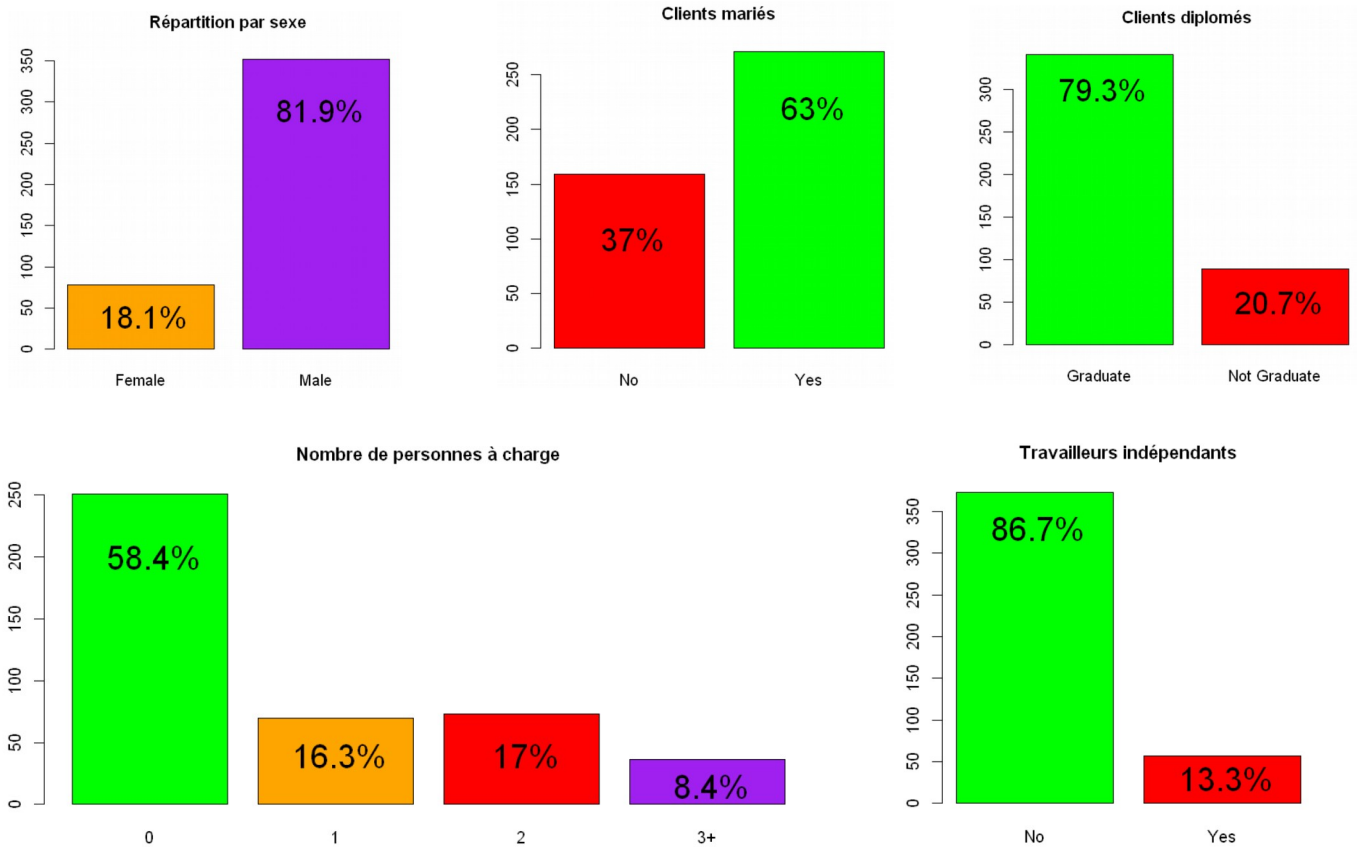
Étonnamment, ces deux clients n'ont pas obtenu leur demande de prêt, ce qui laisse déjà supposer que le modèle employé auparavant par cette banque ne semblait pas axer les octrois sur les revenus du ménage.
Aucun problème concernant la recherche de doublons.

b) Valeurs qualitatives et quantitatives manquantes.

Le choix est fait de supprimer toutes les valeurs manquantes, afin de ne pas imputer par des valeurs fictives et fausser l'analyse exploratoire des données ou l'analyse prédictive.
Nous nous retrouvons alors avec un jeu de 480 clients, soit une perte de 22 % des données initiales.
De plus, on retire une cinquantaine de données du jeu afin d'en créer un nouveau, dans le but uniquement de tester plus tard la fiabilité de la modélisation avec des clients non testés par l'apprentissage et comparer leurs notations crédit connus d'avance avec celles de la modélisation (voir « III. 2) a) Essai »).

2. Analyses univariées et bivariées.

a) Analyses univariées.



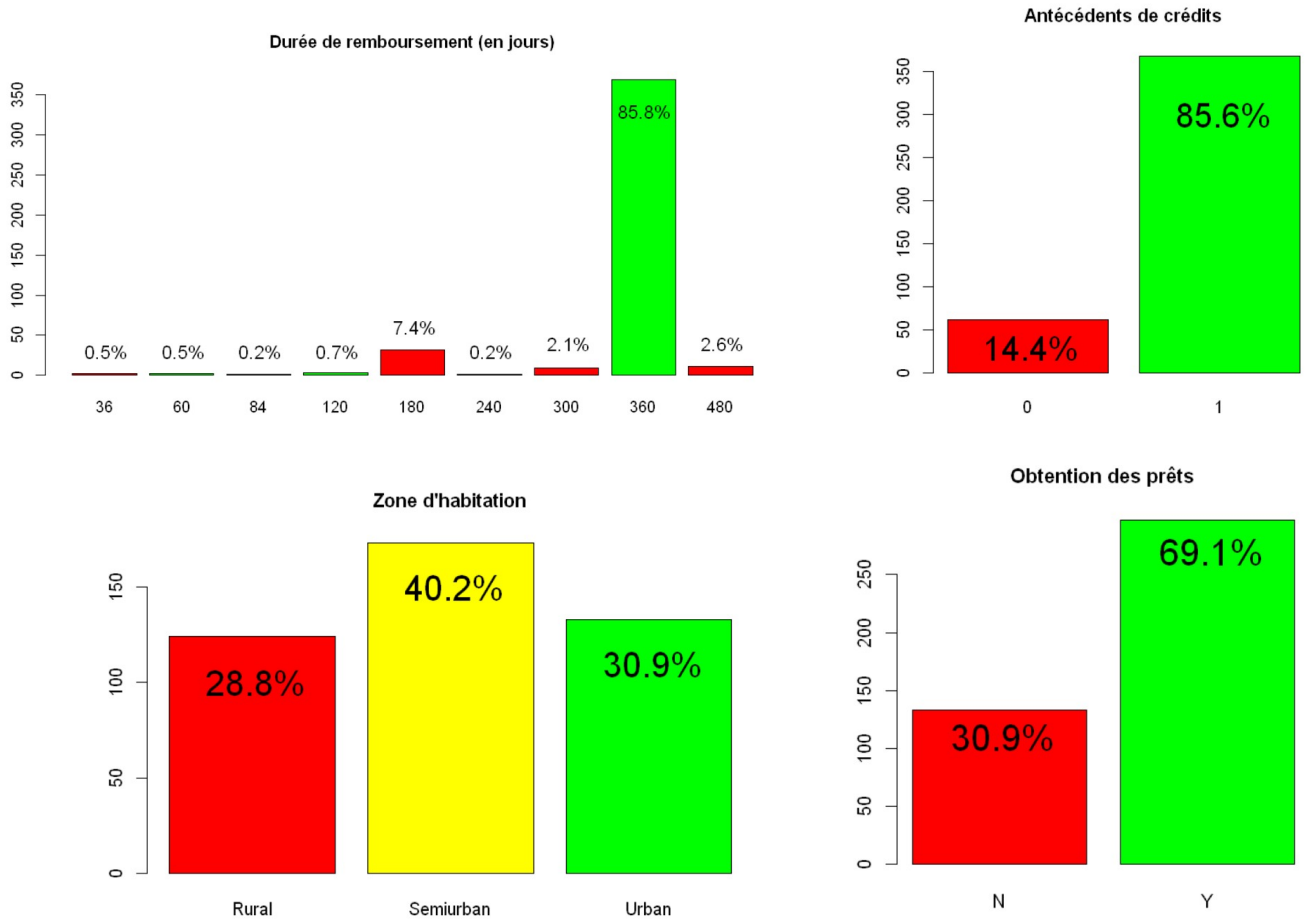


Figure 4 – Répartitions des valeurs nominales en fonction des catégories

b) Analyses bivariées.

