Modèle de Prédiction du Risque de Défaut de Paiement.

Saïd Toubra

20 septembre 2023

1 Introduction

Les services de prêts financiers sont offerts par un large éventail d'entreprises, allant des grandes banques aux institutions financières et aux programmes gouvernementaux de prêt. L'un des principaux objectifs de ces entreprises est de minimiser les risques de défaut de paiement et de s'assurer que les emprunteurs remboursent leurs prêts conformément aux termes convenus.

Pour atteindre cet objectif de manière efficace et méthodique, de nombreuses entreprises se tournent vers l'apprentissage automatique. Cette technologie leur permet de prédire avec précision quelles personnes présentent le plus grand risque de défaut de paiement de leurs prêts, facilitant ainsi le déploiement ciblé de mesures d'intervention appropriées.

Dans cette démarche, nous aborderons l'un des problèmes les plus cruciaux de l'industrie financière en utilisant un ensemble de données unique provenant du *Loan Default Prediction Challenge de Coursera*. Cet ensemble de données comprend un total de 255 347 lignes, représentant des individus, et 16 colonnes contenant diverses variables d'information.

L'ensemble des variables constituant notre base de données sont :

Age (entier) : Âge de l'emprunteur.

Income (entier) : Revenu annuel de l'emprunteur.

LoanAmount (entier) : Montant d'argent emprunté.

CreditScore (entier) : Cote de crédit de l'emprunteur.

MonthsEmployer (entier) : Nombre de mois pendant lesquels l'emprunteur a été employé.

NumCreditLines (entier) : Nombre de lignes de crédit ouvertes par l'emprunteur.

InterestRate (entier) : Taux d'intérêt du prêt.

LoaneTerm (entier) : Durée du prêt en mois.

DTIRatio (réel) : Ratio dette/revenu.

EmploymentType (chaînes de caractères) : Statut professionnel de l'emprunteur (temps plains, temps partiel, travailleur indépendant, sans emploi).

MaritalStatus (chaînes de caractères) : état civil de l'emprunteur (célibataire, marié, divorcé).

HasMontgage (binomial) : si l'emprunteur est un prêt hypothécaire (No/Yes).

HasDependents (binomial) : si l'emprunteur est dépendant (No/Yes).

LoanPurpose (chaînes de caractères) : Le but du prêt (maison, voiture, éducation,autre).

HasCoSigner (binomial) : Si l'emprunteur est cosignataire (No/Yes).

Default (binomial) : Indique s'il y' a défaut de paiement (1/0).

La Figure 1 représente un résumé généré à partir d'un script R qui présente une synthèse de l'ensemble de données.

```
> summary(data)
      Age            Income         LoanAmount      CreditScore    MonthsEmployed   NumCreditLines   InterestRate
 Min.   :18.0   Min.   : 15000   Min.   :  5000   Min.   :300.0   Min.   :  0.00   Min.   :1.000   Min.   : 2.00
 1st Qu.:31.0   1st Qu.: 48892   1st Qu.: 65992   1st Qu.:437.0   1st Qu.: 30.00   1st Qu.:2.000   1st Qu.: 7.76
 Median :43.0   Median : 82524   Median :127429   Median :574.0   Median : 59.00   Median :3.000   Median :13.46
 Mean   :43.5   Mean   : 82485   Mean   :127506   Mean   :574.4   Mean   : 59.51   Mean   :2.501   Mean   :13.49
 3rd Qu.:56.0   3rd Qu.:116053   3rd Qu.:188970   3rd Qu.:712.0   3rd Qu.: 89.00   3rd Qu.:3.000   3rd Qu.:19.23
 Max.   :69.0   Max.   :149999   Max.   :249999   Max.   :849.0   Max.   :119.00   Max.   :4.000   Max.   :25.00

    LoanTerm        DTIRatio           EmploymentType   MaritalStatus   HasMortgage HasDependents    LoanPurpose
 Min.   :12.00   Min.   :0.1000                :    1          :    1      :    1      :    1             :    1
 1st Qu.:24.00   1st Qu.:0.3000   Full-time    :31889   Divorced:42747   No :63822   No :64079     Auto     :25361
 Median :36.00   Median :0.5000   Part-time    :32263   Married :42577   Yes:64094   Yes:63837     Business :25844
 Mean   :36.04   Mean   :0.5002   Self-employed:31739   Single  :42592                             Education:25646
 3rd Qu.:48.00   3rd Qu.:0.7000   Unemployed   :32025                                             Home     :25659
 Max.   :60.00   Max.   :0.9000                                                                  Other    :25406

 HasCoSigner    Default
    :    1   Min.   :0.0000
 No :63998   1st Qu.:0.0000
 Yes:63918   Median :0.0000
             Mean   :0.1165
             3rd Qu.:0.0000
             Max.   :1.0000
             NA's   :1
```

FIGURE 1 – Résumer de l'ensemble de données

L'objectif de notre travail est de développer un modèle prédictif capable d'évaluer le risque de défaut de paiement d'un emprunteur. Pour atteindre cet objectif, nous avons opté pour l'utilisation de la régression logistique. Cette approche nous permettra non seulement de construire un modèle prédictif, mais aussi d'analyser les variables qui le composent afin de mieux comprendre son fonctionnement.

