Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Carthage

École Supérieure de la Statistique et de l'Analyse de l'Information

Rapport de stage d'insertion

Banque de tunisie (BT)

Crédit Scoring : Développement d'un modèle de classification en utilisant la régression logistique

Élaboré par : Bellalah Yesmine

Encadré par : M.BELGASMI Nabil

Juin 2016

Remerciements

Je remercie toutes les personnes qui ont contribué au succès de mon stage.

Tout d'abord, j'adresse mes remerciements à toute l'équipe pédagogique de l'ESSAI pour m'avoir préparé pour ce stage pendant ma première année.

Je tiens à remercier vivement mon maitre de stage, M. Nabil BELGASMI, pour son accueil, le temps passé ensemble et le partage de son expertise au quotidien.

Je remercie particulièrement ma famille pour son aide et encouragement.

Faire mon stage d'insertion à la Banque de Tunisie a été un plaisir, j'ai pu beaucoup apprendre .

Table des matières

Table des figures

1 Introduction

1.1 Le Crédit Scoring

Un crédit scoring est une expression numérique basée sur une analyse au niveau des dossiers de crédit d'une personne, pour représenter la solvabilité de la personne. Un pointage de crédit est principalement basée sur une information de rapport de crédit généralement provenant de bureaux de crédit.

Les prêteurs, comme les banques et les sociétés de cartes de crédit, utilisent les scores de crédit pour évaluer le risque potentiel posé par le prêt d'argent aux consommateurs et à atténuent les pertes dues à la mauvaise dette. Les prêteurs utilisent les scores de crédit pour déterminer qui est admissible à un prêt, à quel taux d'intérêt, et les limites que de crédit. Les prêteurs utilisent également les cotes de crédit pour déterminer quels clients sont susceptibles d'apporter le plus de revenus. L'utilisation du crédit ou de l'identité notation avant d'autoriser l'accès ou l'octroi de crédit est une implémentation d'un système de confiance.

Le crédit scoring ne se limite pas aux banques. D'autres organisations, telles que les sociétés de téléphonie mobile, les compagnies d'assurance, les propriétaires, et les ministères utilisent les mêmes techniques. Le pointage de crédit a aussi beaucoup de chevauchement avec l'exploration de données, qui utilise de nombreuses techniques similaires. Ces techniques se combinent des milliers de facteurs, mais sont similaires ou identiques.

1.2 Entreprise d'accueil : Banque de Tunisie

La Banque de Tunisie ou BT est une banque commerciale privée en Tunisie.
Fondée le 23 septembre 1884, elle est l'une des plus anciennes banques d'Afrique.

Histoire

En 1948, la Banque de Tunisie absorbe la Banque italo-française de crédit et participe en 1951 à la liquidation de la Banca Italiano de Credito dont elle reprend, après l'indépendance, la plupart des agences. Première banque à être installée en Tunisie, elle reste longtemps le seul établissement de crédit bancaire et, pendant longtemps, la seule banque à disposer d'agences dans le pays. En 1963, la Banque de Tunisie ouvre son capital à la Société générale, le Crédit suisse et la Banca Nazionale del Lavoro. En 1968, la Banque de Tunisie rachète les agences tunisiennes de la Compagnie française de crédit et de banque.

Après la révolution tunisienne, la PDG Alya Abdallah, épouse d'Abdelwahab Abdallah, conseiller du président déchu Ben Ali, ainsi que ses proches collaborateurs, sont interdits d'accès à leurs bureaux, la direction générale de la banque redoutant que ces derniers ne se livrent à des malversations financières au profit des anciens membres du régime.

Depuis le 21 janvier 2011, la Banque de Tunisie est placée sous le contrôle de la Banque centrale de Tunisie qui a nommé Habib Ben Sâad comme administrateur provisoire ; le conseil d'administration le nomme PDG le 25 janvier.

Actionnariat

Il s'agit d'une société anonyme dont le capital s'élève à 150 millions de dinars tunisiens. Au 31 décembre 2006, le capital est détenu à 73,03 % par des actionnaires tunisiens et à 22,78 % par des actionnaires étrangers. La BT compte parmi ses actionnaires des banques étrangères de réputation internationale telle que la Banque fédérative Crédit mutuel (20 %) et la Banca Nazionale del Lavoro. Les 13 % du capital appartenant à Belhassen Trabelsi sont confisqués à la suite de la révolution de 2011 et revendus à la Banque fédérative Crédit mutuel, qui devient détentrice de 33% du capital. La Banque fédérative du crédit mutuel est l'actionnaire de référence de la Banque de Tunisie dont elle détient, depuis le 31 décembre 2013, 34,44 % du capital.