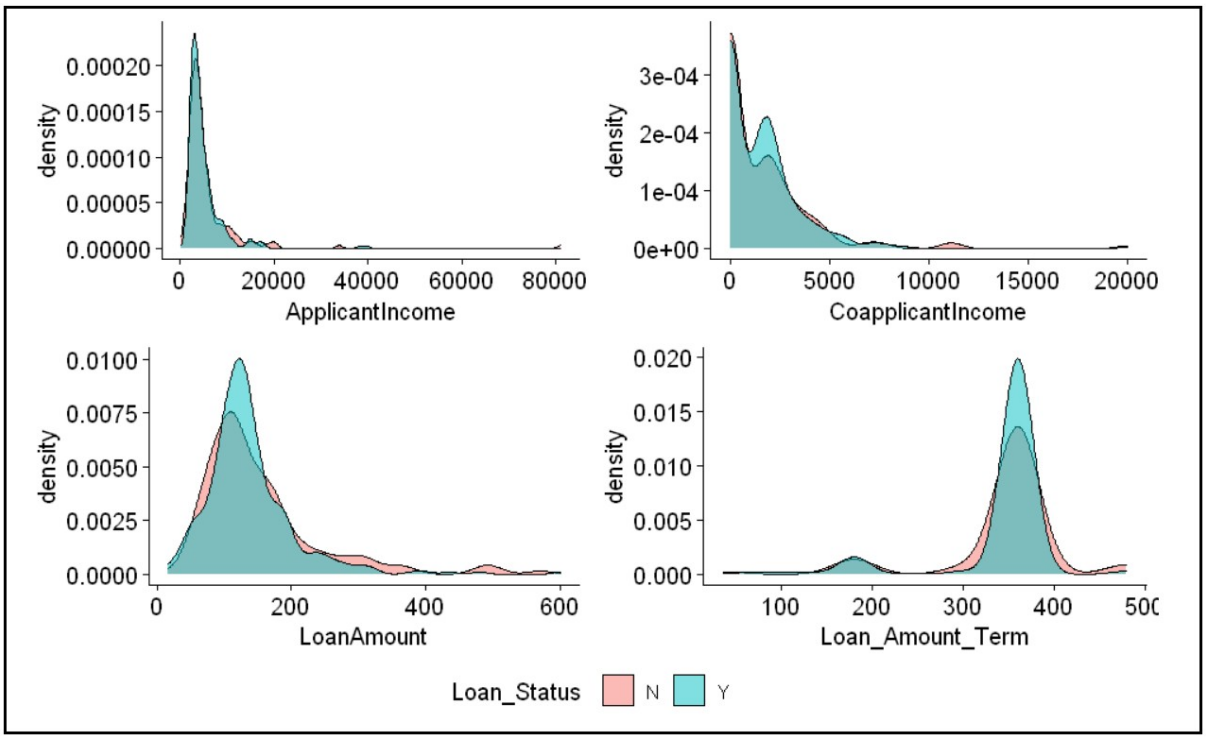


Figure 5 - Courbes de densité de la variable « Statut du prêt » en fonction de chaque variable quantitative.

On observe facilement dans un premier temps que les variables quantitatives n'ont l'air en rien déterminantes dans l'accord des prêts.

Test de normalité :

On effectue maintenant une ANOVA sur les variables quantitatives en fonction de la variable catégorielle « Loan_Status ».
Il faut, avant cela, s'assurer de la normalité de nos variables à expliquer.
Pour cela, on observe tout d'abord un graphe quantile-quantile :

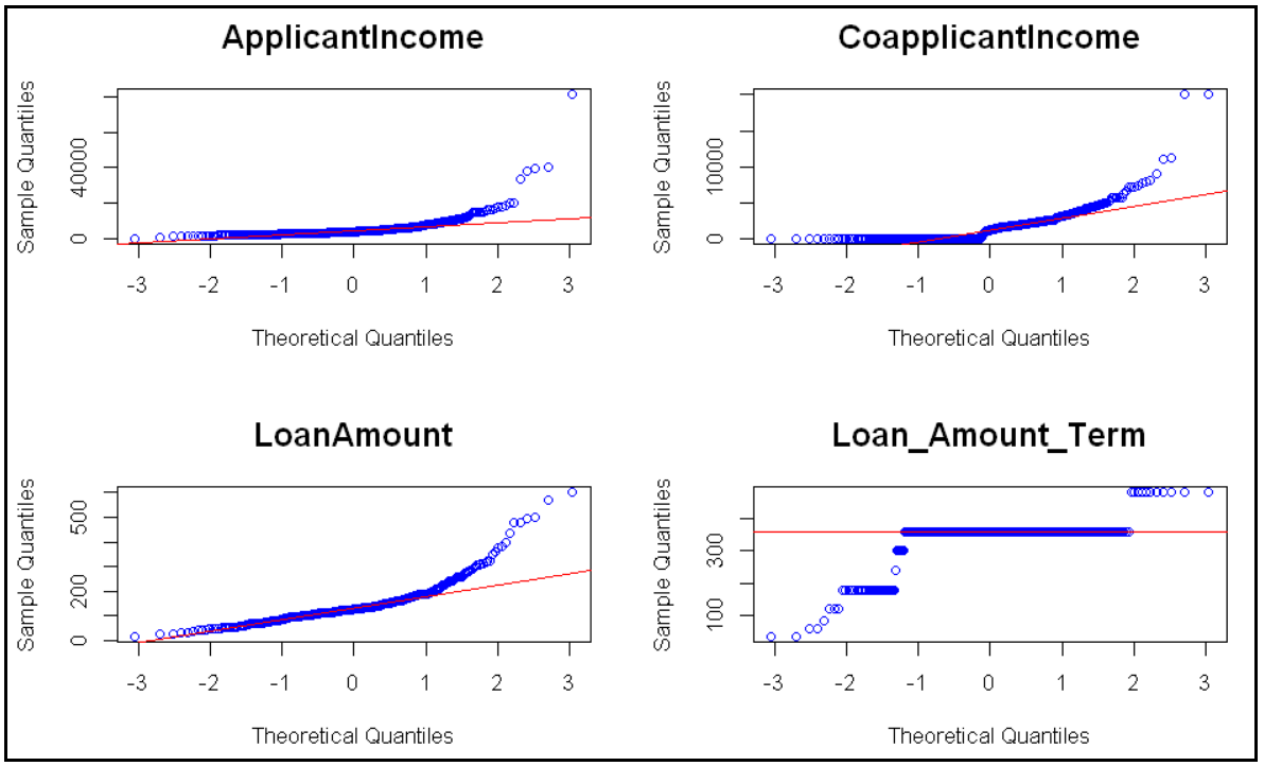


Figure 6 – Graphe quantile-quantile de chaque variable quantitative

Les variables semblent graphiquement suivre une loi normale.

On procède à un test de Shapiro-Wilk pour vérifier que nos variables suivent bien une loi normale (hypothèse nulle H0).
Les tests sont significatifs au seuil de 1 % : H0 est rejetée à chaque fois.
Les quatre variables ne suivent donc pas une loi normale.
Or, étant donné la taille de notre échantillon (supérieur à 30) et d'après le Théorème central Limite, (et au vu des visualisations précédentes) nous pouvons estimer que les variable suivent une loi gaussienne.