Pour ce faire, nous diviserons nos données en deux sous-ensembles distincts : un ensemble de données d'apprentissage (data train) représentant 70% de l'ensemble initial, et un ensemble de données de test (data test) représentant les 30% restants.

```
1     set.seed(1)
2     sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=
    c(0.7,0.3))
3     data_train  <- data[sample, ]
4     data_test   <- data[!sample, ]
5     nrow(data_train)
6     [1] 89562
7     nrow(data_test)
8     [1] 38355
9
```

Nous disposons désormais deux nouvelles bases de données : une base de données d'apprentissage comprenant 89 562 emprunteurs et 16 variables que nous utiliserons pour construire

notre modèle prédictif, ainsi qu'une base de données de teste comprenant 38 355 emprunteurs et 16 variables que nous utiliserons pour évaluer l'efficacité prédictive de notre modèle.

Maintenant que nous avons défini le problème à résoudre et que nous disposons les données nécessaires, nous allons procéder à la construction et à la validation de notre modèle.

2 Construction et validation du modèle

Avant de passer au processus de sélection des variables pertinentes, nous allons créer un modèle qui explique la variable « Default » en fonction de toutes les autres variables. Il s'agit d'un modèle de référence que nous pourrons utiliser pour déterminer si la sélection des variables améliore notre modèle ou non.

```
1     modele <- glm(Default ~ .,data = data_train, family = binomial)
2
```

```
> summary(modele)

Call:
glm(formula = Default ~ ., family = binomial, data = data_train)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.127e+00  8.381e-02 -13.450  < 2e-16 ***
Age                        -1.866e-02  7.135e-04 -26.151  < 2e-16 ***
Income                     -3.932e-06  2.726e-07 -14.423  < 2e-16 ***
LoanAmount                  2.020e-06  1.491e-07  13.546  < 2e-16 ***
CreditScore                -3.731e-04  6.632e-05  -5.625 1.85e-08 ***
MonthsEmployed             -4.403e-03  3.054e-04 -14.417  < 2e-16 ***
NumCreditLines              3.022e-02  9.438e-03   3.202  0.00136 **
InterestRate                3.259e-02  1.604e-03  20.315  < 2e-16 ***
LoanTerm                   -3.382e-04  6.217e-04  -0.544  0.58644
DTIRatio                    1.192e-01  4.568e-02   2.610  0.00904 **
EmploymentTypePart-time     3.456e-01  3.115e-02  11.095  < 2e-16 ***
EmploymentTypeSelf-employed 2.545e-01  3.170e-02   8.030 9.75e-16 ***
EmploymentTypeUnemployed    4.678e-01  3.063e-02  15.276  < 2e-16 ***
MaritalStatusMarried       -2.432e-01  2.604e-02  -9.341  < 2e-16 ***
MaritalStatusSingle        -7.717e-02  2.522e-02  -3.060  0.00222 **
HasMortgageYes             -1.469e-01  2.110e-02  -6.962 3.35e-12 ***
HasDependentsYes           -2.314e-01  2.116e-02 -10.937  < 2e-16 ***
LoanPurposeBusiness         1.034e-02  3.262e-02   0.317  0.75134
LoanPurposeEducation       -4.788e-02  3.303e-02  -1.450  0.14720
LoanPurposeHome            -1.954e-01  3.394e-02  -5.758 8.53e-09 ***
LoanPurposeOther           -4.977e-02  3.313e-02  -1.502  0.13305
HasCoSignerYes             -2.339e-01  2.116e-02 -11.055  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64728  on 89560  degrees of freedom
Residual deviance: 62319  on 89539  degrees of freedom
  (1 observation effacée parce que manquante)
AIC: 62363

Number of Fisher Scoring iterations: 5
```

FIGURE 2 – Modèle de référence.

Avant de nous intéressait aux résultats de notre modèle de référence, nous allons vérifie s'il fait mieux que le modèle ayant comme seule coefficient le constant (modèle trivial). Pour ceux-la, nous allons applique le *test du rapport de vraisemblance*[2] entre notre modèle et le modèle trivial.

