UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Projet de Scoring

Session 2019 - 2020

Par Paul LEYDIER, Louis GRUNENWALD et Wenceslas SANCHEZ

Table des matières

I). Introduction 3

II). Analyse exploratoire des données 4
1). Analyse des variables catégorielles 4
2). Analyse des variables numériques 6

II). Transformation et analyse des données 8
1). Traitement des valeurs manquantes 8
2). Transformation des données 9
3). Analyse des liens entre nos variables 10

III). Construction du meilleur modèle 13
1). Sélection de variable 13
2). Développement des modèles 14

IV). Conclusion : 19

Annexes : 20

I). Introduction

Les modèles de scoring attraient à tous les secteurs d'activités : prédire quel client va partir, prédire quel est le risque qu'un client fasse défaut, prédire l'appétence d'un prospect à nos services, prédire la panne, etc.

C'est une des principales problématiques auxquelles doit faire face un data scientist. En général, c'est un problème complexe car l'évènement considéré peut-être rare, les données à disposition ne sont pas suffisantes, ou il existe des contraintes réglementaires / de business intelligence qui limite l'utilisation de modèle plus intéressant mais moins compréhensible pour le métier.

Vous l'aurez compris, ce projet porte sur une problématique de scoring : prédire le risque de défaut d'un client pour une banque.

Notre approche pour ce projet consiste à générer 2 listes de variables « intéressantes », et de produire pour chaque liste 2 modèles. On pourra alors comparer 4 modèles entre eux. Mais avant d'en arriver à la modélisation, il nous faudra dans un premier temps explorer nos données et les traiter. Effectivement, on se rendra compte que les données à disposition ne sont pas propres, et nécessite certaines transformations afin d'avoir un dataset exploitable. Puis nous nous lancerons dans l'analyse des liens entre nos variables. Cette partie nous permettra de mieux comprendre nos modèles, et ainsi d'apporter des explications au métier à propos de la prédiction sortie.

II). Analyse exploratoire des données

La première étape dans tout projet de data science est l'exploration de nos données. Cette étape est primordiale car elle nous permet à la fois, de comprendre nos données, mais aussi d'identifier des patterns d'analyses, des problèmes etc.

A partir de la Figure 1, on peut déjà remarquer :

- des valeurs manquantes. Il nous faudra donc comprendre ces valeurs manquantes, et surtout les retraiter.
- des variables catégorielles (exemple pour la colonne JOB)
- des variables numériques, continues ou non (exemple pour la colonne CLAGE)

BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
1	1100	25860	39025	HomeImp	Other	10.5	0	0	94.36667	1	9	NA
1	1300	70053	68400	HomeImp	Other	7.0	0	2	121.83333	0	14	NA
1	1500	13500	16700	HomeImp	Other	4.0	0	0	149.46667	1	10	NA
1	1500	NA	NA			NA	NA	NA	NA	NA	NA	NA
0	1700	97800	112000	HomeImp	Office	3.0	0	0	93.33333	0	14	NA
1	1700	30548	40320	HomeImp	Other	9.0	0	0	101.46600	1	8	37.11361

Figure 1 : premières lignes du dataset

1). Analyse des variables catégorielles

Tout d'abord, commençons par analyser nos variables catégorielles. En premier lieu, voici Figure 2 la distribution de notre variable cible, c'est-à-dire la variable qui nous indique si le client a fait défaut ou non. On sait que le fait d'avoir un défaut (i.e. BAD = 1) est censé être un évènement rare, et donc en général plus dur à modéliser. Avoir donc une idée sur la distribution des modalités de notre variable cible nous donnera des pistes sur la méthode à considérer pour modéliser cet évènement.

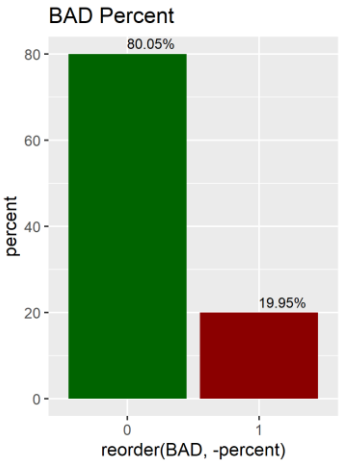


Figure 2 : Distribution de notre variable cible

On peut s'estimer heureux car dans notre dataset, l'évènement faire défaut semble être commun avec environ 20% d'occurrence. Cela facilitera donc la modélisation.

Ensuite, il nous faut analyser toutes les autres variables catégorielles. Dans le dataset, on a considéré 5 variables catégorielles, comme on peut le voir dans sur Figure 3. Pour les variables REASON et JOB, aucun doute, ce sont bien des variables catégorielles. Mais pour les 3 autres, on a des variables numériques non continues. Ainsi, nous avons décidé de les considérer comme des variables catégorielles, en tout cas pour la partie exploration des données, car elles imposent une notion de rang (des variables ordinales). Par exemple, la variable DELINQ désigne le nombre de d'impayés pour l'observation. On peut alors classer les individus selon le nombre d'impayés.

Passons maintenant à l'analyse de la Figure 3. Tout à gauche, on peut observer la distribution des modalités de chaque variable. On remarque dans un premier temps que les valeurs manquantes pour les variables REASON et JOB ne sont pas représentées par des NA

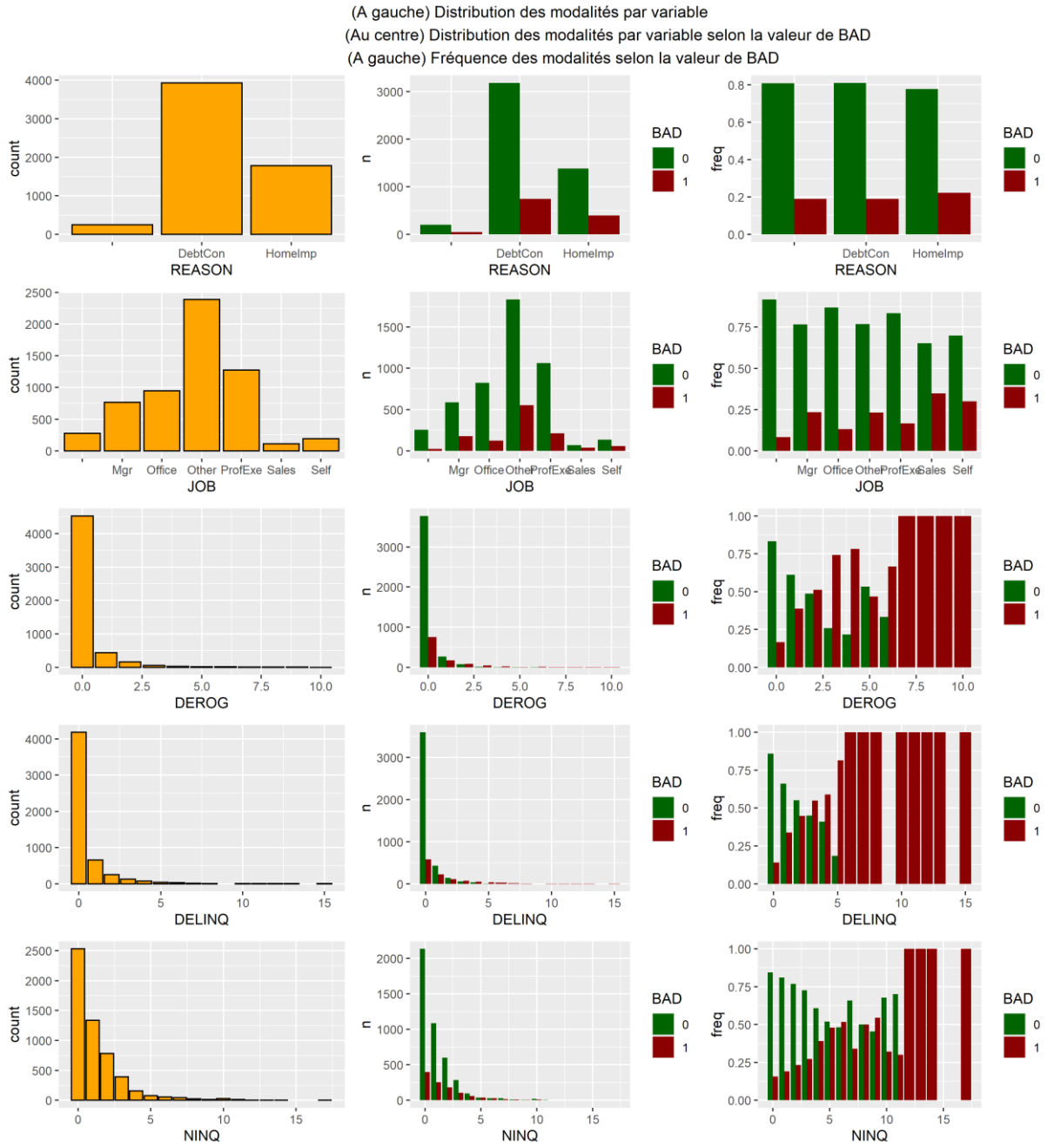


Figure 3 : Distribution des variables catégorielles

mais par une absence de valeur. Par conséquent, on a environ 3% de valeurs manquantes pour la variables REASON et à peu près 5% pour la variable JOB. Par rapport à la variable DEROG, on se rend compte que la plus part des observations n'ont pas eu de problème de paiement (DEROG = 0), et que le fait d'en avoir beaucoup semble inhabituel.