On peut donc leur appliquer la statistique du test de Fisher, qui cherche ici à montrer que les moyennes des modalités propres à chaque variable sont différentes (H1) , afin de prouver la significativité des variables dans la notation crédit.
Au final, seul le test de la variable LoanAmount est significatif au seuil de 5 % (bien que sa p-value de 4.2 % soit relativement proche du seuil de risque).
LoanAmount est donc déterminante dans la notation crédit.

Test de Chi² d'indépendance :

On souhaite maintenant vérifier la corrélation entre les variables qualitatives et Loan_Status. On réalise donc un test de Chi² d'indépendance afin de vérifier quelles sont les variables liées à Loan_Status (Hypothèse alternative H1) au seuil de risque 5 %:

Les p-values des variables Married, Credit_History et Property_Area sont significatives, l'hypothèse nulle est donc rejetée : ces variables étant liées avec Loan_Status, elles sont alors prépondérantes à la notation crédit.

c) Corrélations entre toutes les variables.

Après avoir transformé en valeurs numériques toutes les variables qualitatives, nous regardons leur matrice de corrélation sous forme de carte de chaleur :

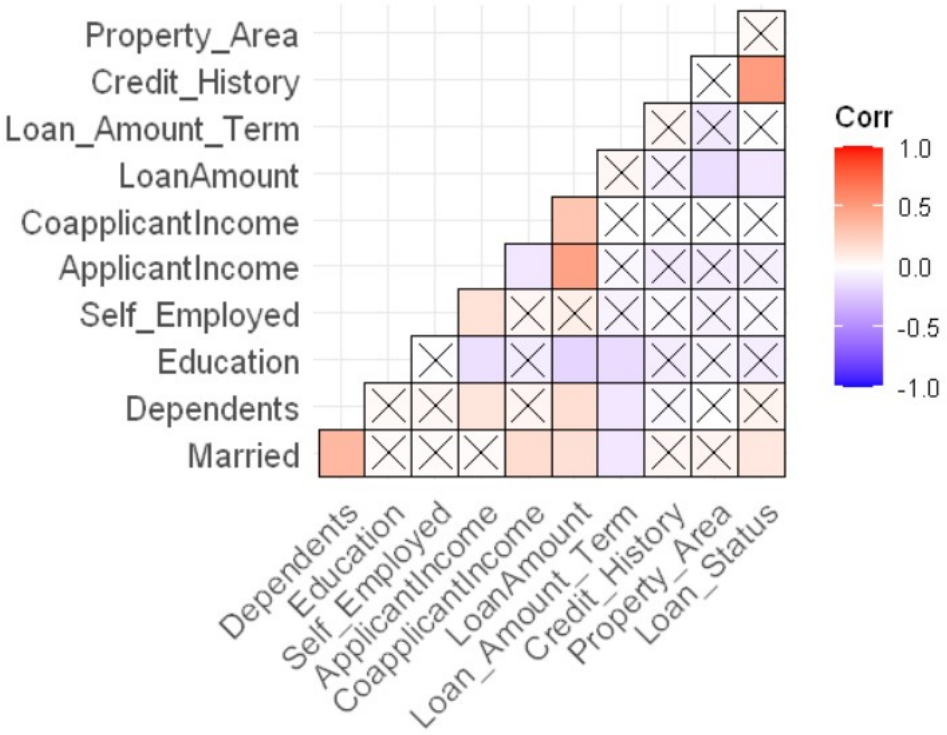


Figure 7 – Corrélation entre toutes les variables sur l'ensemble des prêt

Il y a donc une forte corrélation entre les antécédents de crédit et l'accord du prêt. On retrouve aussi une corrélation entre l'accord du prêt et le statut marital, ainsi que le montant du prêt. On ne retrouve pas en revanche de lien entre Loan_Status et Property_Area (de même, le fait d'être marié, d'avoir des personnes à charges ou d'être diplômé semble être corrélé à la fois avec la durée demandée de remboursement et avec le montant demandé).

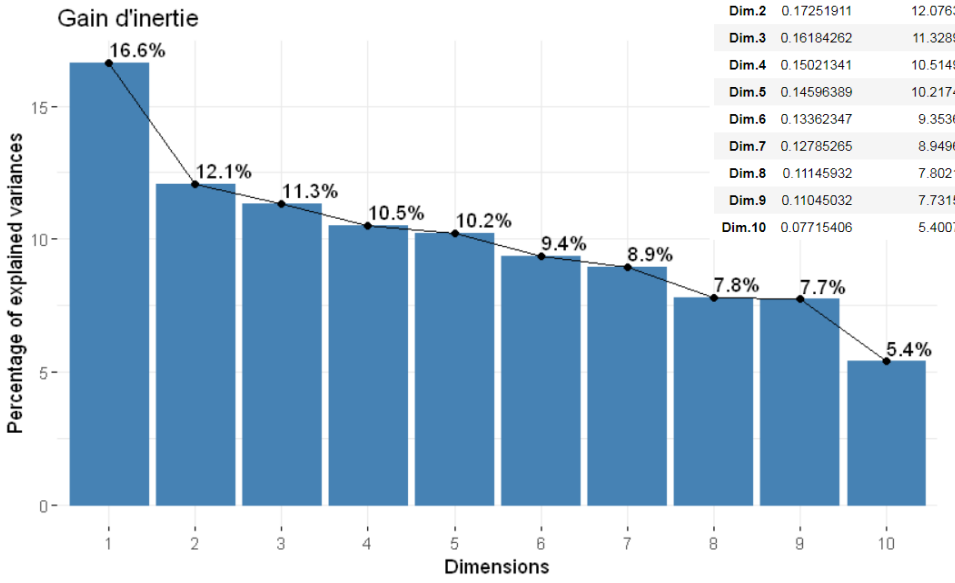
Conclusion: Loan_Status serait fonction des variables LoanAmount, Married, Credit_History et Property_Area.

II Analyse exploratoire des données.

Loan_Status est placé en variable supplémentaire afin de ne pas influencer la construction des composantes et ainsi voir la façon dont se regroupent les modalités des variables autour des deux classes de Loan_Status.

1. Détermination du nombre de composantes principales.

a) Étude de l'inertie.



	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.23749258	16.624481	16.62448
Dim.2	0.17251911	12.076338	28.70082
Dim.3	0.16184262	11.328983	40.02980
Dim.4	0.15021341	10.514938	50.54474
Dim.5	0.14596389	10.217472	60.76221
Dim.6	0.13362347	9.353643	70.11586
Dim.7	0.12785265	8.949685	79.06554
Dim.8	0.11145932	7.802153	86.86769
Dim.9	0.11045032	7.731522	94.59922
Dim.10	0.07715406	5.400784	100.00000

Figure 8 – Gain d'inertie

Selon le critère du coude, qui semble être le plus parcimonieux dans notre étude, 4 axes permettent d'expliquer plus de 50 % de la variabilité totale du nuage de points, ce qui reste acceptable.