Définition 1 (**Test du rapport de vraisemblance ou de la déviance**)**.** La statistique de test est basée sur la différence des rapports de vraisemblance entre le modèle complet et le modèle trivial (sous H_0). On note $\hat{\beta}_{H_0}$ l'estimateur du maximum de vraisemblance contraint H_0 (il s'obtient en supprimant les q premières variables du modèle). On a alors sous :

$$2\left(\mathcal{L}_n(\hat{\beta}) - \mathcal{L}_(\hat{\beta}_{H_0})\right) \underset{\nearrow}{\mathcal{L}} \chi^2_q.$$

où $\mathcal{L}_n$ log-vraisemblance du modèle.

Prise de décision : Si la p-value est inférieure au seuil de signification (généralement fixé à 0,05 ou un autre seuil choisi), on rejette l'hypothèse nulle. Cela signifie que notre modèle fait mieux que le modèle trivial.

Appliquons maintenant ceux-là dans notre modèle en utilisant des scripte R.

```
D <- modele$null.deviance - modele$deviance
[1] 2409.72
ddl <- modele$df.null - modele$df.residual
[1] 21
p_value <- pchisq(D,ddl,lower.tail = FALSE)
[1] 0

```

La p-value étant inférieure à 0.05, nous rejetons l'hypothèse nulle, ce qui suggère que notre modèle performe mieux que le modèle trivial. Nous pouvons donc légitimement utiliser notre modèle comme point de référence. Maintenant que nous avons validé notre modèle de référence, concentrons-nous sur l'interprétation des résultats. Pour faciliter la compréhension des résultats présentés dans la Figure 2, clarifions d'abord le concept de « test de Wald »[3] qui est base sur l'hypothèse selon laquelle les tests de maximum de vraisemblance sont asymptotiquement normaux. Ce test est l'outil utilisé par la fonction « glm » pour calculer les p-values des coefficients.

Définition 2 (Test de Wald)**.** Le *test de Wald* est égale au rapport entre le carré d'un coefficient et sa variance, comme illustré par l'équation suivante :

$$W_{\hat{\beta}_i} = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2_{\hat{\beta}_i}}.$$

La statistique de test de Wald suit une loi du χ^2 à 1 degré de liberté.

Une fois que la statistique de test de Wald est calculée, on la compare à une distribution de probabilité (généralement une distribution du chi-2) pour obtenir une p-value. La p-value mesure la probabilité d'observer une statistique de test au moins aussi extrême que celle calculée, sous l'hypothèse nulle.

Prise de décision : Si la p-value est inférieure au seuil de signification (généralement fixé à 0,05 ou un autre seuil choisi), on rejette l'hypothèse nulle en faveur de l'hypothèse alternative. Cela signifie que le coefficient est considéré comme statistiquement significatif, indiquant un effet non nul de la variable correspondante sur la variable dépendante.

Dans ce sens, les résultat présentés dans la Figure 2 montre que selon le *test de Wald* avec un seuil de signification de 5%, toutes les variables du modèle sont considérées comme pertinentes, à l'exception de la variable « LoanTerm ».

Prenons l'exemple de la variable « Âge ». Le p-value associé est extrêmement faible (2×10^{-16}), ce qui indique une forte significativité au seuil de 5%. De plus, le coefficient de cette

variable est négatif (-1.866×10^{-2}), ce qui signifie que plus une personne est âgée, moins elle a de chances de se retrouver en défaut de paiement. En résumé, au sens du test de Wald, l'âge est une variable pertinente pour prédire le risque de défaut de paiement, et cela suggère que les individus plus jeunes ont un risque plus élevé de défaut de paiement que les personnes plus âgés.

En appliquant un raisonnement similaire à la variable « MaritalStatus », on peut constater que, selon le test de Wald, par rapport aux personnes divorcées, les personnes mariées et célibataires ont moins de chances de se retrouver en défaut de paiement.

Il est essentiel de noter que ces conclusions reposent sur les résultats obtenus grâce à l'application du *test de Wald*. Toutefois, une analyse plus approfondie et une validation du modèle sont nécessaires pour confirmer la pertinence et l'importance de ces variables dans la prédiction du défaut de paiement. Ces résultats préliminaires sont prometteurs, mais ils doivent être soumis à un examen plus détaillé après la sélection et la validation du modèle afin de garantir leur fiabilité.

2.1 Sélection des variables

Il existe plusieurs méthodes de sélection automatique des variables (backward, forward ou stepwise), et dans notre cas, nous avons choisi la méthode « backward ». Cela signifie que nous allons effectuer une sélection de modèle régressive vers l'arrière. il faut noté que le processus de sélection de *modèle backward* commence avec un modèle de régression comprenant toutes les variables explicatives potentielles disponibles ce qui est le cas de notre modèle de référence (Figure 2). Ensuite, il élimine progressivement les variables qui ne contribuent pas de manière significative à l'explication de la variation de la variable dépendante.