Toutefois, la banque est sous-capitalisée bien qu'elle dispose d'importants fonds propres. Elle dispose par ailleurs d'une assise financière solide et affiche les meilleurs indicateurs du secteur avec un taux de couverture des créances classées de 97 %10.

En 1998, la Société générale cède sa participation dans le capital de la Banque de Tunisie essentiellement repris par des actionnaires locaux. Une part de 2 % est reprise par la filiale de l'Agence française de développement. Au 31 décembre 2008, le Crédit industriel et commercial vend sa participation à la Banque fédérale du crédit mutuel, sa maison mère

Toutefois, la banque est sous-capitalisée bien qu'elle dispose d'importants fonds propres. Elle dispose par ailleurs d'une assise financière solide et affiche les meilleurs indicateurs du secteur avec un taux de couverture des créances classées de 97 %10.

En 1998, la Société générale cède sa participation dans le capital de la Banque de Tunisie essentiellement repris par des actionnaires locaux. Une part de 2 % est reprise par la filiale de l'Agence française de développement. Au 31 décembre 2008, le Crédit industriel et commercial vend sa participation à la Banque fédérale du crédit mutuel, sa maison mère.

Réseau

En juin 2012, elle compte 103 agences et 165 distributeurs automatiques de billets répartis sur le territoire tunisien

Président Directeur Général

M. Mohamed Habib BEN SAAD

Directeurs Généraux Adjoints

M. Kamel JANDOUBI et M. Zouhair HASSEN

Le conseil d'administration

Le conseil d'administration est composé de 9 membres, dont deux administrateurs indépendants et un administrateur représentant les intérêts des petits porteurs et ce conformément à la réglementation en vigueur.

Mediateur

M. Mohamed Lotfi LABIDI

2 Régression Logistique

La régression logistique est une technique prédictive, c'est un type de modèle de classification. Elle vise à construire un modèle permettant de prédire / expliquer les valeurs prises par une variable cible qualitative , c'est la variable dépendante , elle est le plus souvent binaire, on parle alors de régression logistique binaire ; si elle possède plus de 2 modalités, on parle de régression logistique multinomiale à partir d'un ensemble de variables explicatives quantitatives et/ou qualitatives (codées), ce sont les variables indépendantes.

Le modèle LOGIT

Y = C (X1,X2 ,... ,Xk) = C (X)

f ne peut pas être une fonction linéaire parce que Y ne prend que deux valeurs.

La régression logistique est basée sur la fonction logistique($\Pi(X)$) qui a toujours une valeur comprise entre 0 et 1.

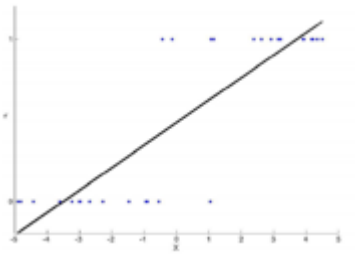

FIGURE 1 – Fonction linéaire

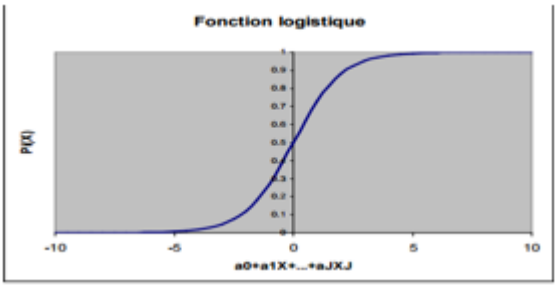

FIGURE 2 – Fontion logistique

Nous pouvons interpréter $\Pi(X)$ comme la probabilité $P(Y = 1/X)$: la probabilité que la variable dépendante est de classe 1, compte tenu des variables indépendantes.

2.1 Hypothèses

— Une relation linéaire et additive entre la variable réponse (dépendante) et les variables prédictives (indépendantes).

— Les variables indépendantes ne doivent pas être corrélées entre elles. L'absence de ce phénomène est connu sous multicollinéarité.

— Il devrait y avoir aucune corrélation entre les termes résiduels (erreur). L'absence de ce phénomène est connu sous autocorrélation.

2.2 Vérification des hypothèses et Solutions

— **Linéarité et additivité**

Si on correspond un modèle linéaire à un non-linéaire, un ensemble de données non-additif, l'algorithme de régression ne parviendrait pas à capturer la tendance mathématiquement, entraînant ainsi un modèle inefficace. En outre, cela se traduira par des prédictions erronées sur un ensemble de données invisible Il faut regarder le plot de valeurs résiduelles vs modelées.