2. Analyse des correspondances multiples.

a) Etude des plans factoriels à observer :

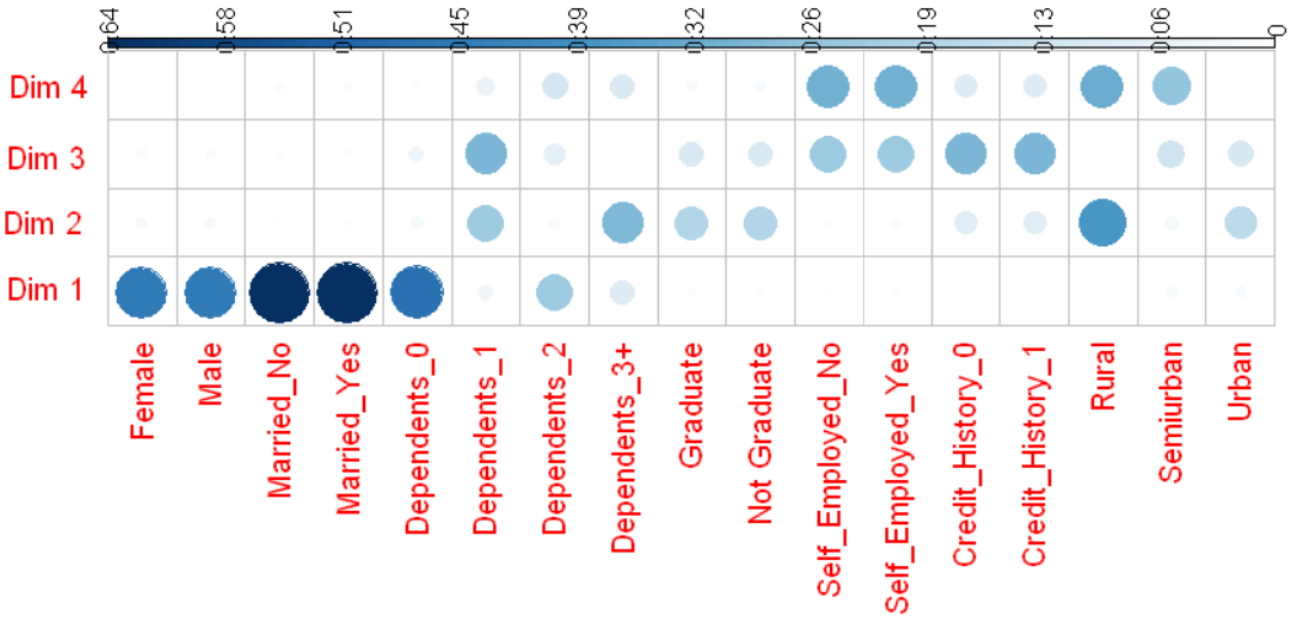


Figure 9 – Qualité de représentation de toutes les modalités sur chacune des composantes choisies

La qualité de représentation des modalités et leur nombre sur la carte de l'ACM déterminent le nombre de plans factoriels utiles à l'analyse : Ici c'est assez compliqué car les qualités de représentations des trois derniers axes sont faibles. Tous les plans sont au final observés pour essayer d'avoir la lecture la plus lisible des individus. Autant les graphes des modalités permettent encore de lire les modalités les plus essentielles à la détermination des groupes de classes, autant aucun graphe des individus ne semble pouvoir nous éclairer.
Nous illustrerons donc notre propos avec les plans factoriels choisis parmi les moins « obscurs » à analyser...

b) Plan factoriel de dimension 1 et 3.

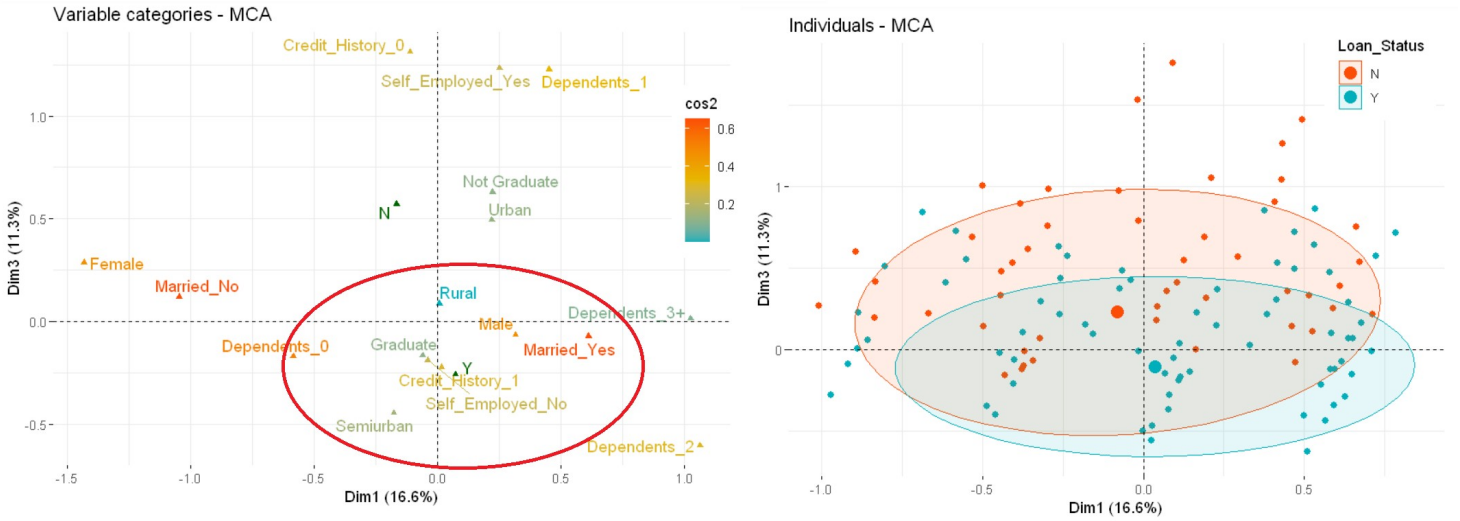


Figure 10 – Graphe des modalités + graphe des individus

On observe un premier groupe de modalité autour de la modalité Yes de Loan_Status : Credit_History_1, Self_Employed_No, Graduate, Semiurban, MarriedYes, Male (Rural n'est pas compté car mal représenté).
Le graphe des points ne permet pas l'identification de deux groupes distincts.

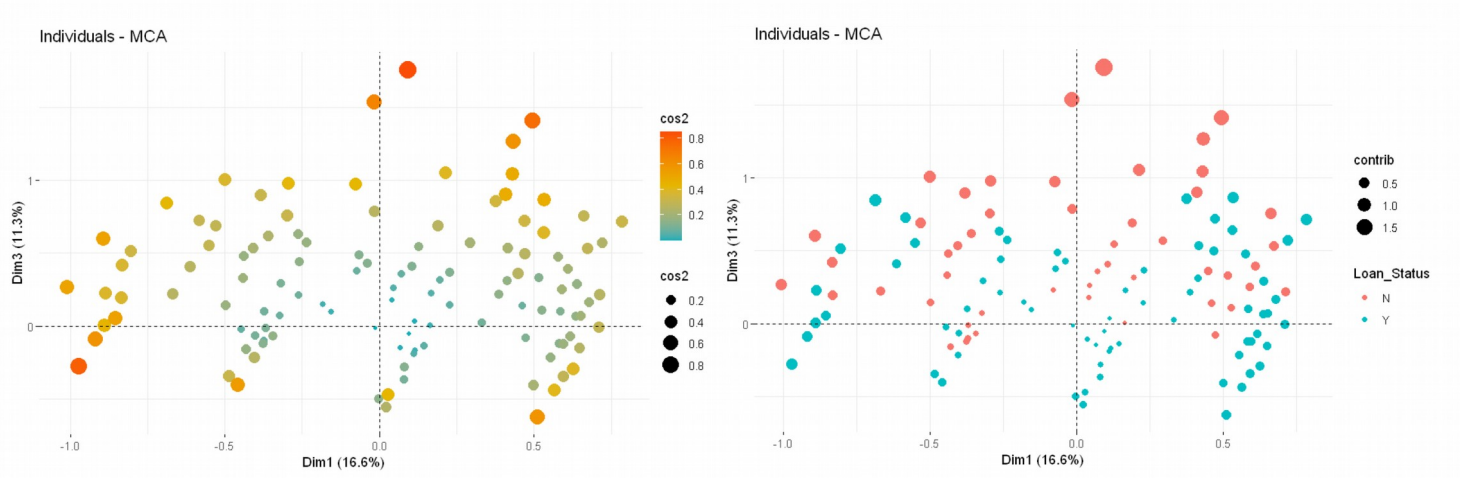


Figure 11 – Graphe de représentation des individus + Graphe de contribution des individus

Le graphe de représentation des individus indique sans surprise que plus les individus s'éloignent du centre, mieux ils sont représentés. Celui de la contribution montre une légère tendance vers le bas des clients ayant obtenu un prêt mais reste toutefois peu éloquent.