Voici le scripte R qui nous permet effectué ces calcules :

```
1  library(MASS)
2  modele.backward_BIC <- stepAIC(modele, data = data_train, direction="
     backward", k = log(nrow(data_train)), trace = FALSE)
3  summary(modele.backward_BIC)
```

Les résultat obtenu sont :

```
Call:
glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
    MonthsEmployed + InterestRate + EmploymentType + MaritalStatus +
    HasMortgage + HasDependents + LoanPurpose + HasCoSigner,
    family = binomial, data = data_train)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.003e+00  7.370e-02 -13.607  < 2e-16 ***
Age                        -1.867e-02  7.133e-04 -26.180  < 2e-16 ***
Income                     -3.931e-06  2.726e-07 -14.421  < 2e-16 ***
LoanAmount                  2.022e-06  1.491e-07  13.560  < 2e-16 ***
CreditScore                -3.741e-04  6.631e-05  -5.642 1.68e-08 ***
MonthsEmployed             -4.394e-03  3.053e-04 -14.391  < 2e-16 ***
InterestRate                3.260e-02  1.604e-03  20.325  < 2e-16 ***
EmploymentTypePart-time     3.458e-01  3.114e-02  11.104  < 2e-16 ***
EmploymentTypeSelf-employed 2.544e-01  3.169e-02   8.028 9.89e-16 ***
EmploymentTypeUnemployed    4.681e-01  3.062e-02  15.285  < 2e-16 ***
MaritalStatusMarried       -2.432e-01  2.604e-02  -9.340  < 2e-16 ***
MaritalStatusSingle        -7.670e-02  2.522e-02  -3.042  0.00235 **
HasMortgageYes             -1.472e-01  2.110e-02  -6.976 3.05e-12 ***
HasDependentsYes           -2.314e-01  2.115e-02 -10.938  < 2e-16 ***
LoanPurposeBusiness         9.903e-03  3.262e-02   0.304  0.76143
LoanPurposeEducation       -4.788e-02  3.303e-02  -1.450  0.14715
LoanPurposeHome            -1.960e-01  3.394e-02  -5.774 7.73e-09 ***
LoanPurposeOther           -5.074e-02  3.313e-02  -1.532  0.12563
HasCoSignerYes             -2.336e-01  2.116e-02 -11.042  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64728  on 89560  degrees of freedom
Residual deviance: 62336  on 89542  degrees of freedom
  (1 observation effacée parce que manquante)
AIC: 62374

Number of Fisher Scoring iterations: 5
```

FIGURE 3 – Modèle après sélection des variables

Nous avons notre nouvelle modèle, et nous allons vérifié s'il fait mieux que notre modèle de référence à l'aide de Test du rapport de vraisemblance.

```
1     D <- modele.backward_BIC$deviance - modele$deviance
2     [1] 17.29828
3     > ddl <- modele.backward_BIC$df.residual - modele$df.residual
4     [1] 3
5     > p_value <- pchisq(D,ddl,lower.tail = FALSE)
6     [1] 0.0006136065
7
```

La p-value vaut 0.61×10^{-3}, elle est fortement significative au seuil de 5%. Cela suggère qu'au sens du test de vraisemblance notre modèle surpasse significativement le modèle de référence. Il serait judicieux de faire recoure à des méthodes de validation de modèle pour évaluer la capacité de notre modèle à prédire avec précision si un emprunteur risque d'être en défaut de paiement ou non.

2.2 Validation du modèle

Pour valider notre modèle, nous allons utiliser un outil appelé le « diagramme de fiabilité »[3]. Ce diagramme permet d'évaluer la cohérence entre les scores modélisé par notre modèle de régression logistique et les scores observés, en particulier la proportion d'observations appartenant à la catégorie cible (c'est-à-dire la proportion de vrais positifs). Cette évaluation se fait en regroupant les données en fonction des scores attribués par le modèle. Le diagramme de fiabilité nous permet ainsi de visualiser la qualité de la calibration du modèle, c'est-à-dire sa capacité à déterminer correctement les probabilités d'appartenance à la classe cible pour différentes tranches de données.

En résumé, le diagramme de fiabilité est un outil essentiel pour vérifier si les scores de probabilité produits par notre modèle de régression logistique sont en accord avec les scores réellement observés, en organisant les données en groupes en fonction des scores prédits. Cela nous permet d'évaluer la qualité de la calibration du modèle et sa capacité à déterminer avec précision les vrais positifs.