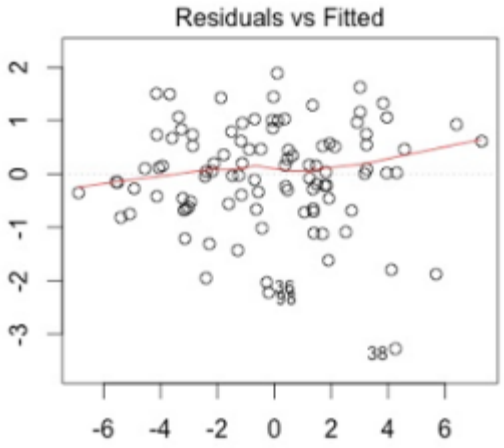

FIGURE 3 – Tendance linéaire

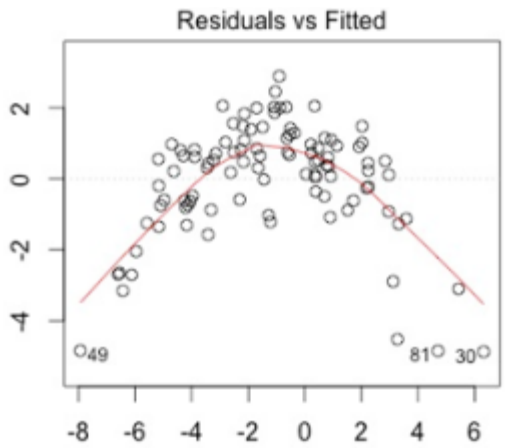

FIGURE 4 – Pas de linéarité

Solution : Pour surmonter le problème de la non-linéarité, on peut faire une transformation non linéaire des prédicteurs tels que $\log X$, $\sqrt{X}$ ou X^2 pour transformer la variable dépendante

— **Multicolinéarité**

— Test de Pearson : limité car la variables indépendante : peut être une combinaison linéaire de plusieurs variables indépendantes , mais pas fortement corrélés avec une quelconque d'entre elles)
— VIF (Variance Inflation Factor) ou Tolérance (1/vif=1- R^2) :
fournit un indice qui mesure combien la variance (le carré de l'écart-type R de l'estimation ^{2}) d'un coefficient de régression estimée est augmentée en raison de la colinéarité. Il est toujours supérieur ou égal à 1. Il n'y a pas de valeur VIF formelle pour déterminer la présence de multicollinéarité. Les valeurs de VIF qui dépassent 10 sont souvent considérés comme indiquant multicollinéarité, mais dans les valeurs des modèles les plus faibles au-dessus de 4 peut être une source de préoccupation.

Solution :
— Augmenter la taille de l'échantillon. Ce diminue généralement les erreurs standards
— Éliminer les variables à forte VIF

— **Autocorrélation**

— Test Durbin Watson (DW) :
chercher Durbin - Watson (DW) statistique. Elle doit se situer entre 0 et 4. Si DW = 2, implique aucune autocorrélation, 0 <DW < 2 implique autocorrélation positive tandis que 2 <DW < 4 indique autocorrélation négative.
L'invention de cette méthode est due à J. Durbin et G. S. Watson 1950. Son principe repose sur une étude de l'autocorrélation et donc du calcul des résidus. En clair, on effectue le rapport entre la somme des différences des résidus à t et t-1 et la somme des résidus (les écarts entre notre modèle de prédiction et les valeurs réelles).
Les hypothèses
On pose ρ comme étant l'autocorrélation de nos valeurs. On pose donc les hypothèses suivantes :
H0 : $\rho = 0$ (pas d'autocorrélation)
H1 : $\rho \neq 0$
Valeur pratique
La valeur pratique d est donc représentative de notre rapport d'autocorrélation. Celui-ci se calcule avec la formule suivante :

d= $\frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2}$
Avec e_t étant l'erreur également appelée résidu et représentative de la différence entre notre modèle de prédiction et les valeurs réelles.

— **Hétéroscédasticité**

— Breusch-Pagan / Cook - Test Weisberg ou White general test
— Plot de valeurs résiduelles vs modelées : Si hétéroscédasticité existe, le plot présenterait un motif de forme d'entonnoir.