c) Plan factoriel de dimension 3 et 4

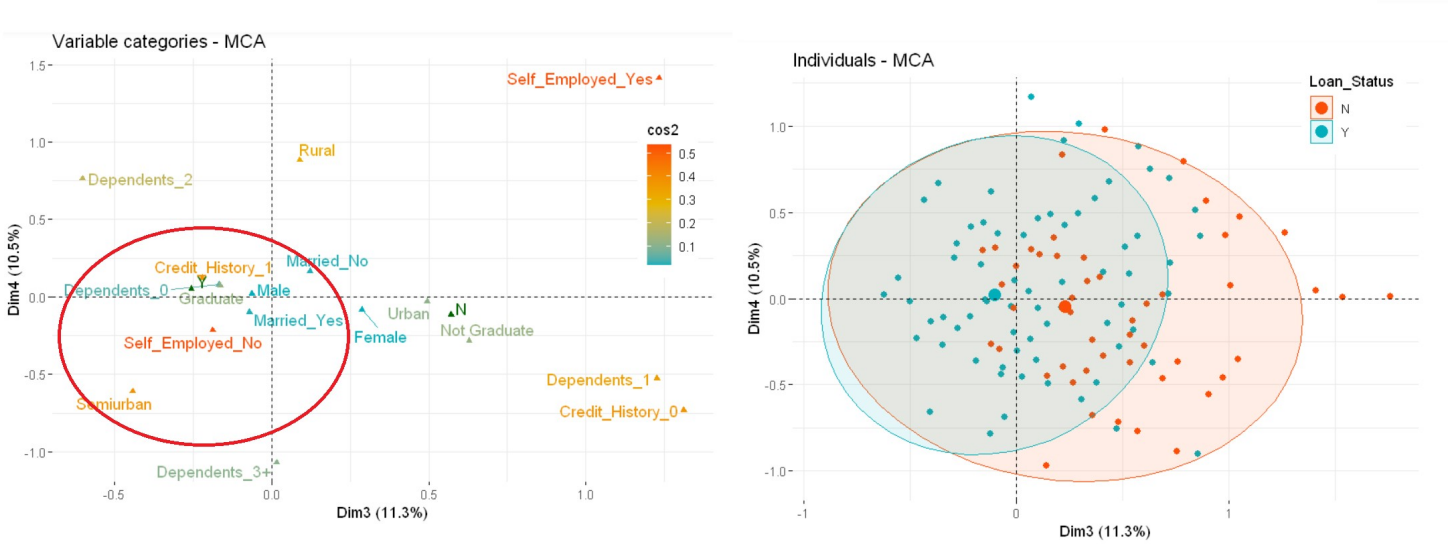


Figure 12 – Graphe des modalités + graphe des individus

Ici, sur ce graphe des modalités, les modalité proches de la modalité Yes de Loan_Status sont: Credit_History_1, Graduate, Married_Yes et Male(bien que mal représentées), Self_Employed_No, Semiurban.
Le graphe des individus ne permet toujours pas une différenciation des classes de Loan_Status, bien que les individus de la modalité Yes soient majoritairement placés sur la gauche.

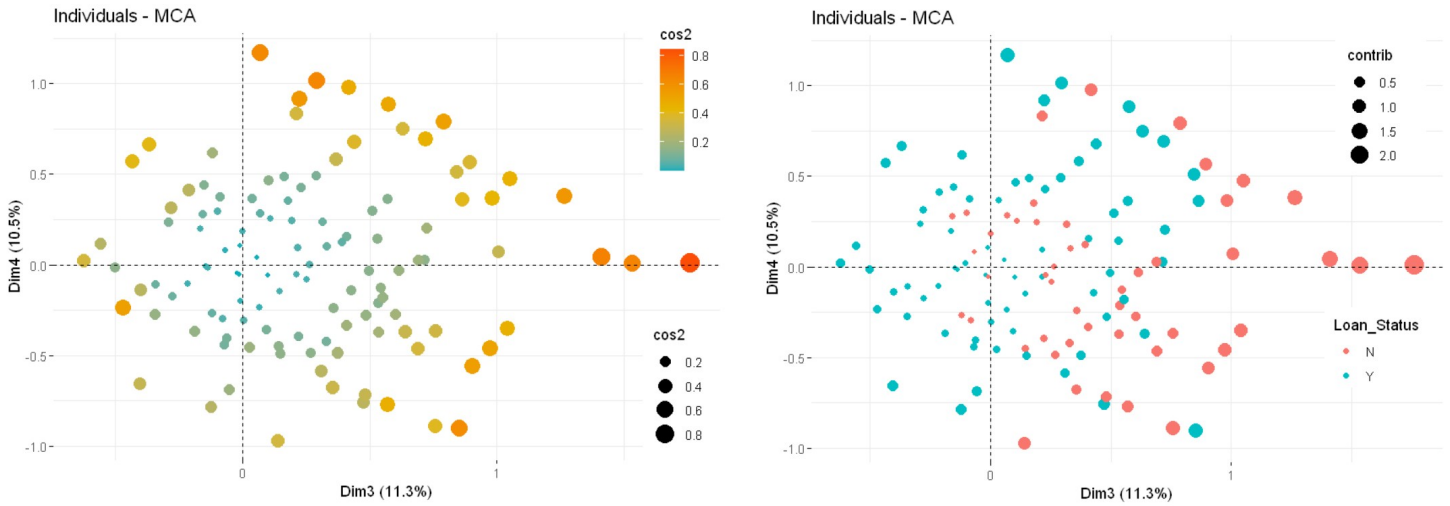


Figure 13 – Qualité de représentation des individus + contribution des individus

La qualité de représentation des individus est très forte à la droite du plan (les clients s'étant vu refusé un prêt). La contribution des individus permet une meilleur séparation des classes de Loan_Status, sans être encore une fois très probant.

Conclusion : On ne peut donc ici que regretter l'aspect limitatif de la visualisation de l'exploration des données. L'étude des graphes des modalités a tout de même permis de conforter certains déterminants qui auraient joués un rôle dans l'accord des prêts, comme le fait d'avoir eu par le passé un antécédent de crédit ou d'être marié : la banque semble privilégier de manière évidente les personnes qui ont un jour contracté une dette.

Les graphes des individus apportent peu de compréhension. Néanmoins on peut deviner plus facilement les clients dont le prêt est refusé (le groupe des prêts accordés étant contenu dans celui des refusés).
Cela nous permettra donc d'envisager cette certaine souplesse de l'octroi de prêt abordé plus haut et donc de répondre au « cahier des charge » fixé dans l'optique de la simulation prévu sur le site internet.

III Analyse prédictive.

1. Modélisation à l'aide d'une régression logistique multiple à variable binaire.

a) Modélisation.

```
# Paramétrage du processus d'apprentissage par validation croisée stratifiée:
train.control <- trainControl(method = "cv", number = 5)
```

Figure 14 – Validation croisée stratifiée

Choix de la validation croisée stratifiée :

On commence par paramétrer le processus d'apprentissage en procédant par validation croisée stratifiée sur un nombre de partitions choisi au départ (ici 5).
Le jeu de donnée contient 430 clients. Le choix de la validation croisée est justifiée ici par le fait qu'elle est utilisée pour évaluer l'ajustement du modèle lorsque la taille de l'ensemble de données est limitée.
Elle divise de manière itérative l'ensemble de données en deux parties: un jeu de test et un jeu d'entraînement.
En choisissant 5 partitions, notre modèle sera donc entraîné sur 80 % des données et ce 5 fois (soit 344 clients à chaque fois, ce qui sera amplement satisfaisant).
Le choix du nombre de divisions a un impact sur le biais (la différence entre la valeur moyenne / attendue et la valeur correcte - c'est-à-dire l'erreur) et la variance. En règle générale, moins il y a de divisions, plus la variance est faible et plus le biais / l'erreur est élevé (et vice versa).
La stratification sert à réarranger les données afin de s'assurer que chaque partition possède la même fréquence distributive de classe.
Au final, dans ce processus, tous les clients auront au moins servi une fois dans un jeu de test et un jeu d'entraînement tout en respectant le principe selon lequel on ne fait pas de test sur des données qui ont servi à l'entraînement.
Les erreurs de prédiction de chacun des jeux de test sont ensuite moyennées pour déterminer l'erreur de prédiction attendue pour l'ensemble du modèle.

```
# Méthode de sélection pas à pas "Both", basée sur le critère d'Akaike:
model <- train(Loan_Status ~ ., data = pret, method = 'glmStepAIC', family = 'binomial', trControl = train.control)
model
```

Figure 15 – Entraînement du modèle

Explication de la méthode employée :
On va donc appliquer un modèle linéaire généralisé sur le jeu de données puisque les critères suivants sont vérifiées :
- La variable Loan_Status est une variable qualitative binaire puisqu'elle contient deux modalités à savoir Yes et No. Elle sera donc la variable à expliquer.
Dans le cas d'une variable binaire, la variable suit une loi binomiale, appartenant à la famille exponentielle.