Voici un script en R permettant de tracer la courbe de fiabilité :

```
1
2  diarammeDeFiablite <- function(modele,data,varReponce){
3    #probabilit s d'affectation
4    scores <- predict(modele, newdata = data, type = "response")
5    print(summary(scores))
6      #groupes de d coupage
7      decoupe <- cut(scores,breaks=seq(from=0.0, to=1.0, by=0.2),include.
     lowest=T)
8      print(table(decoupe))
9      #moyenne des probabilit s par groupe
10   mean.scores <- tapply(scores,decoupe,mean)
11   print(mean.scores)
12   #proportion des positifs par groupe
13   prop.pos <- apply(table(decoupe,varReponce),1,function(x){x[2]/sum(x)
     })
14   print(prop.pos)
15   #diagramme de fiabilit?
16     plot(mean.scores,prop.pos,main="Diagramme de fiabilit ",type="b",
     xlim=c(0,1),ylim=c(0,1))
17   abline(a=0,b=1)
18   }
19 diarammeDeFiablite(modele.backward_BIC,data_test,data_test$Default)
```

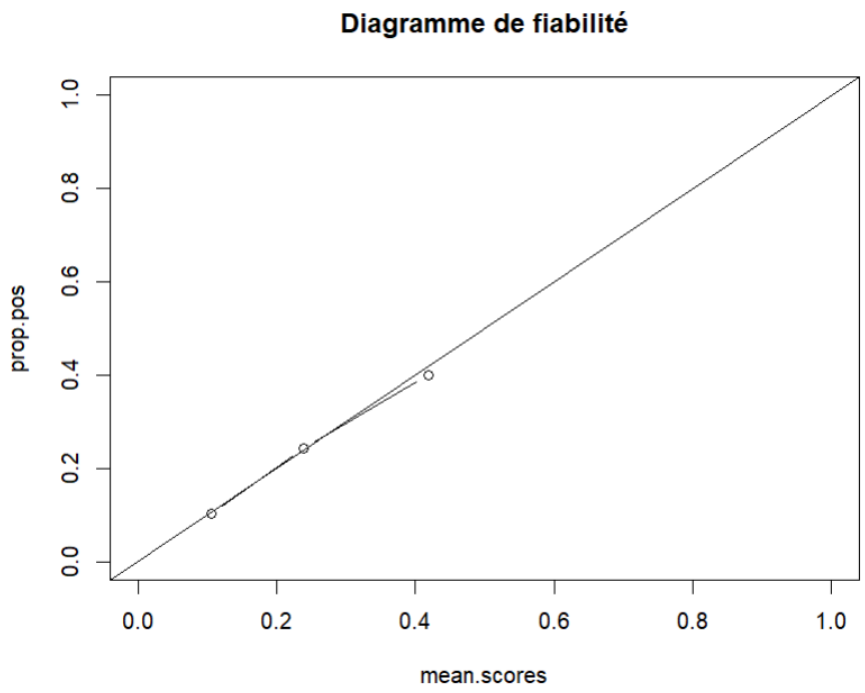

FIGURE 4 – Diagramme de fiabilité

Nous avons observé que les points sur la courbe de fiabilité semblent former pratiquement une droite, ce qui suggère que notre modèle est bien calibré. Cette observation a des implications importantes. Premièrement, cela indique que notre modèle n'est pas en surapprentissage, ce qui signifie qu'il ne s'est pas trop adapté aux données d'entraînement au point de perdre sa généralisation. Deuxièmement, cela suggère que notre modèle est capable de calculer avec précision les scores d'appartenance aux classes, ce qui est essentiel pour des prédictions fiables.

Maintenant, pour renforcer ces conclusions, nous allons utiliser un outil statistique appelé le test de « Hosmer & Lemeshow »[1]. Ce test repose sur le même principe que le diagramme de fiabilité que nous avons récemment tracé. Il nous permettra de quantifier la calibration de notre modèle de régression logistique et de confirmer nos observations précédentes.

Vous pouvez maintenant poursuivre en appliquant le test de « Hosmer & Lemeshow » à nos données pour une évaluation plus approfondie de la calibration de votre modèle.

```
1  testHosmerLemeshow <- function(modele,data,varReponce){
2  library(ResourceSelection)
3  scores <- predict(modele.backward_BIC, newdata = data_test, type = "
     response")
4  #appel de la fonction
5  ResourceSelection::hoslem.test(unclass(data_test$Default)-1,scores)
6  }
7  testHosmerLemeshow(modele.backward_BIC,data_test,data_test$Default)
```

Résultat du test :

```
1     Hosmer and Lemeshow goodness of fit (GOF) test
2     data:  unclass(data_test$Default) - 1, scores
3     X-squared = 445008, df = 8, p-value < 2.2e-16
4
```

On constate que la valeur du p-value est extrêmement faible (2.2×10^{-16}, ce qui est fortement significative au seuil de 5%. Cette constatation indique que les scores observés sont en parfait accord avec les scores modélisés par notre modèle. Par conséquent, nous pouvons conclure que notre modèle est bien calibré, car il fournit des scores qui reflètent de manière réaliste la probabilité d'appartenance aux différentes classes. Cette validation statistique renforce notre confiance dans la performance et la fiabilité de notre modèle.