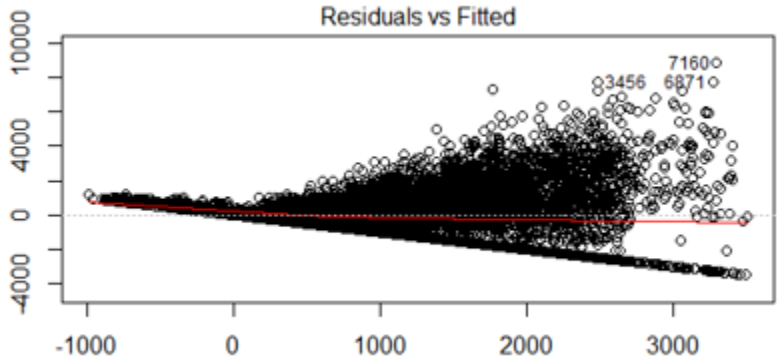

FIGURE 5 – Plot de valeurs résiduels vs valeurs modelées

Solution :
— Transformer la variable réponse telle que $\log Y$ ou $\sqrt{Y}$
— Utiliser la méthode de moindre carrés pondérés

— **Normalité de la distribution des termes d'erreur**

— Les tests de normalité : En statistique, les tests de normalité permettent de vérifier si des données réelles suivent une loi normale ou non.
Il y a plusieurs tests de noramlité : Shapiro-Wilk, Kolmogorov-Smirnov ...

2.3 Avantages et inconvénients

— **Avantages**
• Aucune relation linéaire entre la variable indépendante et les variables dépendantes doit être assumée.
• Peut gérer toutes sortes de relations.
• La variable indépendante peut être ordinale ou nominale et n' pas besoin d'être métrique (intervalle ou un ratio à l'échelle).
• Les variables dépendantes et les résidus ne doivent pas être distribués normalement. • Homoscédasticité n'est pas nécessaire.
— **Inconvénients**
• Nécessite un échantillon de grande taille pour obtenir des résultats stables.

3 Partie Pratique

3.1 Outil

On a utilisé le logiciel R (R studio) car

• c'est un logiciel multi-plateforme, qui fonctionne aussi bien sur des systèmes Linux, Mac OS X ou Windows

• R est un Open Source (c'est un logiciel libre), développé par ses utilisateurs et modifiable par tout un chacun

• Le premier logiciel d'analyse de données (plus de 49% des entreprises l'utilisent)

3.2 Base de données

On dispose de 2 bases de données Training.csv et Test.csv , la première sert a créer le modèle et la 2ème à le tester , les deux ont 500 observations (les crédits) et 21 variables , la variable réponse est " Creditability " et les autres variables sont les variables indépendantes dont 3 sont continues Age..years., Credit.Amount Duration.of.Credit..month.,et le reste est catégoriel : Payment.Status.of.Previous.Credit, Purpose, Value.Savings.Stocks, Length.of.current.employment, Instalment.per.cent,Sex...Marital.Status, Guarantors, Duration.in.Current.address, Most.valuable.available.asset, Account.Balance, Concurrent.Credits,Type.of.apartment, No.of.Credits.at.this.Bank, Occupation,No.of.dependents, Telephone, Foreign.Worker.

3.3 Analyse exploratoire des données (EDA)

On a réalisé une analyse exploratoire sur la base Training . Au début, on s'intéressait au valeurs manquantes , on les a visualisé et on a constaté qu'il n'y a aucune valeur manquante.

```
missmap(data,main="missing values per variable",col = c("white","blue"))
```

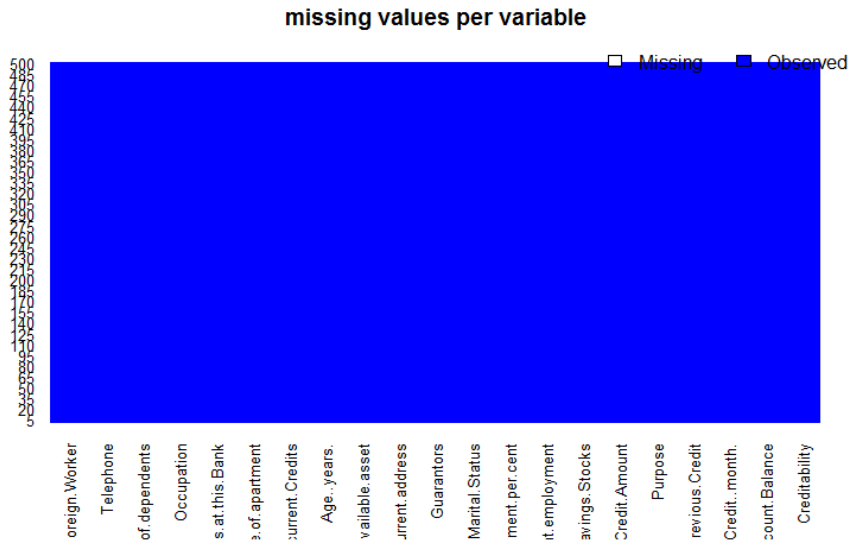

FIGURE 6 – Valeurs manquantes par variable

Ensuite on a commencé par visualiser la variable réponse "Creditability"