- Les variables explicatives (ou prédicteurs) sont donc les variables quantitatives que l'on a étudiées auparavant et sont au nombre de 11.
Elles sont susceptibles d'influencer l'issue de la variable à expliquer.
- La fonction de lien modélise le logarithme du rapport de côtes. Elle est appelée logit.

On utilise la méthode pas à pas « both » (ou stepwise), mélange de méthode Backward et Forward : C'est la méthode la plus performante et la plus utilisée en pratique. On part de toutes les variables disponibles et on enlève au fur et à mesure les variables non significatives puis on les remet à nouveau et ainsi de suite.

Ici on s'appuie sur le critère d'Akaïke (ou AIC): Ce critère traduit la complexité des modèles en les pénalisant au fur et à mesure que l'on rajoute des variables (donc des paramètres à estimer) et permet donc de satisfaire le critère de parcimonie.
Le modèle fait en sorte de minimiser l'AIC, mais il peut y avoir des exceptions quand il estime que des paramètres supplémentaires permettent une meilleur approximation de la variable explicative.

```
Generalized Linear Model with Stepwise Feature Selection

430 samples
 11 predictor
  2 classes: 'N', 'Y'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 344, 344, 344, 345, 343
Resampling results:

  Accuracy   Kappa
  0.8068852  0.4740128
```

Figure 16 – Résumé sommaire de la modélisation

Ici on observe que le modèle entraîné et testé par validation croisé stratifié a un taux de succès de 80.6% (en rouge) ce qui est correct, et un coefficient Kappa de Cohen (qui quantifie la concordance entre deux tests) de 47.4%, ce qui pourrait sûrement être amélioré.

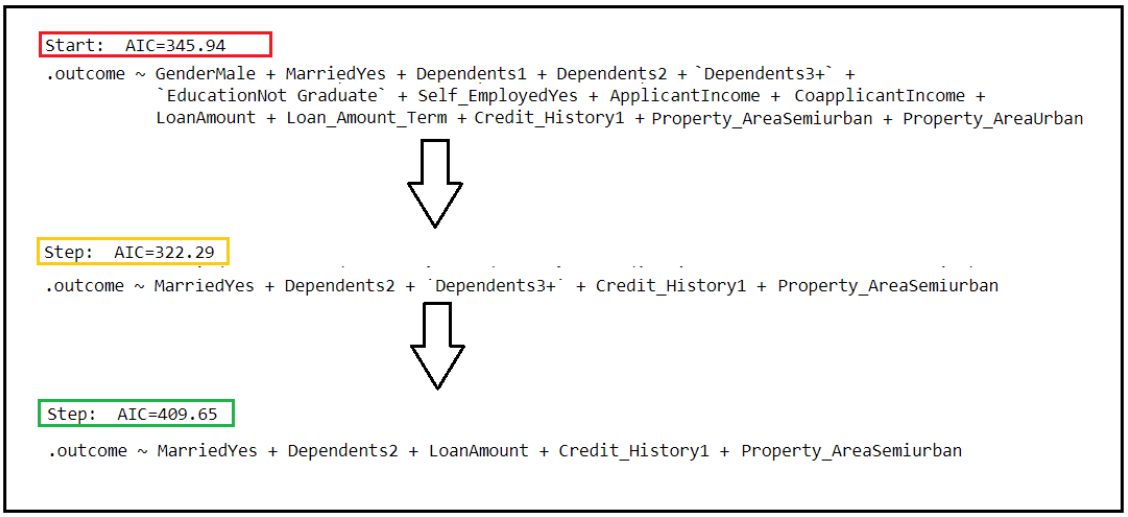
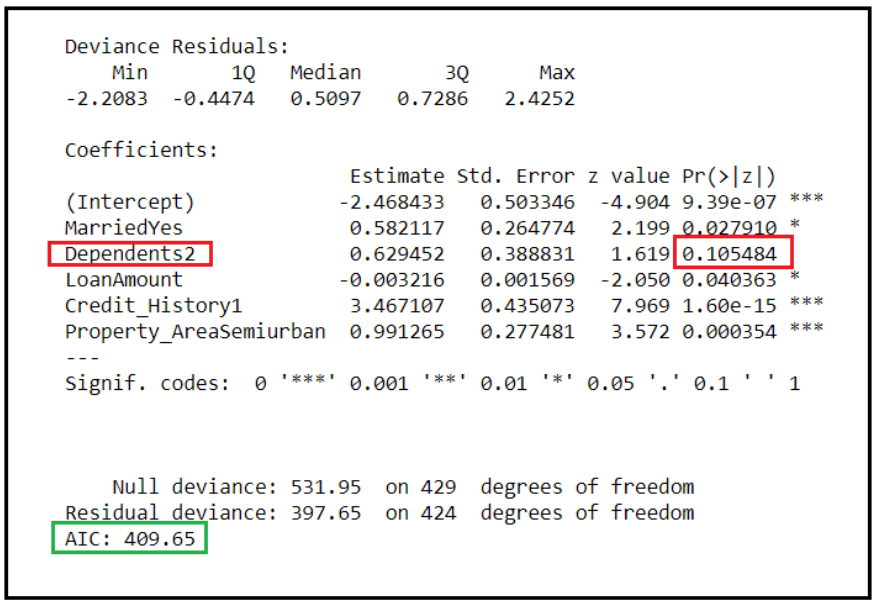


Figure 15 – Étapes de recherche

On voit que le modèle a bien débuté avec toutes les variables explicatives et s'est terminé après une soixantaine d'itération avec 5 variables dont on savait déjà que quatre d'entre elles étaient corrélés à la variable binaire : le modèle a tenu compte au final de la modalité Dependents_2.

On constate également l'évolution du critère d'Akaïke : le modèle s'est finalisé sur un critère plus grand qu'au départ, bien qu'il soit passé par un stade où le critère était minimisé. Il a donc estimé, malgré la hausse de l'AIC, que ces prédicteurs modélisaient mieux la variable à expliquer Loan_Status.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2083  -0.4474   0.5097   0.7286   2.4252

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.468433   0.503346  -4.904 9.39e-07 ***
MarriedYes               0.582117   0.264774   2.199 0.027910 *
Dependents2              0.629452   0.388831   1.619 0.105484
LoanAmount              -0.003216   0.001569  -2.050 0.040363 *
Credit_History1          3.467107   0.435073   7.969 1.60e-15 ***
Property_AreaSemiurban   0.991265   0.277481   3.572 0.000354 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



    Null deviance: 531.95  on 429  degrees of freedom
Residual deviance: 397.65  on 424  degrees of freedom
AIC: 409.65
```

Figure 16 – Coefficients des paramètres et p_values

Il est intéressant de constater que la méthode de recherche des variables explicatives retient au final la variable Dependents_2, dont la p-value est supérieur à 10%.

Ces 5 variables permettront de déterminer la courbe sigmoïde de la fonction de répartition de la loi logistique logit.

b) Évaluation statistique de la régression.

On vérifie si le modèle de régression logistique fournit un ajustement adéquat pour les données :
Le test de rapport des vraisemblances qui compare la vraisemblance entre le modèle courant et le modèle saturé (le modèle dans lequel nous avons tous les paramètres) et le calcul de la p-value donnent la significativité globale du modèle.
Ici la statistique du rapport de vraisemblance suit une loi du chi2 à 5 degrés de libertés.
On cherche à rejeter l'hypothèse nulle H0 disant que tous les coefficients de la relation linéaire sont nulles.
Le test est significatif au seuil de risque de 5 %, on rejette bien l'hypothèse nulle. Le modèle est donc globalement significatif: il existe bien une relation entre les variables explicatives et la variable expliquée.

c) Courbe ROC.

Le modèle ayant été validé à plus de 80 % par les cinq variables explicatives, il est donc satisfaisant, on peut alors entraîner un nouveau modèle autour de ces cinq seules variables.