2.3 Analyse des performances prédicatives du modèle

Nous souhaitons évaluer dans quelle mesure notre modèle attribue des scores plus élevés aux individus appartenant à la modalité cible par rapport à l'autre modalité. Pour ce faire, nous allons tout d'abord, tracer la courbe ROC (Receiver Operating Characteristic)[2] . Cette courbe nous permettra de visualiser la performance de notre modèle de classification en analysant sa capacité à distinguer les vrais positifs des faux positifs à différents seuils de probabilité.

Ensuite, nous allons calculer l'AUC (Area Under the Curve)[2]. L'AUC représente la probabilité qu'un individu choisi au hasard appartenant à la modalité cible ait un score de probabilité plus élevé qu'un individu choisi au hasard appartenant à la modalité négative. Si l'AUC est égal à 1, cela signifie que tous les individus de la modalité cible ont des probabilités plus élevées que tous les individus de la modalité négative.

En résumé, en traçant la courbe ROC et en calculant l'AUC, nous évaluons la capacité de notre modèle à discriminer efficacement entre les deux modalités, en mesurant la probabilité que les individus de la modalité cible aient des scores de probabilité plus élevés que ceux de la modalité négative.

```r
courbROC <- function(modele,donnees,varReponce){
 #utilisation de la librairie ROCR
 library(ROCR)
 #probabilit s d'affectation
 scores <- predict(modele, newdata = donnees, type = "response")
 #construire un objet prediction avec les scores et la var. cible
 pred <- ROCR::prediction(scores,varReponce)
 #objet perfromance mesur e
 graphs_roc <- ROCR::performance(pred,measure="tpr",x.measure="fpr")
 #graphique courbe ROC
 ROCR::plot(graphs_roc,xlab="taux de faux positifs",ylab="Taux de
 vrais positifs ",main = "Courbe ROC",col="darkblue")
 abline(a=0,b=1)

 #calcul de l'AUC avec la librairie ROCR - objet performance
 auc_roc <- ROCR::performance(pred,measure="auc")
 print(paste("AOC = ", auc_roc@y.values))
}
courbROC(modele.backward_BIC,data_test,data_test$Default)
```

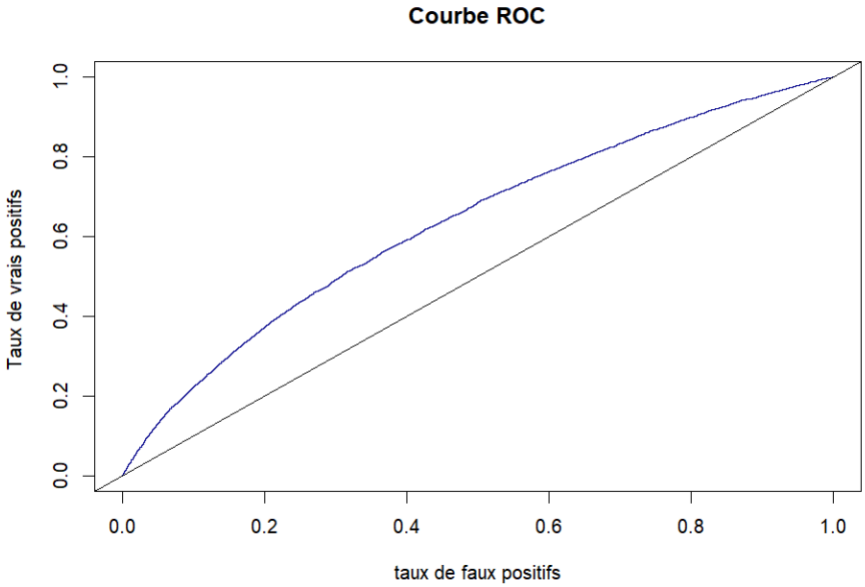

FIGURE 5 – Courbe ROC

L'aire sous la courbe est égale à 0.634, ce qui signifie que notre modèle a une précision de 63% pour prédire efficacement les individus présentant un risque de défaut de paiement. D'un point de vue statistique, cette performance est significative. Cependant, il est important de noter que l'appréciation de l'efficacité du modèle peut varier en fonction du contexte et des attentes d'un spécialiste. Nous espérons que cette précision soit suffisante pour répondre aux besoins spécifiques du domaine d'application, mais une évaluation plus approfondie par un expert du domaine pourrait être nécessaire pour confirmer son adéquation.