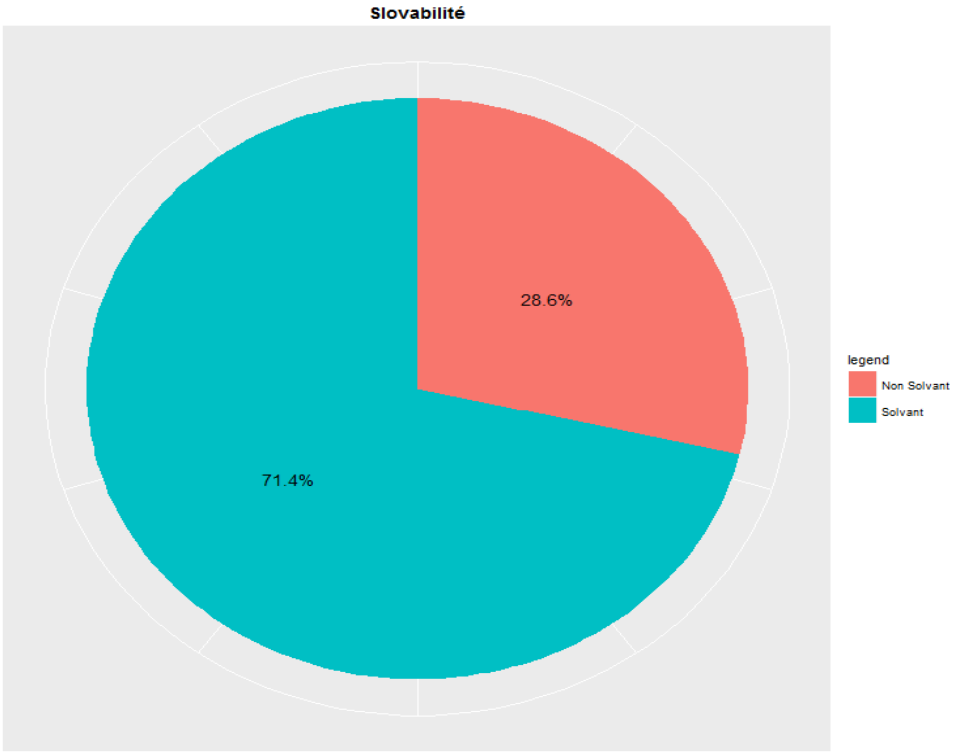

FIGURE 7 – Sovabilité

Par la suite , on a visualisé les variables continues (boxplot , fonction densité et histogramme) et les variables catégorielles (camembert)

— Variables continues

— Age " Age..years."

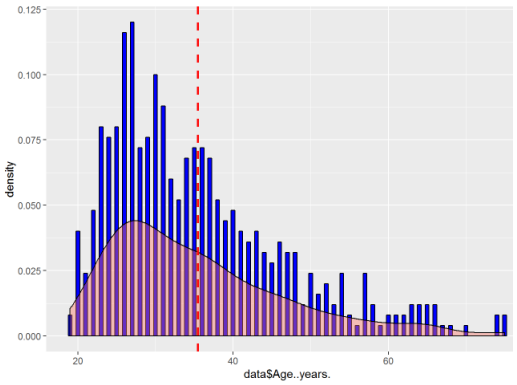

FIGURE 8 – Histogramme de Age..years.

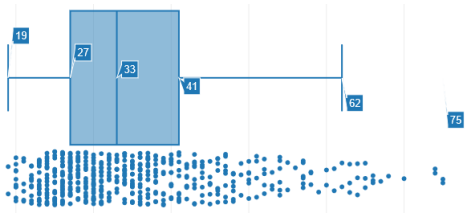

FIGURE 9 – Boxplot et nuagede points de Age..years.