```r
# Nous pouvons donc utiliser à présent cette modélisation:
new_model <- train(Loan_Status ~ Married + Credit_History + Property_Area + LoanAmount + Dependents,
                   data = pret, method = 'glm', family = 'binomial', trControl = trainControl('none'))
```

Figure 17 – Modèle finale

Le score (la probabilité qu'un client ait sa demande de prêt acceptée) est calculé et aidera à la création du programme permettant la simulation.

```
# Calcul du score, probabilité pour un client que sa demande de prêt soit accordée:
score <- predict(new_model, pret, type = 'prob')[, "Y"]
quantile(score)
```

0%	0.0264013334250081
25%	0.64224941432258
50%	0.779224584870131
75%	0.864857477198893
100%	0.955613090572917

Figure 18 – Calcul du score

L'aire sous la courbe ROC du modèle qui en résulte est de 80.13 %:

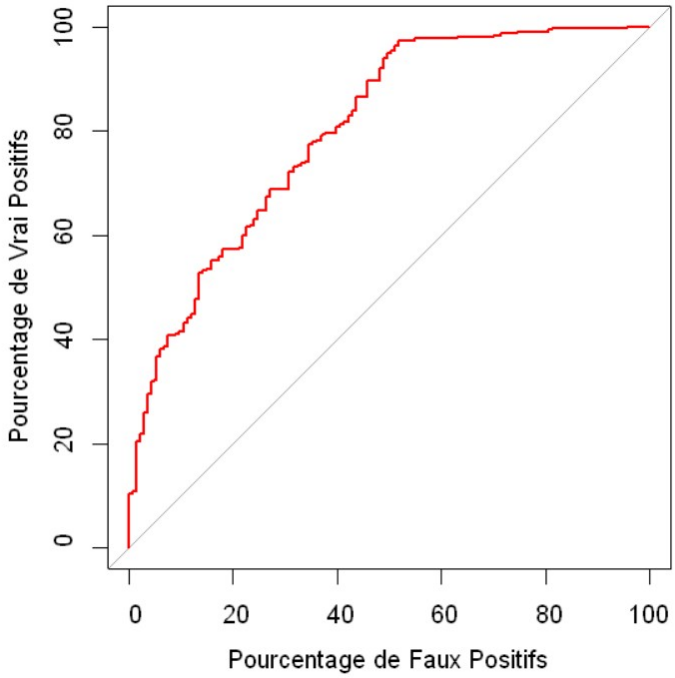


Figure 19 – Courbe ROC

On peut traduire ce résultat ainsi : En présence de deux demandes de prêt, l'un accordé l'autre pas, la probabilité que le modèle désigne correctement le prêt accordé est de 80.1%.

2. Test de notation du crédit.

a) Essai.

On peut maintenant écrire un programme qui servira à la simulation du site internet et le tester sur le jeu de donnée que l'on avait extrait au tout début et qui comportent 50 clients, dont 15 refus et 35 accords de prêts. On connaît les résultats des demandes. On peut donc les comparer avec les résultats du modèle :

	Loan_ID	score	Loan_Status
431	LP002785	0.745089718162566	Y
432	LP002788	0.078096738153420 5	N
433	LP002789	0.0902253795550881	N
434	LP002792	0.923326679307856	Y
435	LP002795	0.850150366479991	Y
436	LP002798	0.886047878429072	Y
437	LP002804	0.879389724718855	Y
438	LP002807	0.94553190274359	Y
...	...	...	...
474	LP002964	0.846220629272017	Y
475	LP002974	0.774415772645909	Y
476	LP002978	0.683587202302271	Y
477	LP002979	0.810326586271777	Y
478	LP002983	0.682878203745699	Y
479	LP002984	0.833240243240471	Y
480	LP002990	0.129548756195168	N

Figure 20 - Prédiction-essai