Au centre, on peut observer la distribution en fonction de la valeur de la variable cible. Il est intéressant de considérer des distributions condistionnelles car elles peuvent nous permettre de rendre compte visuellement des différences de comportement selon la valeur de la cible. Dans notre cas, il n'est pas pertinent de regarder ces distributions car la fréquence des modalités de la variable BAD ne

sont pas égales, et aucunes des caractéristiques ne semblent sortir du lot. C'est pourquoi nous avons la colonne de droite, qui représente la distribution par modalité en fonction de la valeur de la cible. On peut alors remarquer pour la variable DEROG :

- les valeurs extrêmes sont toutes associées à la modalité BAD = 1 (i.e. a fait défaut). Peut-être que les valeurs ne sont pas extrêmes et font juste partie des caractéristiques d'un individu qui fait défaut.
- plus le nombre de dérogation augmente, plus la fréquence de personne ayant fait défaut augmente. C'est intéressant, et pas anodin, car cela semble explicable. En effet, peut-être qu'avant de faire défaut, le client a plusieurs retards de paiement.

On remarque les même éléments pour les variables DELINQ et NINQ.

2). Analyse des variables numériques

Passons aux variables numériques, en adoptant la même méthode d'analyse. On a décidé dans un premier temps de nous concentrer sur la distribution de ces variables.

C'est ce que nous avons représenté en Annexe 1. Par exemple, on peut se rendre compte que la plus part des crédits se situent entre 2 000 et 25 000 d'unité monétaire (aucune devise n'est spécifié), que le plus gros de la distribution de la valeur des maisons se situe entre 30 000 et 250 000. Chose intéressante, certaines variables ont des distributions asymétrique (avec une queue vers la droite), c'est-à-dire que la moyenne est supérieure à la médiane. C'est ce qu'on observe pour la variable VALUE, LOAN, MORTDUE et YOJ. Probablement qu'effectuer une log transformation de ces variables pourrait-être intéressant pour l'analyse, car elle nous permettrait de transformer la distribution en gaussienne. Finalement, on observe un certain nombre de valeurs aberrantes en apparance, en observant les queues des distributions. On verra prochainement à l'aide de boxplots s'il s'agit réelement d'outliers.

Dans un deuxième temps, toujours sur l'Annexe 1, on a regardé la distribution selon la valeur de la variable BAD. Initialement, nous voulions savoir si il était possible de dissocier les distributions selon la valeur de la variable cible. Dans notre cas, aucune distribtuion n'est différenciable selon la valeur de BAD.

On a énoncé précédemment la possibilité que nos variables contiennent des valeurs aberrantes, et donc dans la Figure 4, nous avons tracé les boxplots de nos variables avec et sans la considération de notre variable cible.

Et en effet, toutes les variables ont bien des outliers. Mais, pour nous, aucune ne représente des valeurs aberrantes. En effet, nous n'avons pas le contexte de ces prêts, donc on ne peut pas dire si un LOAN a 80 000 est une aberration. Si cela avait été un prêt à la consommation pour de l'électroménager par exemple, alors probablement que 80 000 serait aberrant. Mais pour une voiture ce n'est pas le cas. Pareil pour la VALUE, car une maison peut très facilement valoir 750 000. Pour YOJ, travailler 40 ans pour la même entreprise n'est pas commun, mais cela ne peut être considéré comme absurde. Enfin pour finir, la variable DEBTINC est exprimée en pourcentage, donc un ratio de 200 (i.e. une dette 3 fois supérieure au revenu) n'est pas aberrant non plus.

Ce qui est maintenant pertinent, c'est de voir à quelle modalité de BAD ces valeurs particulièrement fortes appartiennent. C'est pourquoi nous pouvons observer un boxpot par variable selon la valeur de la cible. Au-delà de pouvoir attribuer un outlier à une modalité de la cible, elle nous permet d'observer plus facilement la répartition des valeurs qu'avec les histogrammes de l'Annexe 1.

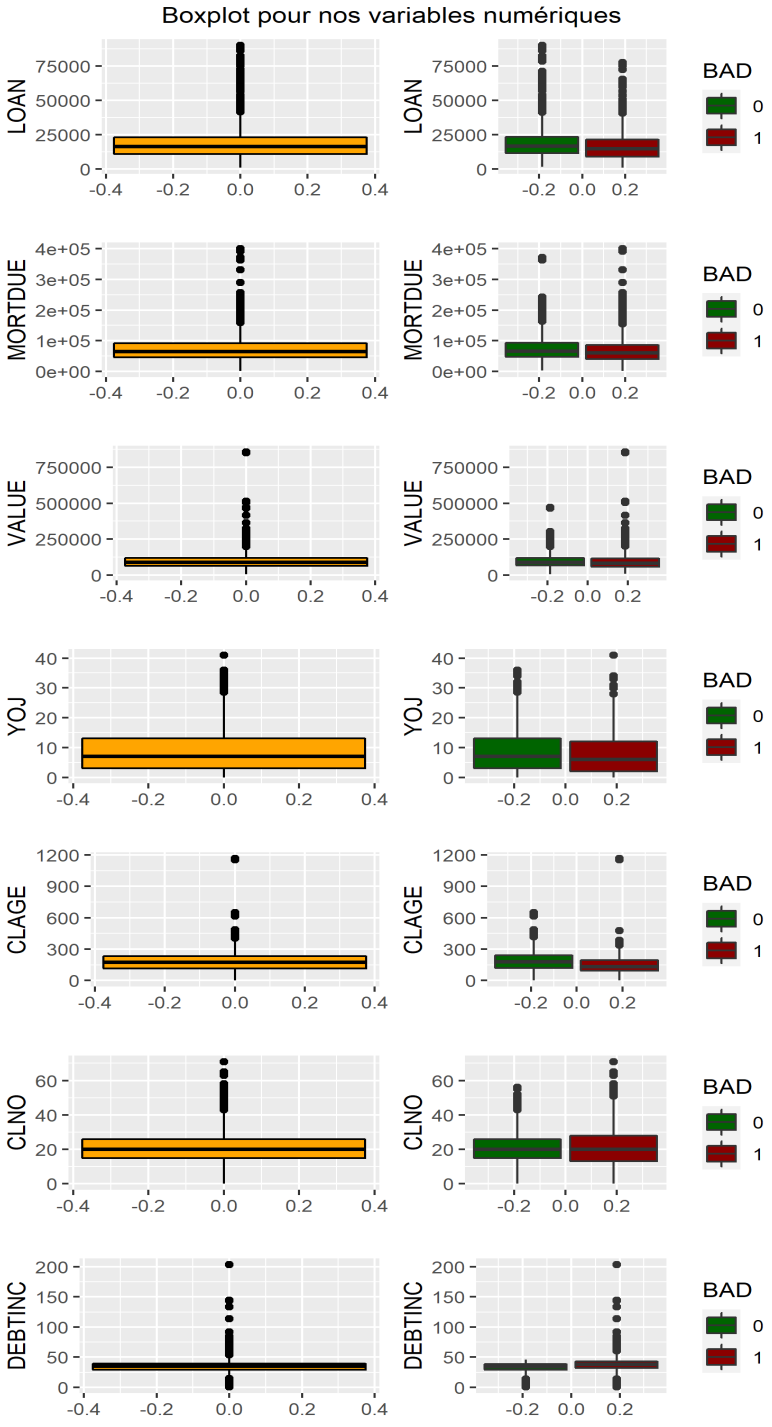


Figure 4 : Boxplot

De cette façon, on a l'occasion d'expliquer certaines observations qui peuvent sembler aberrantes comme étant une caractéristique d'un client qui fait défaut. Mais comme expliqué lors de l'étude des histogrammes (Annexe 1), il est difficile de différencier la distribution selon la modalité de BAD. Néanmoins, nous avons un contre-exemple avec la variable DEBTINC. Une personne qui fait défaut a

tendance à être surendettée, et c'est ce qu'on peut remarquer sur le boxplot : toutes les valeurs jugées aberrantes pour cette variable sont associées au fait de faire défaut. Cela pourrait être une des caractéristiques importantes pour prédire le défaut d'un client.

Ainsi, de cette analyse, on peut en ressortir 2 grands points :

1) Graphiquement, il est compliqué de déterminer une caractéristique associée au défaut ou au non-défaut. Ce qui pourrait nous être défavorable lors de l'utilisation de modèles linéaires simples.
2) Lors de notre analyse, nous nous sommes d'avantage reposés sur une connaissance métier que sur une interprétation graphique de nos valeurs.