Après avoir construit et validé notre modèle, nous allons maintenant nous pencher sur l'analyse des variables qui le composent.

3 Analyse des variables du modèle

Dans cette section, nous allons analyser les variables présentes dans le modèle des résultats de la Figure 3. Il est important de noter qu'il existe plusieurs façons d'interpréter ces résultats. Il est fort probable que si nous avions consulté un spécialiste, il nous aurait suggéré d'explorer d'autres caractéristiques ou d'examiner nos variables sous un autre angle. Cependant, dans le cadre de cette étude, nous avons choisi de nous concentrer sur trois éléments statistiques essentiels.

Tout d'abord, nous examinerons les *p-valeurs* pour déterminer si les variables sont significatives selon le test de Wald. Ensuite, nous nous pencherons sur les *coefficients* afin de comprendre à quel moment une variable caractérise un défaut de paiement ou non. Enfin, nous utiliserons les *Old Ratios* pour évaluer le surpoids des variables dans le modèle.

Pour mesurer les *Old ratios*, nous allons souvent utiliser la fonction suivante.

```
1  old.Ratio <- function(coefficients,variables){
2    print(paste("Old ratio : ", exp(coefficients[variables,"Estimate"]) ))
3    print(paste("intervalle de confiance : ", exp(coefficients[variables,"
     Estimate"] - qnorm(0.975)*coefficients[variables,"Std. Error"]), exp(
     coefficients[variables,"Estimate"] + qnorm(0.975)*coefficients[
     variables,"Std. Error"])) )
4
5  }
```

3.1 Variables quantitatives

Âge : Même interprétation que l'exemple dans le modèle témoin (Section 2).

Income : Le p-value associé est exceptionnellement bas (2×10^{-16}), ce qui indique une très grande significativité au seuil de 5% selon le test de Wald. De plus, le coefficient est négatif (-1.867×10^{-2}), ce qui suggère que lorsque les revenus annuels de l'emprunteur sont plus élevés, le risque de défaut de paiement diminue. En d'autres termes, une augmentation des revenus annuels est associée à une réduction du risque de défaut de paiement.

Pour aller plus loin dans notre analyse, nous allons diviser notre variable de revenu en plusieurs niveaux. Le niveau 1 représente les individus dont les revenus annuels sont supérieurs au premier quartile, le niveau 2 englobe ceux dont les revenus sont supérieurs à la médiane, et enfin, le niveau 3 concerne ceux dont les revenus sont supérieurs au troisième quartile. Cette approche nous permettra de déterminer jusqu'à quel niveau de revenu il faut s'attendre pour minimiser les chances de défaut de paiement.

```
1     niv1 <- ifelse(data_train$Income >= 48892,1,0)
2     niv2 <- ifelse(data_train$Income >= 82524,1,0)
```

```
3     niv3 <- ifelse(data_train$Income >= 116053,1,0)
4     summary(glm(data_train$Default ~ niv1+niv2+niv3,data = data_
    train,family=binomial))
5
```

```
Call:
glm(formula = data_train$Default ~ niv1 + niv2 + niv3, family = binomial,
    data = data_train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.76046    0.01889 -93.176   <2e-16 ***
niv1        -0.35229    0.02866 -12.293   <2e-16 ***
niv2         0.01471    0.03035   0.485    0.628
niv3        -0.04444    0.03055  -1.455    0.146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64728  on 89560  degrees of freedom
Residual deviance: 64486  on 89557  degrees of freedom
  (1 observation effacée parce que manquante)
AIC: 64494

Number of Fisher Scoring iterations: 4
```

FIGURE 6 – Résultat par rapport des niveaux de salaire

Les résultats de la Figure 6 indiquent que les personnes qui ont un salaire supérieur au premier quartile (48 892) des montants de salaire que nous avons dans nos données ont effectivement moins de chances de se retrouver en défaut de paiement. En revanche, avoir un salaire supérieur au deuxième quartile (82 524) ne semble pas réduire de manière significative les risques de défaut de paiement par rapport à ceux qui gagnent plus que le premier quartile.

En résumé, il semble que pour réduire significativement les risques de défaut de paiement, il suffit d'avoir des revenus supérieurs au premier quartile. Avoir des revenus supérieurs au troisième quartile par rapport au deuxième quartile, ou au deuxième quartile par rapport au premier quartile, ne semble pas apporter de réduction significative supplémentaire des risques de défaut de paiement supplémentaire.