```
       Predit          Sensibilité: 100
Actuel  N  Y           Spécificité: 53.33
     N  8  7           Précision: 83.33
     Y  0 35           F-mesure: 90.91
```

Figure 21 – Matrice de confusion

La matrice de confusion indique 7 faux positifs :
- la sensibilité (probabilité de prédire correctement que le prêt est accordé) est maximale.
- la spécificité (probabilité d'avoir prédit correctement un prêt refusé) est d'un peu plus d'une fois sur deux.
Malgré cela, la précision de 83 % et la moyenne harmonique de 90 % restent tout à fait correcte.

b) Simulation.

On passe maintenant à la simulation proprement dite sur un jeu de données inconnues de 50 clients.

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360	NA	Urban
LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban

Figure 22 – Échantillon-test

Après un nettoyage des valeurs, il ne reste que 37 clients. Les données sont analysées par le modèle :

Loan_ID	score	Loan_Status
LP001015	0.773289988733954	Y
LP001022	0.764141200177223	Y
LP001031	0.823642362263827	Y
LP001051	0.678697271155592	Y
LP001054	0.748737767631716	Y
LP001055	0.858169543761897	Y
LP001056	0.150628378617232	N
LP001067	0.831222573500111	Y
...	...	...
LP001219	0.635249587007604	Y
LP001220	0.778879211220089	Y
LP001221	0.655852403884163	Y
LP001226	0.907423735584945	Y
LP001230	0.793581469490938	Y
LP001231	0.674474429102871	Y
LP001242	0.837883444209099	Y

Figure 23 – Prédictions du test

Sur 37 clients, nous obtenons 3 refus et 34 accords.

Nous observons les 3 refus pour essayer de voir les critères qui ont pu inciter le modèle à rejeter la demande :

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
7	LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural	N
20	LP001153	Male	No	0	Graduate	No	0	24000	148	360	0	Rural	N
28	LP001203	Male	No	0	Graduate	No	3150	0	176	360	0	Semiurban	N

Figure 24 – Observation des clients dont le prêt est refusé

Sans surprise, ces trois clients n'ont jamais eu d'antécédents de crédit.
Le modèle semble donc bien appliquer les priorités appliquées par la banque.

Vérification graphique : Les clients testés sont en bleu foncé.

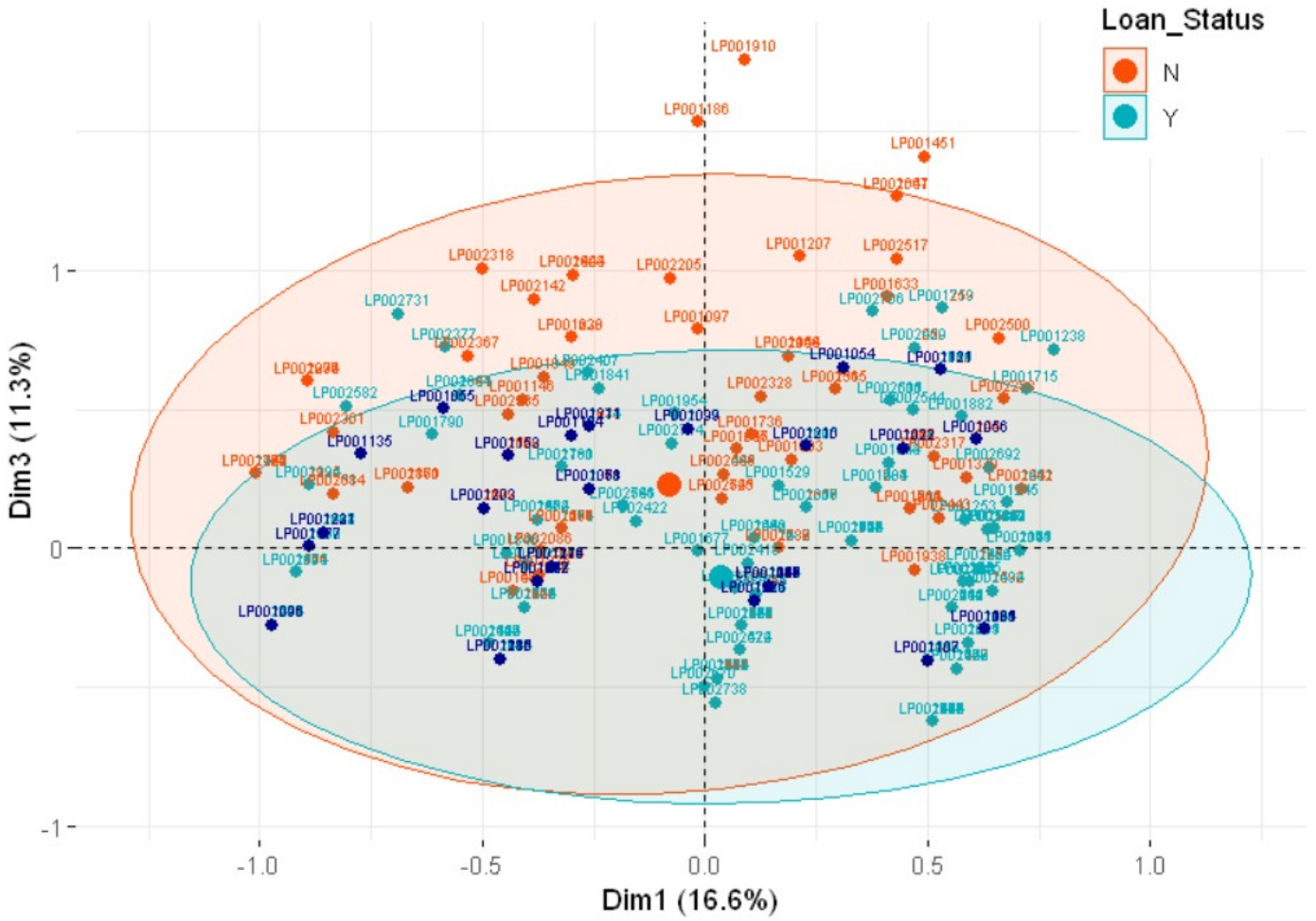


Figure 25 - Graphe des individus des dimension 1 et 3.

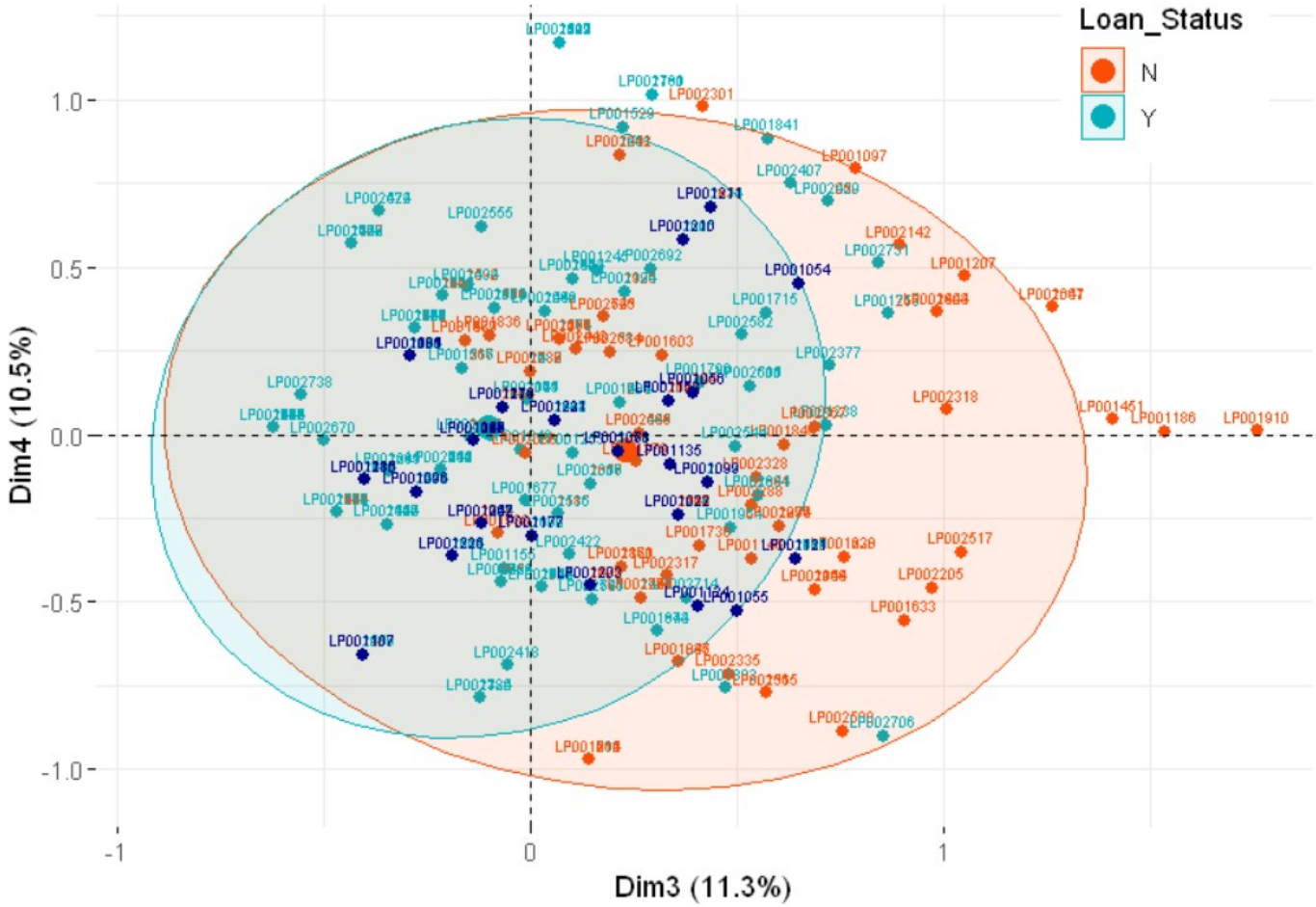


Figure 26 - Graphe des individus des dimension 3 et 4.

On peut constater que les clients dont le prêt est refusé sont très proches du centre de l'ellipse rouge (figure 26).

La banque pourra, si elle veut se référer à ces graphes, proposer éventuellement un entretien aux clients en fonction des demandes qui auront intégré l'ellipse bleu sans être trop proche du centre de l'ellipse rouge.

De façon plus restrictive, elle pourra les choisir en fonction de leur position sur les graphes: en dessous de l'axe des abscisses pour le graphe de la figure 25 et à gauche de l'axe des ordonnées pour le graphe de la figure é6.

CONCLUSION :

L’octroi de prêt s’est donc fait selon le statut marital, le nombre de personnes à charge, la zone habitée, les antécédents de crédit et le montant demandé.

Nous avons pu voir que les différents types d’approches en jeu ont mis en exergue la difficulté d’analyser finement les déterminants permettant d’expliquer l’octroi d’un prêt.
L’analyse par correspondance multiple a ici montré ses limites.
Au final, notre algorithme a pu affiner un modèle dont les critères de sélection n’étaient pas si intuitifs.

Cela montre que les domaines tels que l’intelligence artificielle et le machine learning apportent des solutions rapides et efficaces.