II). Transformation et analyse des données

1). Traitement des valeurs manquantes

Pour être honnête, l'analyse que nous venons de vous présenter reste incomplète étant donné que nous n'avons pas pris en considération, pour les variables numériques du moins, les valeurs manquantes. A l'aide de la Figure 5, nous pouvons observer la disposition des valeurs manquantes, numériques uniquement :

Nous pouvons alors remarquer que la variable avec le plus de valeur manquante est DEBTINC, qui malheureusement était la seule variable avec laquelle nous avions pu associer certaines de ses valeurs

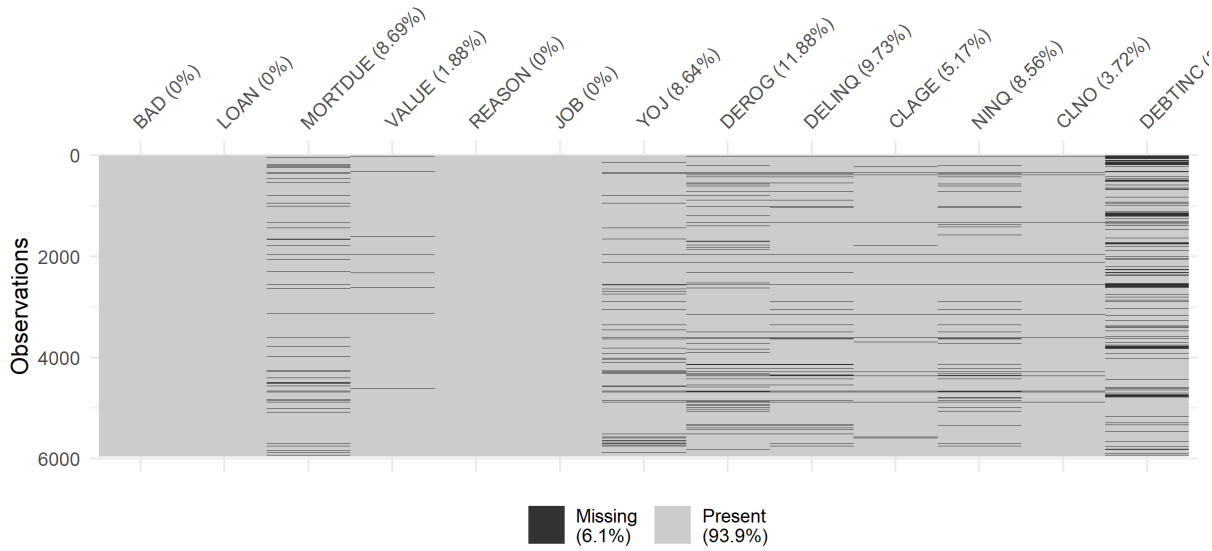


Figure 5 : Position des valeurs manquantes par variable numérique

à la variable cible. Or, le raisonnement est peut-être biaisé par le manque de valeur. Deuxièmement, on peut observer des « patterns » de valeurs manquantes. C'est-à-dire que visuellement, on a des plusieurs traits qui traverse toutes les variables numériques. Cela semble nous indiquer que nous sommes en présence pour la plupart des valeurs manquantes à des MAR (missing at random) et des MNAR (missing not at random). Traiter ces valeurs manquantes sera donc plus compliqué, surtout si on souhaite éviter de générer du biais pour notre future analyse.

Dans un premier temps, on a cherché à savoir si les valeurs manquantes par colonnes étaient plus associées aux observations qui ont fait défaut ou non. Par conséquent, nous avons tracé la disposition des valeurs manquantes par modalité de BAD (Annexe 2). Premier fait marquant, nous avons 65% de valeurs manquantes pour la variable DEBTINC quand BAD = 1, contre 10% quand BAD = 0, alors que sans considérer la valeur de BAD, nous avions 23% de valeurs manquantes. Ainsi, la plupart des NA de cette colonne sont liées au fait d'avoir fait défaut. Probablement que les personnes qui ont fait défaut n'ont pas fourni leurs informations à propos de leur endettement ; mais, quand on supprime les lignes qui possèdent plus de 37%[1] de valeurs manquantes, nous n'avons plus ce déséquilibre de NA pour la colonne DEBTINC (Annexe 3), sans pour autant détraquer les proportions[2] pour les modalités de la variable cible.

Il nous faut donc traiter les valeurs manquantes. Après avoir supprimé les lignes qui possédaient plus de 37% de NA, nous avons :

- Remplacé par la moyenne pour les variables MORTDUE, VALUE, YOJ, CLAGE et DEBTINC (variables numériques continues)
- Remplacé par la médiane pour les variables DEROG, DELINQ, NINQ et CLNO (variables numériques non continues)
- Remplacé par la valeur « missing_value » pour les variables JOB et REASON (variables catégorielles)

2). Transformation des données

Pour finir on a cherché à comparer l'effet d'une log transformation pour certaines variables sur la forme de leur distribution (Annexe 4) : Le fait d'appliquer un log nous permet de fit une gaussienne. On a donc tracé un QQPlot avant et après la log transformation, qui permet de comparer la position de notre distribution par rapport à différents quantiles de la distribution gaussienne, il permet donc de comparer la similarité entre 2 distributions (ici une loi normale et la distribution de notre variable). Avant la transformation, on observe un QQPlot avec une courbe « convexe », ce qui représente bien la distribution asymétrique avec une queue vers la droite. Il y a énormément à dire sur ces graphiques, mais on va se concentrer sur le graphique après la log transformation. Nous remarquons alors que pour la variable YOJ[3], cette log transformation nous éloigne d'une distribution gaussienne, avec beaucoup de valeurs nulles. On n'appliquera donc pas cette transformation sur cette variable (en plus de la distribution de sa transformation qui s'éloigne d'une gaussienne). Pour toutes les autres variables, les queues s'éloignent un peu de la loi normale, mais l'ensemble fit bien avec la gaussienne.

C'est pourquoi cette log transformation a été appliqué aux variables LOAN, VALUE et MORTDUE, et pas YOJ pour les raisons qui ont été cités précédemment.

[1] Ce seuil correspond à la valeur inférieure du ratio 5/13, choisi pour éviter d'avoir des valeurs manquantes simultanément dans les colonnes DEROG, DELINQ, CLAGE, NINQ et CLNO.
[2] ~80% pour BAD = 0
[3] La variable YOJ possède des valeurs nulles, donc nous avons préféré faire une log + 1 transformation.

Maintenant que nous avons traité les valeurs manquantes, la suite du projet est de scaler nos données pour éviter les problèmes d'échelles, et d'encoder nos variables catégorielles.

Les deux variables facteurs de notre jeu de données, à savoir REASON et JOB, sont des variables catégorielles disposant de peu de valeurs uniques. En effet, la variable REASON contient les modalités suivantes :

- HomeImpt
- DebtCon
- missing_value

Et la variable JOB :

- Office
- Sales
- Mgr
- ProfExe
- Other
- missing_value

Ainsi, ces deux variables sont éligibles à deux techniques simples de préprocessing permettant de les rendre compréhensible par un modèle linéaire : la labellisation et le OneHotEncoding (aussi connu sous le nom de dummying). Mais nous avons de la chance car le merveilleux langage R permet de typer les variables catégorielles comme factor, ce qui nous offre la possibilité de passé outre l'encodage.

Après l'encodage, nous avons appliqué du standard scaling (retrait de la moyenne, puis division du tout pas l'écart-type) à nos variables numériques dans le but d'obtenir une meilleure interprétabilité de nos variables, et surtout des paramètres des régressions linéaires que nous allons appliquer par la suite. En effet, si les variables disponibles sont à la même échelle, alors les coefficients de la régression linéaire le sont aussi, et peuvent être comparés l'un à l'autre.

En Annexe 5, nous pouvons constater que le standard scaling conserve la distribution de nos variables, et que seule l'échelle est impactée par cette transformation.

3). Analyse des liens entre nos variables

Nous avons maintenant des données propres, qui vont nous servir à modéliser la probabilité de défaut. Mais avant, on va chercher à observer les liens entre nos variables, pour avoir à la fois une direction pour la modélisation et surtout des pistes d'analyse. L'idée de ce point et de pouvoir analyser les variables entres elles, la plupart du temps face à la variable cible mais pas uniquement. Cette analyse n'a pas pour but de faire de la sélection variable, mais de nous donner les outils pour comprendre pourquoi telle ou telle variable n'a pas été sélectionné par nos méthodes de feature selection.

Le premier lien que nous avons décidé d'étudier était celui entre toutes nos variables catégorielles et la variable cible. C'est pourquoi nous nous sommes orientés vers un test du Chi^2 d'indépendance, qui nous permet de tester l'absence de lien entre 2 variables, dans notre exemple JOB-BAD et REASON-BAD. On va donc tester l'hypothèse suivante H0 : « X et Y sont indépendants » (i.e. les valeurs de Y dépendent de X).

Dans notre cas, nous avons obtenu une p-value de 0.0125 pour le test de l'hypothèse « REASON et BAD sont indépendants », et une p-value de 0 pour le test « JOB et BAD sont indépendants ». Par conséquent, pour un seuil de confiance de 5%, nous sommes obligés de refuser l'hypothèse H0, car la p-value est inférieure à notre seuil. On peut alors en conclure que nos 2 variables explicatives REASON et JOB ne sont pas inintéressantes à utiliser car les valeurs que peut prendre la variable BAD dépend des valeurs que prennent REASON et JOB (avec un risque de se tromper de 5%).

Le deuxième lien que nous avons décidé d'étudier est simplement la corrélation entre toutes les variables numériques (on exclut donc la variable cible). C'est ce que nous pouvons observer à travers la Figure 6 ci-dessous. Nous avons choisi de calculer aussi la corrélation de Spearman car comme l'analyse exploratoire nous l'avait montrée, nous avions des outliers (même si d'un point de vu