— Montant du crédit "Credit.Amount"

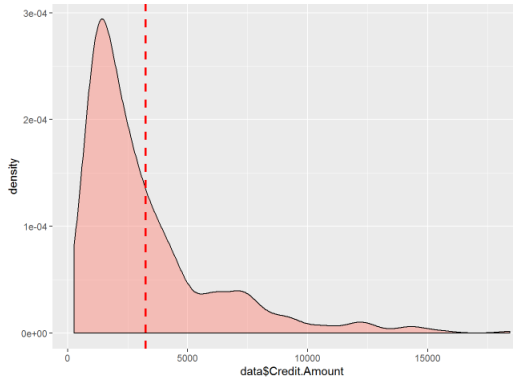

FIGURE 10 – Fonction densité de Credit.Amount

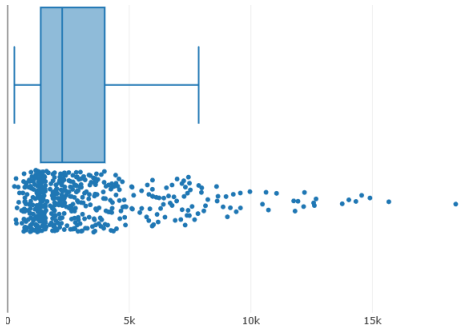

FIGURE 11 – Boxplot et nuage de points de Credit.Amount

— Durée du crédit ”

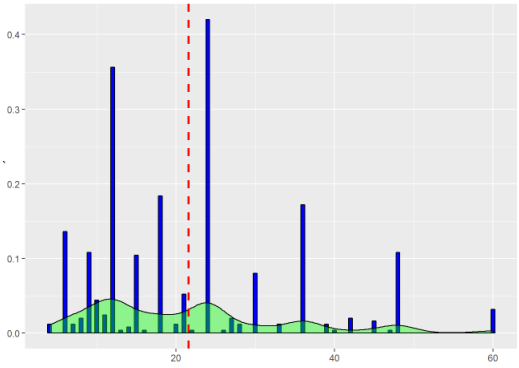

FIGURE 12 – Histogramme de la variable durée de crédit

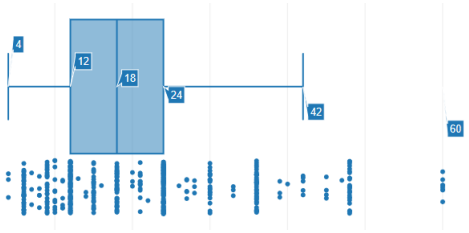

FIGURE 13 – Boxplot et nuage de points de la variable durée de crédit

— Variables catégorielles
Pour chaque variable on a représenté des camemberts dynamiques