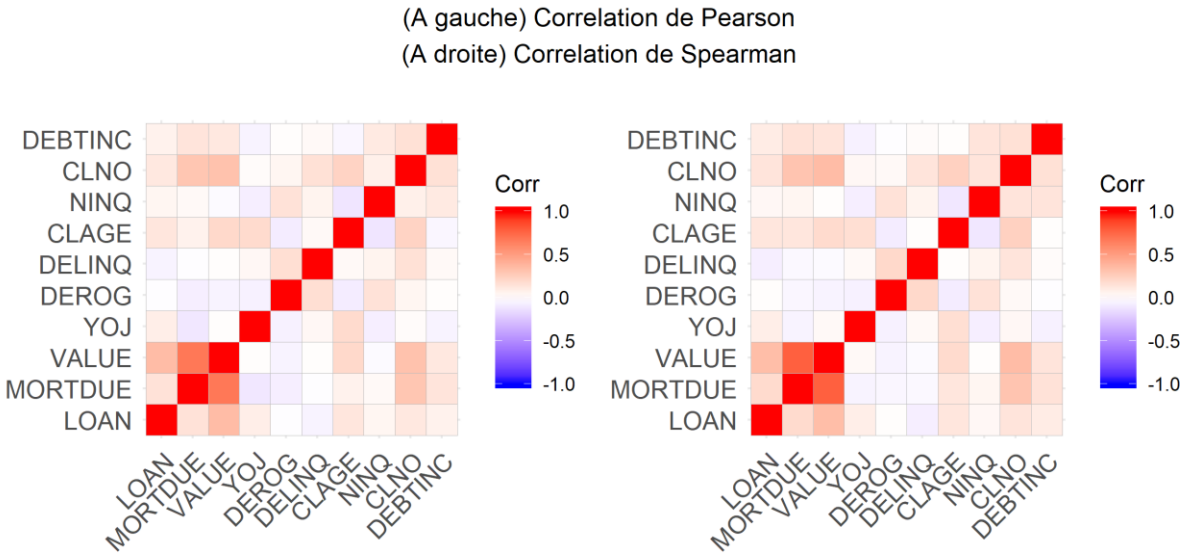


Figure 6 : Corrélation plot

« Business » ce n'étaient pas des observations extrêmes) et par conséquent, pour éviter d'avoir un biais dans notre analyse nous avons décidé de compléter avec une corrélation du rang ; étant donné que nous avions décidé de garder ces observations dans notre jeu de données. A première vue, nous n'avons pas de différence de couleur entres les cellules de nos 2 graphique de corrélation, ce qui est une bonne chose étant donné qu'on va pouvoir se concentrer sur un des graphiques.
On se rend compte que la variable VALUE fortement corrélé à la variable MORTDUE, ce qui est logique, étant donné VALUE exprime la valeur de la maison, et que MORTDUE représente la valeur du prêt hypothécaire. Il faudra faire attention plus tard à vérifier qu'il n'y a pas de colinéarité entre ces variables. On a aussi une corrélation élevée entre LOAN et VALUE, mais mis à part ces deux cas, rien d'anormal. Pour compléter cette analyse, voici un Variance Inflation Factor analysis (VIF), qui va nous permettre d'observer les problèmes de colinéarité et de multi-colinéarité. On cherche donc à savoir si une variable peut s'exprimer en fonction des autres, et peut alors biaiser notre modèle et notre analyse. Voici Figure 7 ci-dessous les résultats de notre VIF :

```
> print(vif(reg))
             GVIF Df GVIF^(1/(2*Df))
LOAN     1.345468  1        1.159943
MORTDUE  1.917474  1        1.384729
VALUE    2.215180  1        1.488348
REASON   1.209107  2        1.048615
JOB      1.363214  6        1.026157
YOJ      1.089324  1        1.043707
DEROG    1.033791  1        1.016755
DELINQ   1.116669  1        1.056726
CLAGE    1.180547  1        1.086530
NINQ     1.071085  1        1.034932
CLNO     1.305893  1        1.142757
DEBTINC  1.077688  1        1.038117
```

Figure 7 : Sortie R du VIF

Le VIF permet « d'estimer de combien la variance d'un coefficient est augmentée en raison d'une relation linéaire avec d'autres prédicteurs »[4].

On va regarder la sortie GVIF. Si la valeur est supérieure à 2.5 ou 5 (cela dépend des statisticiens), alors on a de la multi-colinéarité sévère dans nos variables. Dans notre cas aucune valeurs de dépasse les 2.2 pour la variable VALUE. Cela veut dire que la variance de coefficient est de près 2 fois plus forte que la variance que l'on aurait pu observer si ce facteur était absolument décorrélé de toutes les autres variables. Cela explique alors la forte corrélation que nous avons pu observer entre VALUE et MORTDUE. Mais son VIF est inférieur à 2.5, donc nous allons tout de même la conserver et nous verrons si cette variable sera sélectionné ou pas par le processus de feature selection.

Nous pouvons aussi remarquer que toutes les autres variables ont un VIF qui avoisine 1, ce qui caractérise des variables, peu corrélées entre elles, ce que nous avions pu observer Figure 6.

Jusque-là, nous avons pu étudier le lien entre nos variables catégoriques et la variable cible, et le lien entre toutes les variables numériques. Le problème pour le moment, c'est que nous n'avons pas encore considéré l'étude du lien entre la variable cible et les variables numériques. Vu que nous devons étudier le lien entre une variable catégorique et une variable numérique (pairwise analysis), il nous semble pertinent d'entamer une ANOVA, c'est-à-dire une analyse entre une variable catégorielle et une variable numérique. Pour le coup, la variable catégorielle est représentée par notre cible BAD. Pour rappel, elle possède 2 modalités : 1 si l'observation a fait défaut, 0 sinon. Par conséquent, dans ce cas précis, notre ANOVA revient à faire un test de Student de différence de moyenne entre les 2 modalités de BAD. On va donc tester l'hypothèse H0 : « Les moyennes de X par modalité de Y sont égales » (i.e. la moyenne de la variable numérique considéré et la même peu importe que le client ait fait défaut ou non)

Nous avons représenté les résultats sur la Figure 8 ci-dessous. Voici les clefs pour comprendre ce graphique :

- Les boxplots représentent la dispersion des valeurs de la variables numérique considéré par modalité de BAD.
- Le titre de chaque graphique représente le résultat du test de Fisher. Ainsi, « Toutes les moyennes de LOAN sont différentes » signifie que nous avons rejeté l'hypothèse H0, c'est-à-dire que les moyennes de LOAN par modalité de BAD sont statistiquement différentes. Cela correspond alors a une p-value inférieur à notre seuil de 5%.

[4] http://larmarange.github.io/analyse-R/multicolinearite.html

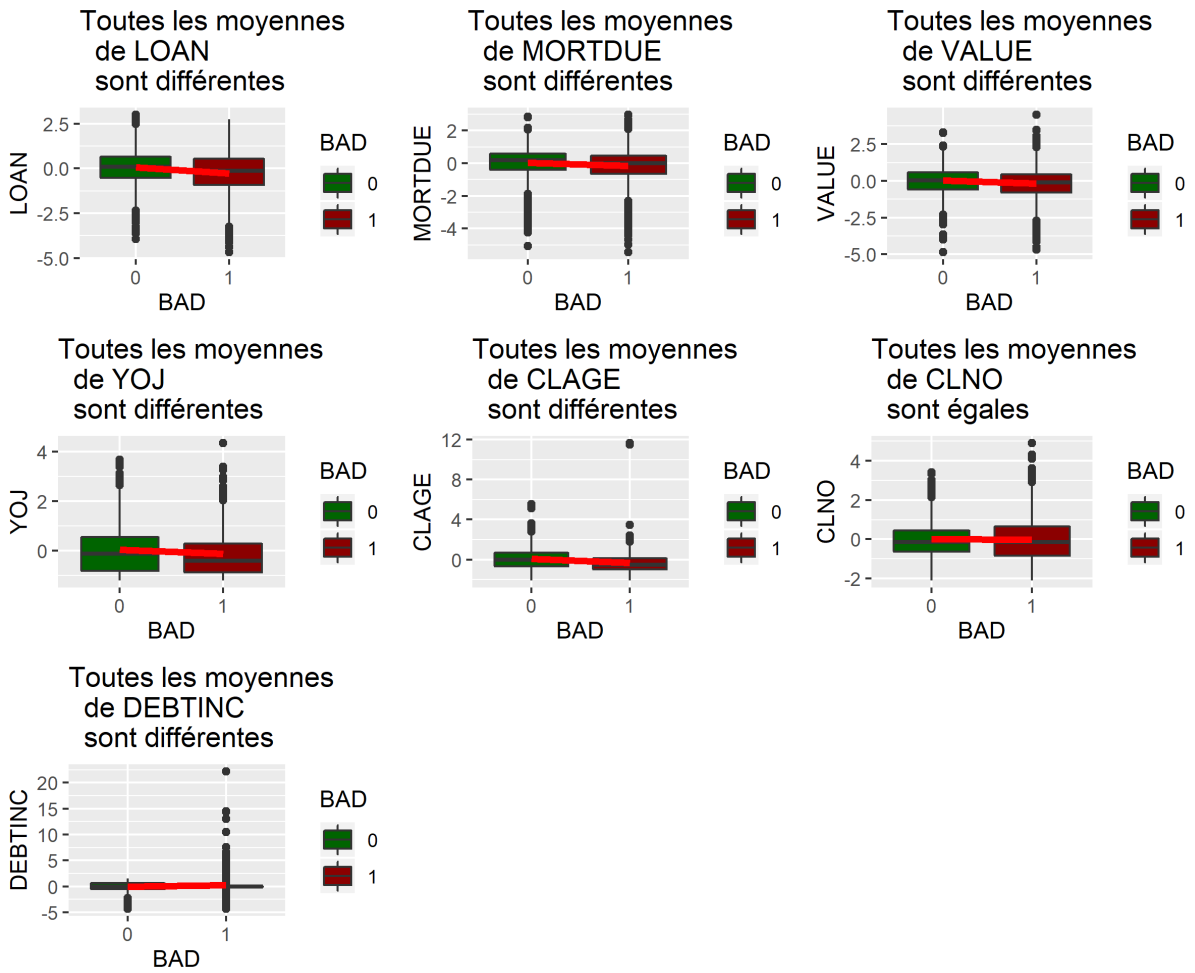


Figure 8 : ANOVA avec les boxplots par variables numériques

- La ligne rouge relie les moyennes que l'on compare dans le test. Si le segment est horizontal et que l'on observe par modalité des dispersions similaire, alors il est fortement probable que le test nous fasse accepter l'hypothèse H0 (i.e. les moyennes par modalité sont égales)

Pour toutes les variables sauf une, on rejette l'hypothèse H0. Seule la variable CLNO, c'est-à-dire le nombre de crédit, ne nous permet pas de refuser H0. C'est-à-dire que peu importe qu'un client ait fait défaut ou non, en moyenne ils ont le même nombre de crédit.

III). Construction du meilleur modèle

Enfin nous sommes arrivés à la modélisation du risque de défaut. Le but est de challenger 4 modèles que nous allons construire à l'aide de différentes méthodes de sélection de variable ; et représenter dans un logit, et dans un probit ces meilleures variables pour comparer notre ensemble de modèle.

1). Sélection de variable

Nous commençons tout d'abord les méthodes de sélection de variable dans le but d'utiliser uniquement les variables les plus pertinentes dans l'explication de notre variable cible (BAD). Cela nous permettra de diminuer la variance du modèle, et donc notre modèle sera plus précis quand appliqué à des données réelles.

Etant donné que le nombre de variables disponibles dans notre jeu de données est relativement faible (13 variables, 20 si l'on compte les différentes modalités des facteurs), une recherche pas à pas ne sera pas nécessaire. Nous allons donc utiliser des recherches exhaustives, qui nous assureront d'obtenir les meilleures combinaisons de variables.

Nous avons utilisé les critères AIC (Critère d'Information d'Akaike) et BIC (Critère d'Information Bayésien) pour la sélection des variables, et avons comparé les résultats.

Pour le critère AIC, les variables suivantes ont été sélectionnées : LOAN, MORTDUE, REASON, JOB, YOJ, DEROG, DELINQ, CLAGE, NINQ, CLNO et DEBTINC. Autrement dit, la variable VALUE n'a pas été retenue, ce qui semble cohérent avec le fait qu'elle soit fortement corrélée avec MORTDUE. L'AIC obtenu par bestglm en utilisant cette meilleure combinaison de variable est de 4327.09.

Pour le critère BIC, les variables suivantes ont été sélectionnées : LOAN, MORTDUE, JOB, DEROG, DELINQ, CLAGE, NINQ et DEBTINC.

Nous pouvons ainsi observer que le critère BIC a été plus sévère et a retenu moins de variables (8 variables) que le critère AIC (11 variables). Cela semble logique car le BIC est reconnu pour pénaliser plus fortement le nombre de paramètres à estimer que l'AIC.

Si on met en exergue l'analyse de lien avec ces résultats, on se rend compte que la variable REASON a disparu, et comme par hasard, c'était la variable catégorielle qui avait la p-value la plus forte lors de notre test d'indépendance avec la variable cible ; VALUE était une variable très corrélée à MORTDUE ; CLNO était la seule variable qui ne rejetait pas l'hypothèse d'égalité des moyennes pour les modalités de BAD. Une autre variable n'a pas été prise en compte, mais nous ne sommes pas en mesure de comprendre pourquoi : YOJ. Néanmoins, c'était la seule variable dont la distribution de sa log transformation différait très fortement de celle de la loi normale.

2). Développement des modèles

Nous nous situons maintenant à l'étape finale d'implémentation du modèle. En effet, après avoir analysé, traité et formaté notre jeu de données, nous pouvons enfin le modéliser. Pour ce faire, nous nous appuierons sur deux jeux de données, correspondant aux deux méthodes de sélection de variables. Ces dataset ont été développés dans la partie précédente.

Et donc sur chacun de ces jeux de données, nous réaliserons un modèle probit et un modèle logit. Ces deux modèles font partis de la famille des modèles linéaires généralisés : ils suivent tous deux une loi binomiale, mais se différencient dans leur fonction de lien, en effet, celle du logit est de la forme : $g(u) = \log\left(\frac{u}{1-u}\right)$, tandis que le probit aura une forme telle que : $g(u) = \Phi^{-1}(u)$.

Pour finir, une dernière étape est nécessaire afin de savoir si la modélisation est de bonne qualité ou alors que celle-ci ne prédit pas ou peu. Nous diviserons nos deux jeux de données en deux jeux de test et deux jeux d'entrainement, le jeu d'entraînement permettra au modèle d'apprendre en faisant la relation entre les variables explicatives et la variable cible (BAD). Le jeu de test permettra quant à lui d'utiliser notre modèle, préalablement entraîné, sur le jeu de test afin d'en sortir une probabilité. Cette probabilité obtenue se verra appliquer un seuil qui répartira les probas continues en 2 modalités qualitatives, 1 pour défaut ou 0 pour non-défaut. Cette classification ainsi obtenue sera comparée à la vraie cible de la table de test. On pourra ainsi, par des outils tels que la courbe ROC, déterminer si le modèle a bien prédit ou à contrario si celui-ci relève pus de l'aléatoire (exemple AUC = 0.5).

Par conséquent, nous avons devoir développer 4 modèles (2 datasets pour 2 modèles différents). La première étape sera de choisir le type de modèles le plus pertinent. Ainsi, pour commencer, voici Figure 9 une de nos sorties R quand au training du modèle logit sur le dataset AIC, c'est-à-dire les variables qui ont été sélectionnées sous le critère AIC dans notre recherche exhaustive.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2485  -0.5752  -0.3976  -0.2531   3.9343

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.46702    0.13867 -17.791  < 2e-16 ***
LOAN               -0.26311    0.04901  -5.368 7.97e-08 ***
MORTDUE            -0.17345    0.05067  -3.423 0.000619 ***
REASONHomeImp       0.12217    0.10553   1.158 0.246982
REASONmissing_value 0.33764    0.28599   1.181 0.237754
JOBmissing_value   -0.96725    0.41169  -2.349 0.018802 *
JOBOffice          -0.62933    0.18377  -3.425 0.000616 ***
JOBOther            0.10740    0.13939   0.771 0.440996
JOBProfExe          0.13179    0.15958   0.826 0.408886
JOBSales            1.02452    0.30979   3.307 0.000942 ***
JOBSelf             0.73435    0.25823   2.844 0.004458 **
YOJ                -0.10428    0.05105  -2.043 0.041058 *
DEROG               0.59238    0.05983   9.900  < 2e-16 ***
DELINQ              0.78354    0.04605  17.015  < 2e-16 ***
CLAGE              -0.48224    0.05565  -8.665  < 2e-16 ***
NINQ                0.18043    0.02439   7.399 1.37e-13 ***
CLNO               -0.08834    0.05165  -1.710 0.087236 .
DEBTINC             0.42593    0.05876   7.249 4.20e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4208.7  on 4262  degrees of freedom
Residual deviance: 3188.2  on 4245  degrees of freedom
AIC: 3224.2