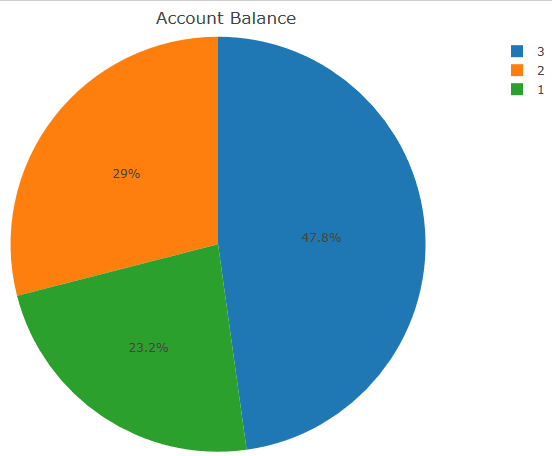

FIGURE 14 – Camembert de la variable Account Balance

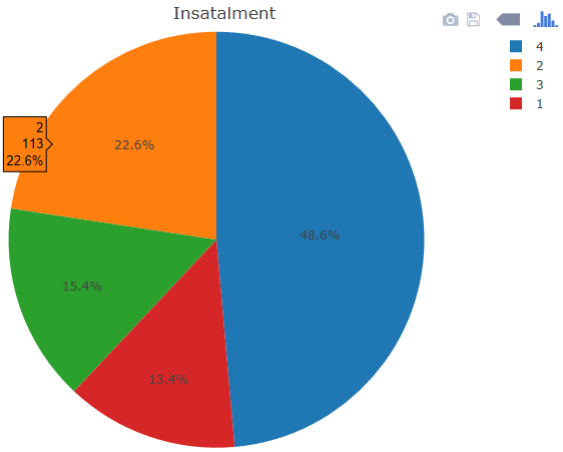

FIGURE 15 – Camembert de la variable Instalment

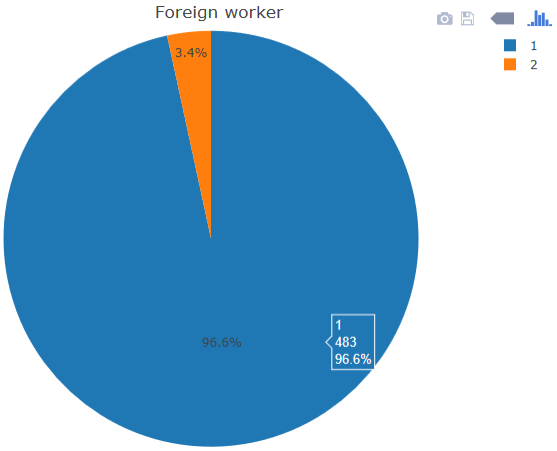

FIGURE 16 – Camembert de la variable Foreign Worker

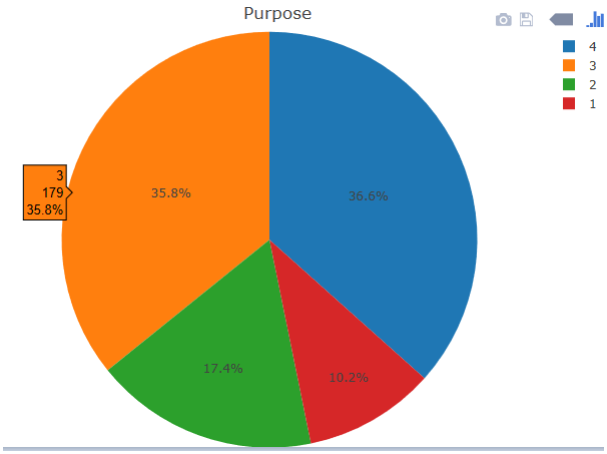

FIGURE 17 – Camembert de la variable Purpose

Finalement on a représenté la matrice de corrélation entre les variables continues

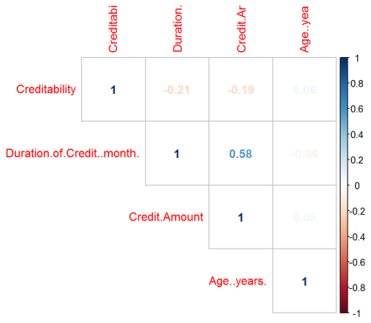

FIGURE 18 – Matrice de corrélation

3.4 Implémentation de la régression logistique

— 1ère étape : création d'un modèle basique en utilisant la commande "glm"

— 2ème étape : Amélioration du modèle
On a pensé à utiliser la méthode stepwise pour diminuer le nombre de varibles dépendantes dans le sens de garder que les variables significatives pour le modèle . Par la suite, on refait un modèle qui dépend seulement des varibales choisies par la commande "step"

```
m2=glm(formula = Creditability ~ Account.Balance + Duration.of.Credit..month. +
                 Payment.Status.of.Previous.Credit + Purpose + Credit.Amount +
                 Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent +
                 Sex...Marital.Status + Duration.in.Current.address + Concurrent.Credits +
                 No.of.dependents, family = binomial(link = "logit"), data = data)

summary(m2)
```

FIGURE 19 – Nouveau modèle

— 3ème étape : Vérification des hypothèses

• Multicolinéarité :en utilisant la méthode vif , on voit clairement qu'il n'y a pas de multicolinéarité (le vif de chaque variable ne dépasse pas 2)

```
library(car)
fit2=vif(m2)
fit2
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## Account.Balance                 1.295531  2        1.066871
## Duration.of.Credit..month.      1.640814  1        1.280943
## Payment.Status.of.Previous.Credit 1.293804  2        1.066515
## Purpose                         1.434471  3        1.061977
## Credit.Amount                   2.035508  1        1.426712
## Value.Savings.Stocks            1.279945  3        1.041994
## Length.of.current.employment    1.451263  3        1.064039
## Instalment.per.cent             1.518533  3        1.072105
## Sex...Marital.Status            1.422817  2        1.092162
## Duration.in.Current.address     1.476716  3        1.067127
## Concurrent.Credits              1.111842  1        1.054439
## No.of.dependents                1.256576  1        1.120971
```

FIGURE 20 – Appliquer la méthode VIf sur le nouveau modèle

• Autocorrélation : on a employé le test de Durbin Watson , ce qui a donné une p-valeur égale à 0.985(>>> 0.05) donc l'hypothèse d'autocorrélation est à rejeter.

```
durbinWatsonTest(m2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.00410068      1.990913   0.958
##  Alternative hypothesis: rho != 0
```

FIGURE 21 – Tester l'autocorrélation

3.5 Évaluation du modèle

Après avoir importé la base du test et lui appliquer le modèle de classification crée , on l'évalue.