Number of Fisher Scoring iterations: 5
```

Figure 9 : Sortie logit sur dataset AIC

On peut y interpréter plusieurs éléments :

- La déviance des résidus représente la distribution des résidus dans notre modèle. En règle générale, un résidu représente une erreur, soit le décalage entre les données réelles et le modèle créé.
- Les coefficients représentent le coefficient calculé pour chacune des features. Celui-ci est déterminé avec l'estimateur du maximum de vraisemblance. Le but sera de maximiser la log-vraisemblance de notre modèle. C'est ce que fait la commande GLM pour déterminer chacun des coefficients. Le but de cette optimisation est de déterminer les paramètres pour lesquels le gradient s'annule. On peut observer que les paramètres ont été bien calculés dans notre modèle, étant donné qu'il a convergé. Ces coefficients peuvent être interprétés de différentes façons, d'une part, on peut interpréter l'influence d'un coefficient sur le modèle après transformation par sa fonction de lien. Aussi, on peut observer l'importance de chacun de ces coefficients dans notre modèle. Cette importance ou significativité des coefficients est un test réalisé afin de savoir si le coefficient sert à notre modèle. En effet, un coefficient significativement différent de 0 apporte à lui seul de l'information, tandis qu'un coefficient non significativement différent de 0 a des chances de n'apporter aucune information et de rajouter du bruit. Ce test est réalisé sur le positionnement de la t-value par rapport à un seuil,

relevant d'un quantile de distribution de loi, qui accepte ou rejette l'hypothèse de significativité. Cette t-value est calculée par le rapport du coefficient estimé sur son écart-type.

- La « Null deviance » est l'ajustement du modèle avec l'intercept pour un degré de liberté fixé à n-1. Elle peut être interprétée comme une valeur de chi-2. En effet, elle nous permet de mesurer l'écart entre la distribution obtenue par l'utilisation unique de « l'intercept », vis à vis d'un modèle complet.
- La métrique de déviance est obtenue en différenciant la vraisemblance d'un modèle saturé (considéré comme parfait) à celle de notre modèle actuel. Celle-ci sera toujours positive étant donné que plus la vraisemblance est élevée meilleure elle est, alors le modèle parfait aura toujours la plus grande vraisemblance, de plus, plus elle sera faible mieux l'ajustement sera. La métrique est donnée dans la sortie GLM par « Residual Deviance ».

On peut voir ici que la plupart de nos coefficients sont significatifs, au seuil de 5%. Cependant, il est à noter que 5 de nos coefficients ne le sont pas. Ce qui est intéressant à voir, c'est que 4 d'entre elle sont des variables dummisées. Cela provient peut-être du fait qu'une des modalités n'est pas intéressante.

Chose plus étonnante, la variable CLNO n'est que très peu significative : si le seuil d'acceptation de 5% était choisi, elle ne le serait pas statistiquement différente 0. Cette variable n'a par conséquent que peu d'importance dans notre modèle. Ce qui, pour rappel était notre conclusion par rapport au test ANOVA effectué précédemment. On peut donc affirmer que le nombre de crédit d'un client n'a pas d'impact sur sa solvabilité.

De plus, nous pouvons observer un « Residual Deviance » de 3188 ; ce qui semble de prime abord assez élevé, cependant cela n'est que relatif. En effet, la métrique étant liée à la fonction de vraisemblance et donc intrinsèquement à la taille de notre échantillon, nous devons réaliser un test pour savoir si le modèle est adéquat.

On pose H0 : « le modèle est adéquat » versus H1 : « il ne l'est pas ». A partir d'un test du Chi-2, la p-value est très proche de 1, par conséquent nous acceptons l'hypothèse H0 (Annexe 6).

Tous ce qu'on vient de faire, c'était uniquement pour un des quatre modèles. On a pu alors décrire et comprendre les sorties pour tous les modèles. Le but n'est pas d'étudier tous les coefficients pour tous les modèles, mais de pouvoir challenge les modèles entre eux. On cherche alors à sélectionner 1 ou 2 modèles, les plus pertinents, ceux avec le BIC le plus faible. C'est pourquoi la Figure 10 représente les valeurs des AIC et des BIC pour chacun des modèles que nous avons développé.

Feature_selected	Model	AIC	BIC	test_dev
AIC	Logit	3224.209	3338.648	1
AIC	Probit	3231.605	3346.044	1
BIC	Logit	3226.138	3315.146	1
BIC	Probit	3237.580	3326.588	1

Figure 10 : AIC et BIC pour tous les modèles

On va donc sélectionner le modèle logit, avec les données sélectionnée grâce au BIC. En effet, on se rend compte que c'est le modèle qui minimise le BIC, et qui par la même occasion, a un des AIC les plus faible. Nous préférons nous fier au BIC car nous préférons pénaliser le nombre de variable du modèle. C'est pourquoi le reste de l'analyse se portera sur ce modèle.

Dans un premier temps, nous avons tracé sa courbe ROC, qui représente la valeur spécificité et de la sensibilité quand on fait bouger le seuil de catégorisation. En effet, notre modèle produit une probabilité, et c'est à nous de choisir un seuil qui permet au mieux de catégoriser nos observations. On fait varier le seuil en fonction de certaines métriques comme la sensitivité ou la précision. Le problème, c'est qu'étudier un modèle en fonction du seuil choisi dans l'absolue est une erreur car il ne nous permet de comparer ce modèle aux autres. La courbe ROC nous permet ainsi de calculer 2 métriques en fonction de la valeur du seuil (que l'on fait varier), d'avoir une métrique universelle qui nous aide à comparer les modèles. Dans notre cas, on regarde l'aire sous la courbe ROC, c'est l'AUC. C'est ce que nous pouvons voir sur la Figure 11 :

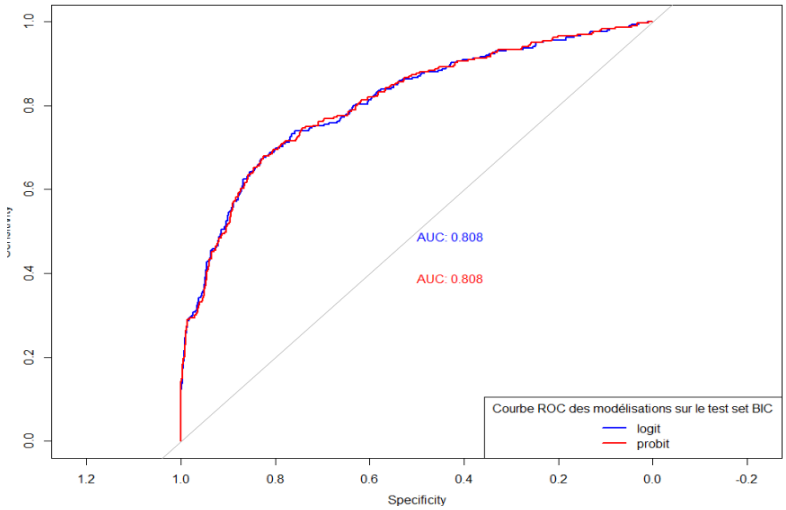


Figure 11 : Courbe ROC pour le dataset BIC

Voici les courbes ROC pour les modèles calculés pour le dataset BIC (recherche exhaustive avec BIC). Nous allons nous concentrer sur la courbe bleue qui est le modèle choisi précédemment, même si on remarque que les 2 courbes se superposent. Pour un classifier, nous avons une bonne AUC (à voir avec les autres membres de la promotion). Probablement que les transformations que nous avons effectuées sur le dataset étaient bénéfiques, autant que la méthode de sélection des variables.

Pour comparer, avec le dataset AIC, nous avons une AUC de 0.811 pour les 2 modèles. On a donc une métrique plus faible pour le modèle que nous avons choisi grâce au critère BIC. Il est probable que le fait d'avoir plus de variable dans le modèle AIC améliore l'AUC, mais cela risque d'augmenter le bruit par la même occasion lors des prédictions. On reste donc sur notre position.

Maintenant que nous avons idée de la performance du modèle, il nous faut déterminer le seuil. Pour cela, nous avons décidé de choisir le seuil qui maximise le F1-score, qui désigne la moyenne harmonique entre la sensitivité et la précision. Nous remarquons sur la Figure 12 que le seuil qui maximise le F1 est 0.67. C'est-à-dire que selon les caractéristiques d'une observation, une probabilité ressortit par le modèle supérieur à 0.67 sera considéré comme un défaut.

A partir de ce seuil, on peut alors tester la pertinence de ce seuil et du modèle. On va alors tester notre modèle sur l'échantillon de test. Voici donc Figure 13, la matrice de confusion, pour le modèle logit sur le dataset BIC, pour le seuil de 0.67. On peut alors calculer une sensibilité de 84% et une précision de 28%. Etant donné la proportion d'observations ayant fait défaut dans le dataset (environ 20%), notre précision nous indique que le modèle prédit mieux que le hasard (la probabilité d'aller chercher