— **Courbe ROC**
La fonction d'efficacité du récepteur, plus fréquemment désignée sous le terme « courbe ROC » (de l'anglais receiver operating characteristic, pour « caractéristique de fonctionnement du récepteur ») dite aussi caractéristique de performance (d'un test) ou courbe sensibilité/spécificité, est une mesure de la performance d'un classificateur binaire. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs (fraction des positifs qui sont effectivement détectés) en fonction du taux de faux positifs (fraction des négatifs qui sont détectés (incorrectement)).

```
plot(fitperf,col="blue",lwd=2,main="ROC Curve for Logistic regression ")
abline(a=0,b=1,lwd=2,lty=2,col="black")
```

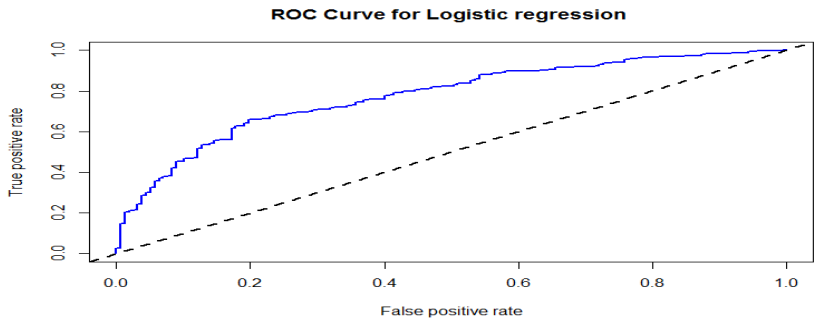

FIGURE 22 – Courbe ROC du modèle

— **AUC**
C'est la surface sous la courbe ROC (Area Under the Curve en anglais), c'est un outil pertinent pour mesurer la performance d'un classifieur et possède de nombreux avantages par rapport aux mesures de rappel et précision par classe : la performance est indiquée par une seule mesure et ne dépend pas des populations des classes.

```
perf=performance(fitpred, "auc")
perf@y.values[[1]]
```

```
## [1] 0.7757888
```

FIGURE 23 – La valeur de l'AUC

— **Matrice de confusion**
La matrice de confusion, dans la terminologie de l'apprentissage supervisé, est un outil servant à mesurer la qualité d'un système de classification.
Chaque colonne de la matrice représente le nombre d'occurrences d'une classe estimée, tandis que chaque ligne représente le nombre d'occurrences d'une classe réelle (ou de référence). Les données utilisées pour chacun de ces groupes doivent être différentes.

Un des intérêts de la matrice de confusion est qu'elle montre rapidement si le système parvient à classifier correctement.
pour ce modèle, on a obtenu cette matrice de confusion

	test$Creditability		
model_pred_creditability	0	1	Row Total
0	72	41	113
	37.584	17.203	
	0.637	0.363	0.226
	0.459	0.120	
	0.144	0.082	
1	85	302	387
	10.974	5.023	
	0.220	0.780	0.774
	0.541	0.880	
	0.170	0.604	
Column Total	157	343	500
	0.314	0.686	

FIGURE 24 – Matrice de confusion

On a également calculé par la suite le pourcentage d'erreur pour ce modèle ce qui a donné 25.2% de valeurs qui a été prédites à tort.

```
mean(model_pred_creditability!=test$Creditability)
```

```
## [1] 0.252
```

FIGURE 25 – Poucentage d'erreur

4 Conclusion

A travers ce stage on a eu l'occasion de découvrir une parmi les méthodes de scoring telle que la régression logistique.

La régression logistique est un modèle classique et basique entrainant un résultat d'aide à la décision (dans ce cas la solvabilité du créditeur). Ce modèle peut être amélioré avec les techniques de boosting.

Il existe d'autres modèles qu'on peut créer pour ce résoudre ce problème. On cite le réseau de neurones, analyse linéaire discriminante, Adaboost...

Mieux encore, on peut combiner plusieurs modèles pour créer un seul plus efficace et plus performant.

5 Webographie

[1] www.http://scg.sdsu.edu/logit_r/

[2] www.http://https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package utm_content=bufferacfca&utm_medium=social&utm_source=linkedin.com&utm_campaign=buffer

[3] www.http://https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-pl ?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+ Vidhya%29

[4] www.https://www.coursera.org/learn/regression-modeling-practice

[5] http://www.kdnuggets.com/2016/02/ensemble-methods-techniques-produce-improved-machine-learning. html