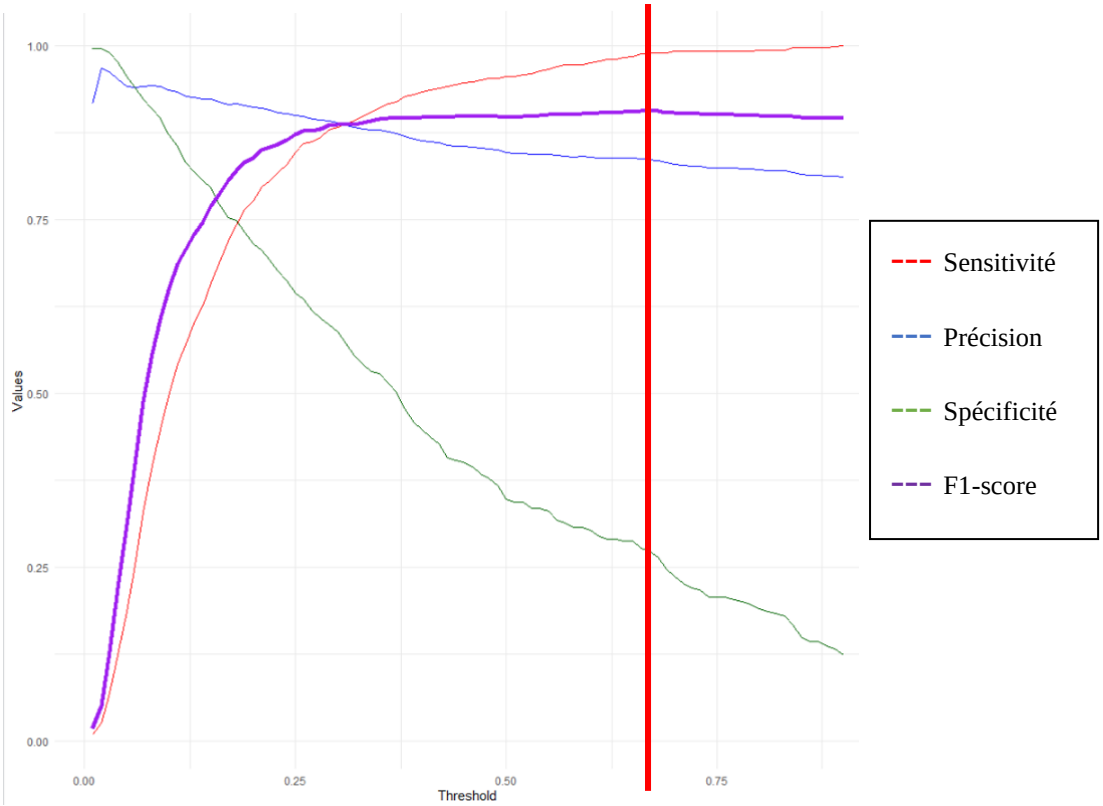


Figure 12 : Valeurs des métriques en fonction du seuil

	Prédiction		
Réalité		Ne fait pas défaut	Fait défaut
	Ne fais pas défaut	1107	213
	Fait défaut	16	86

Figure 13 : Matrice de confusion pour le modèle logit et le dataset BIC

au hasard un client dans notre dataset et de tomber sur un client en défaut est de 1 chance sur 5). Pour l'autre modèle, les métriques ne sont pas significativement différentes de celles-ci. Ce qui nous rassure dans le choix du modèle étant donné que l'autre modèle n'améliore pas la précision si la sensibilité avec plus de variable. Principe de parcimonie oblige, nous restons camper sur notre choix du modèle logit avec le dataset BIC.

IV). Conclusion :

Dans ce projet de scoring bancaire, nous avons mis l'accent sur l'analyse des données. Nous avons mis l'accent et pas en vain. En effet, à de nombreuses reprises, nous avons expliqué le comportement d'un modèle, d'une méthode grâce à nos analyses, soit exploratoire des données, soit de lien. Au sein de n'importe quelle entreprise, à partir du moment ou nous avons des gens du métier en face de nous, nous nous devons de simplifier les méthodes, et de nous faire comprendre. C'est pourquoi de notre côté, près des 3/5 du projet est orienté sur la compréhension des données.

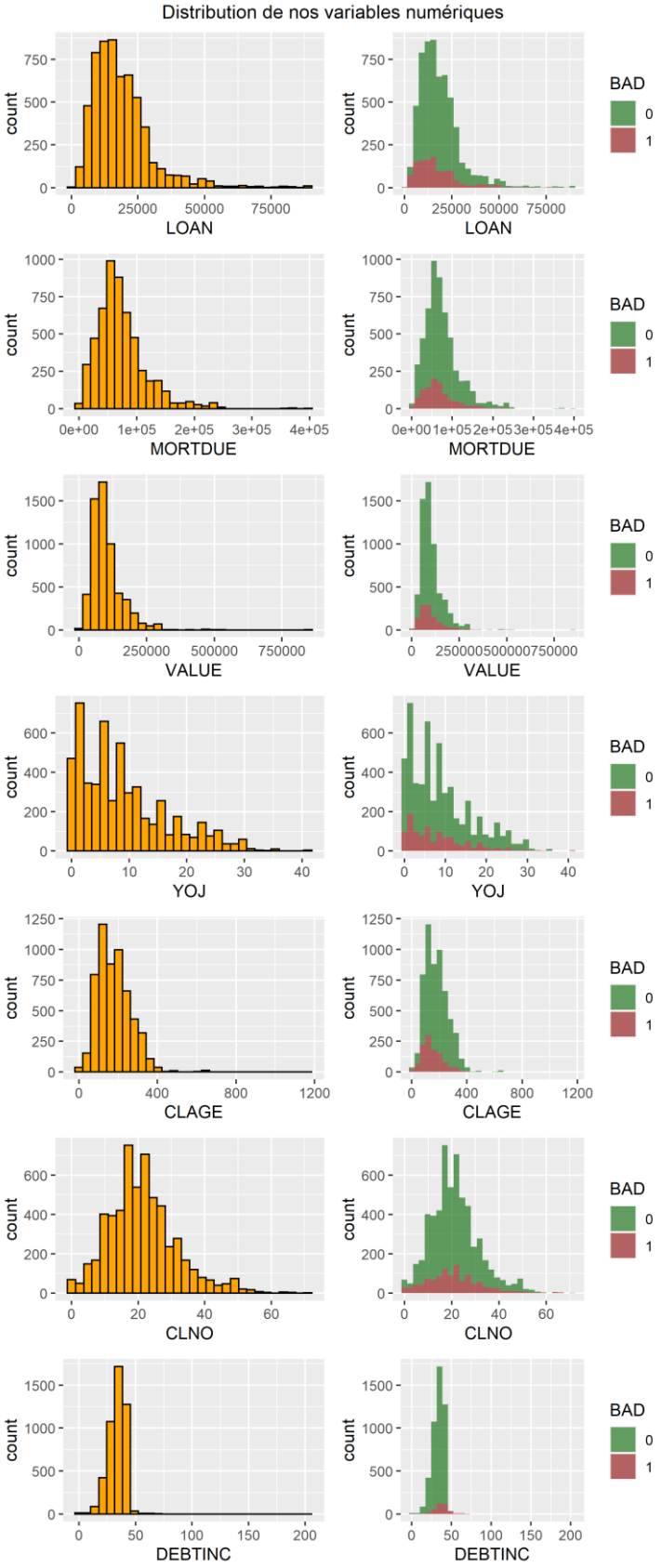
Mais ce n'est pas pour autant qu'il faut négliger la modélisation. Au contraire, se donner autant de mal sur l'analyse des données pour à la fin mal définir le modèle est une perte de temps. Dans le cadre de la régulation des établissements bancaires, nous nous sommes attardés sur les modèles classiques de prédiction binaire, qui sont facilement auditables par le régulateur, mais aussi plus facilement compréhensible par le métier. De plus, expliquer au client que son prêt lui a été refusé à cause de la classification faite par un XGBoost serait difficile.

Cependant, il y a quelques années est apparu la Shap Value ou encore le LIME, qui sont des outils nous permettant de comprendre un XGBoost, et tous les autres modèles dit de « boîte noire ». Mais leur démocratisation se fait lente, et leurs entrées dans le secteur bancaire ne se fera pas avant quelques dizaines d'années.

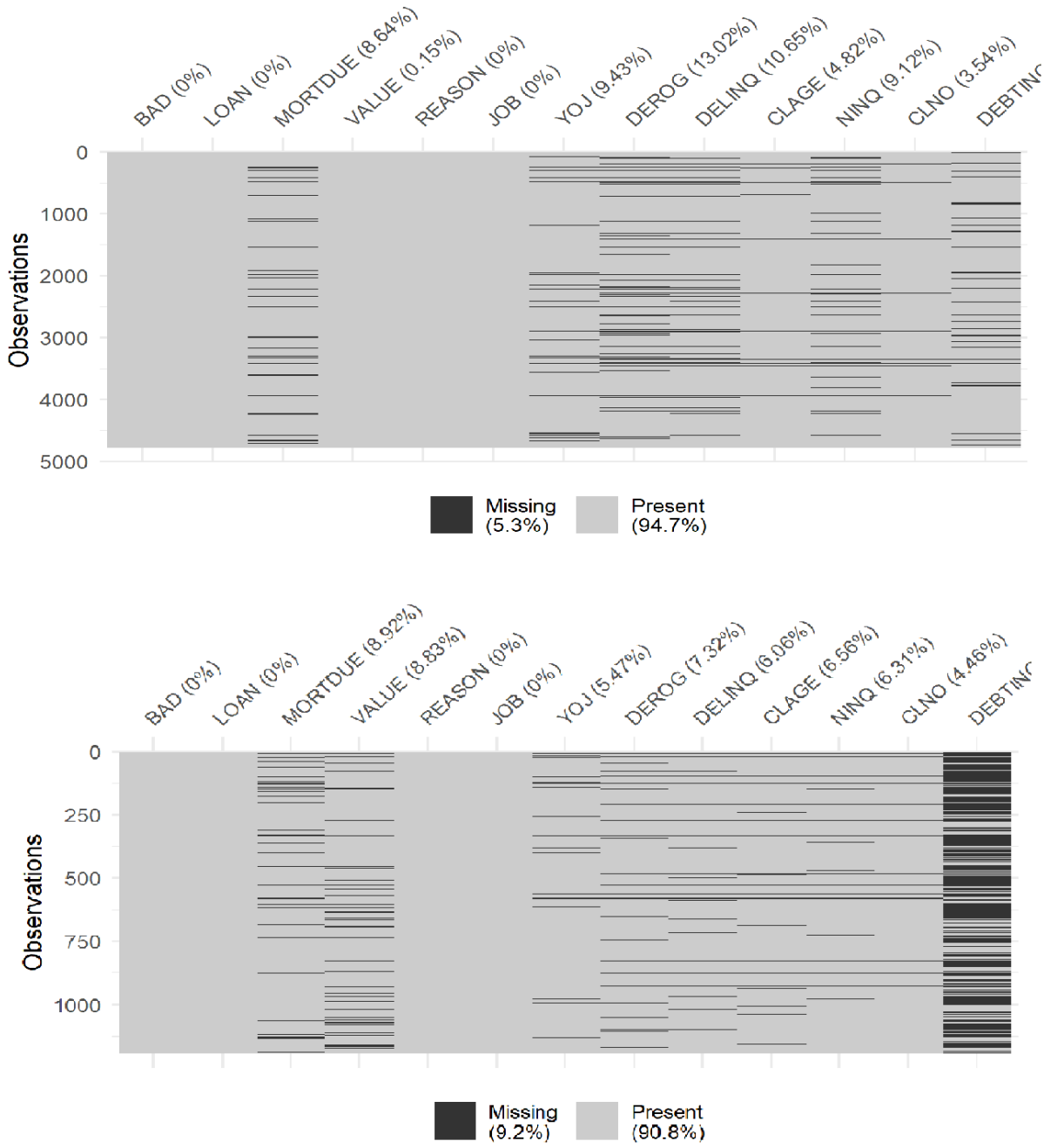
Notre projet s'inscrit donc dans cette veine d'un dessein réglementaire. En effet, nous aurions eu probablement de meilleures performances avec des modèles plus complexes (et nous aurions pu les comprendre), mais le but, de notre point de vu en tout cas, est de respecter ce qui est rigueur au sein des établissements bancaires, pour à la fois mieux mettre en exergue les méthodologies, ses points forts et ses points faibles.

En espérant que vous avez apprécié cette lecture.

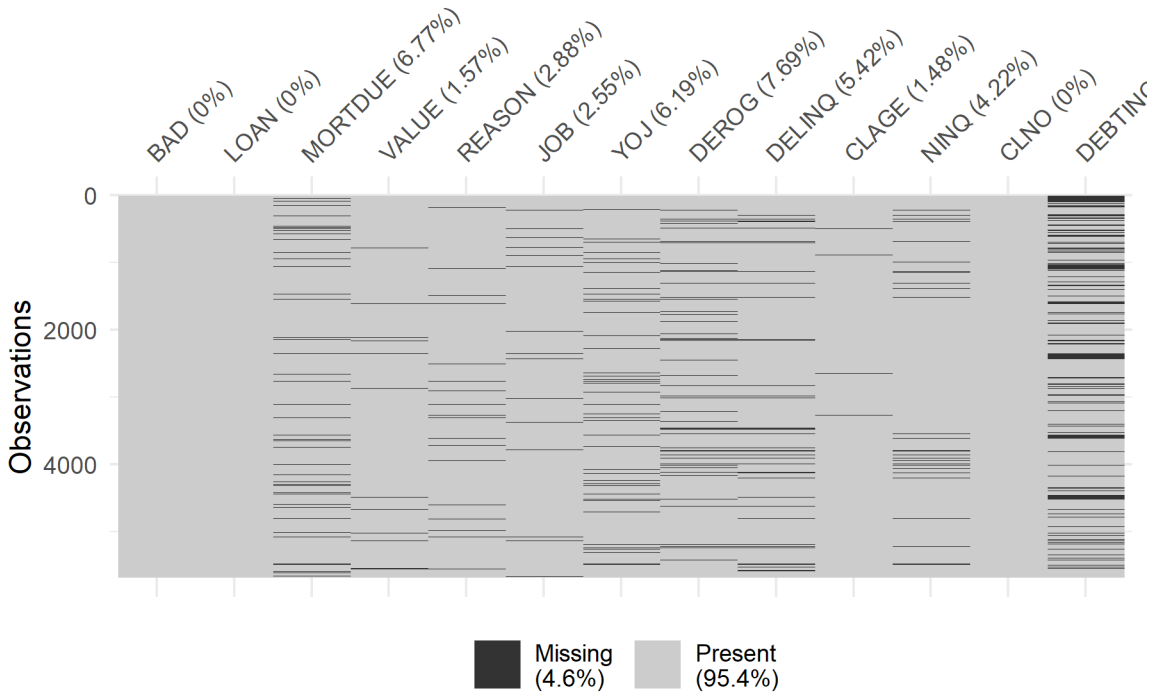
Annexes :



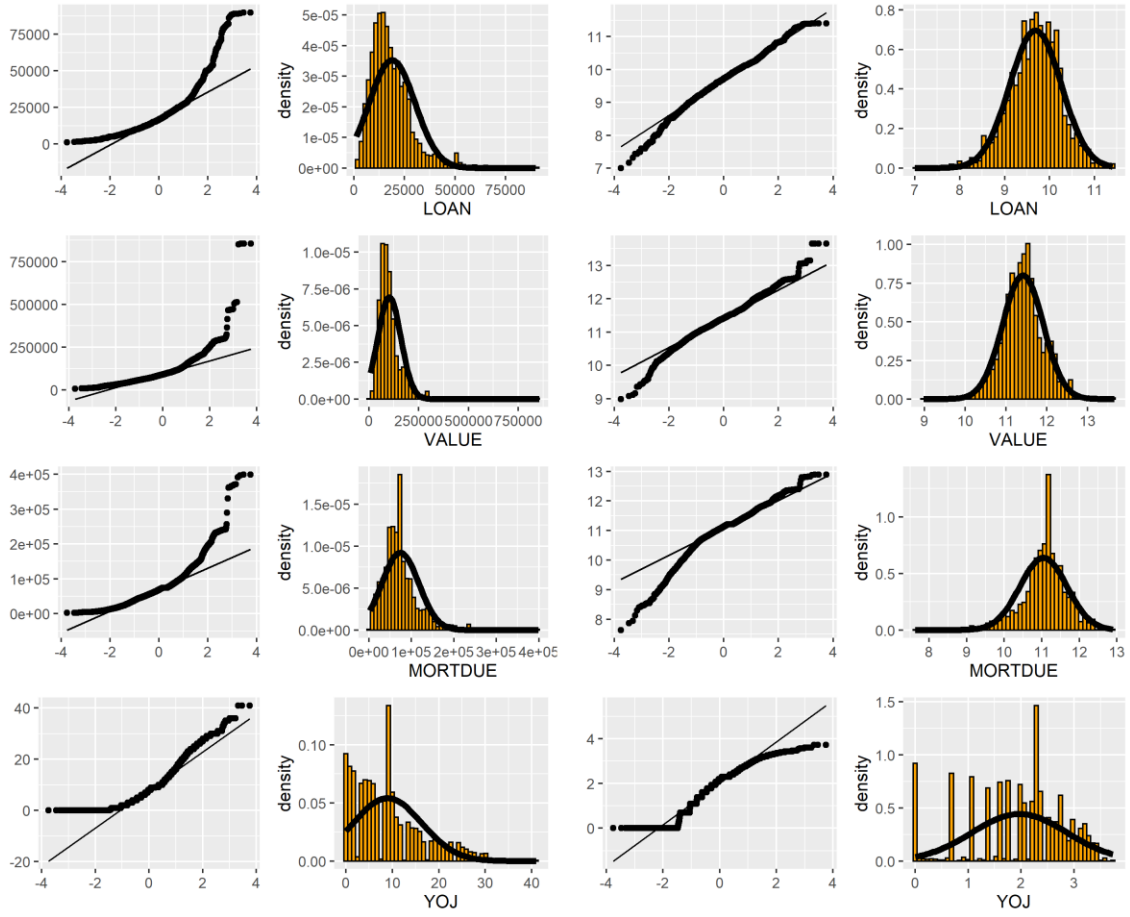
Annexe 1: Distribution des variables continues



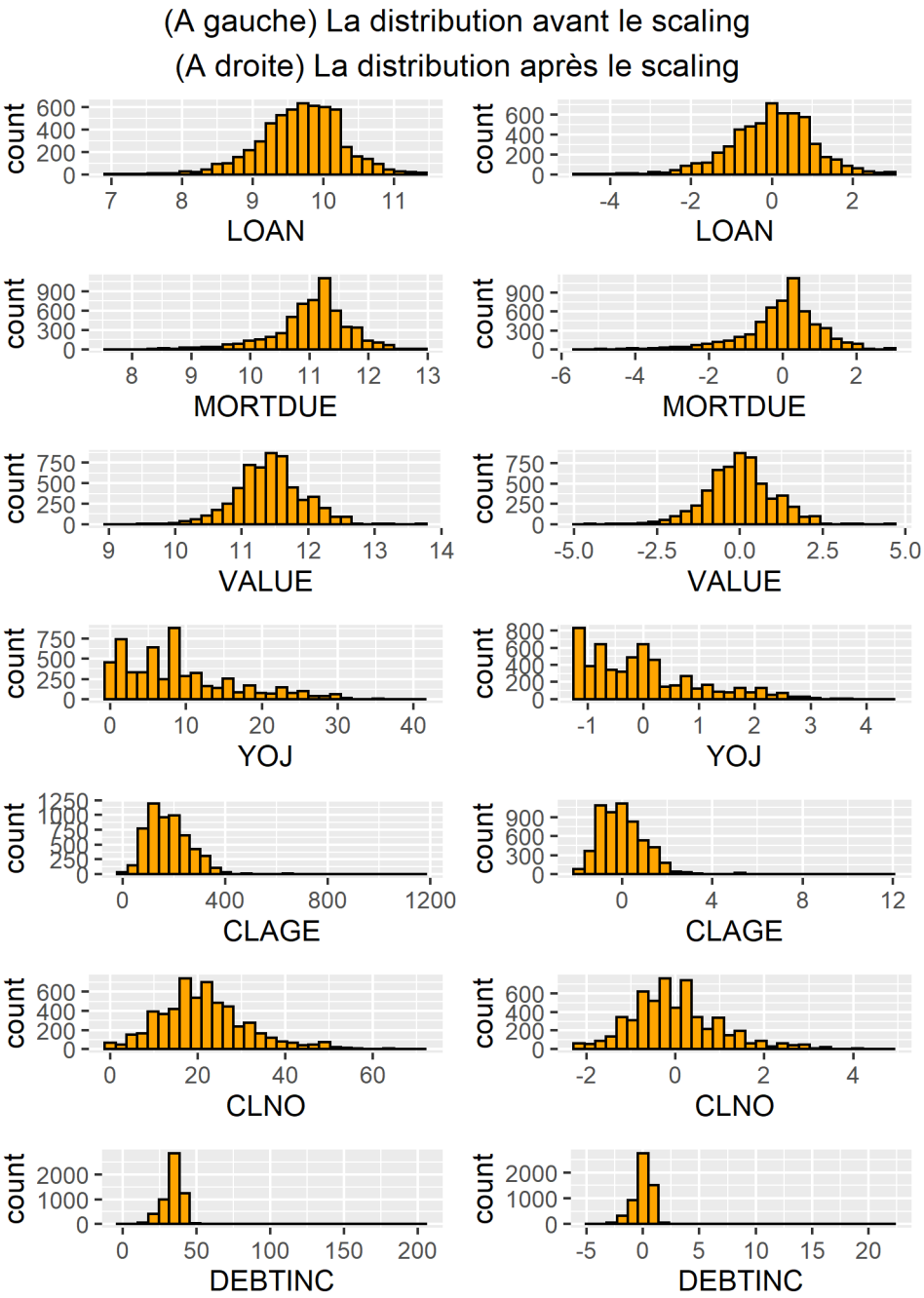
Annexe 2 : (En haut) Valeurs manquantes pour BAD = 0, (En bas) Valeurs manquantes pour BAD = 1



Annexe 3 : Variables manquantes après suppresion des lignes avec 37% ou plus de NA



Annexe 4 : QQPLot avant / après log transformation



Annexe 5 : Distribution avant / après le scale des variables numériques continues

```
## Test de déviance
logit_AIC$df.resid
] 4245
logit_AIC$deviance
] 3188.209
1-pchisq(logit_AIC$deviance,logit_AIC$df.resid)
] 1
```

Annexe 6 : Test de déviance du modèles logit sur le dataset AIC