LoanAmount : De même, le p-value associé est extrêmement faible (2×10^{-16}), ce qui indique une forte significativité au seuil de 5%. En ce qui concerne le coefficient associé à la variable « LoanAmount », il est positif (2.022×10^{-6}), ce qui suggère que plus le montant d'argent emprunté est élevé, plus le risque de défaut de paiement de l'emprunteur augmente.

Pour quantifier l'ampleur de l'augmentation du risque lorsque la variable « LoanAmount » augmente d'une unité, nous utiliserons le « Old Ratio », Cela nous permettra d'évaluer de manière précise l'impact de cette variable sur le risque de défaut de paiement.

```
odele.LoanAmount <- summary(glm(data_train$Default ~ LoanAmount
,data = data_train,family=binomial))
oldRatio <- exp(modele.LoanAmount$coefficients[LoanAmount,
Estimate])
oldRatio
[1] 1.000002

```

Le « Old Ratio » associé à la variable « LoanAmount » est de 1.0. Étant donné que la relation entre les montants empruntés est linéaire, ce résultat indique que d'un point de vue statistique, chaque fois que le montant emprunté augmente d'une unité, le risque de défaut de paiement augmente d'un facteur de 1.0.

En résumé, ces résultats nous informent que non seulement le risque de défaut de paiement augmente à mesure que le montant emprunté augmente, mais également que chaque augmentation d'une unité dans le montant emprunté est associée à une augmentation du risque de défaut de paiement d'un facteur de 1.0. Cela souligne l'importance de la variable « LoanAmount » dans la prédiction du risque de défaut de paiement.

3.2 Variables qualitatives binomial

HasMortgage : Selon les résultats de la Figure 3, le p-value associé à la variable « HasMortgage » est très faible (3.05×10^{-12}), ce qui indique une forte significativité au seuil de 5%. De plus, le coefficient associé à cette variable est négatif (-1.472×10^{-1}). Cela signifie qu'au sens du test de Wald, si l'emprunteur a un prêt hypothécaire, il a statistiquement moins de chances de se retrouver en défaut de paiement.

Pour mesurer l'ampleur de cette réduction du risque, nous allons effectuer une régression logistique de la variable « Default » en fonction de la variable « HasMortgage », puis nous appliquerons la fonction « old.Ratio » pour quantifier le surpoids de risque. Cette analyse nous permettra de mieux comprendre l'impact de la détention d'un prêt hypothécaire sur le risque de défaut de paiement.

```
modele.HasMortgage <- summary(glm(Default ~ HasMortgage,data =
data_train, family = binomial))
old.Ratio(modele.HasMortgage$coefficients,HasMortgageYes)

[1] Old Ratio :  0.870688535026521
[1] intervalle de confiance :  [0.835899048291613,0.906925933910
333]

```

On voit bien que le « Old Ratio » est de 0.871, ce qui signifie que si l'emprunteur a un prêt hypothécaire, il a environ 0.129 fois moins de chances de se retrouver en défaut de paiement. Cette observation valide les résultats du modèle présentés dans la Figure 3. En d'autres

termes, la détention d'un prêt hypothécaire est un facteur significatif de réduction du risque de défaut de paiement, comme indiqué par le coefficient négatif et le faible p-value.

HasDependents : Le p-value associé à la variable « HasDependents » dans la Figure 3 est très faible (2×10^{-16}), ce qui indique une forte significativité au seuil de 5%. De plus, le coefficient associé à cette variable est négatif (-2.314×10^{-1}). Ainsi, au sens du test de Wald, si l'emprunteur a des personnes à charge (est dépendant), il a statistiquement moins de chances de se retrouver en défaut de paiement.

Pour mesurer l'ampleur de cette réduction du risque, nous allons utiliser le « Old Ratio ». Cela nous permettra de quantifier le surpoids de risque associé à la variable « HasDependents » et de mieux comprendre son impact sur le risque de défaut de paiement.

```
1     modele.HasDependents <- summary(glm(Default ~ HasDependents,data
      = data_train, family = binomial))
2     old.Ratio(modele.HasDependents$coefficients,HasDependentsYes)
3
4     [1] Old Ratio :  0.798549323602168
5     [1] intervalle de confiance :  [0.766544709050987,0.831890188133
     912]
6
```

Le « Old Ratio » est de 0.798, ce qui signifie que les emprunteurs ayant des personnes à charge (sont dépendants) ont environ 0.798 fois plus de chances d'être en défaut de paiement que les emprunteurs sans personnes à charge (sont indépendants). Cette observation confirme les résultats du modèle présentés dans la Figure 3.

Maintenant, après avoir examiné l'impact individuel des variables telles que le prêt hypothécaire et les personnes à charge, nous pouvons nous demander s'il existe une interaction entre ces deux variables. Cela signifierait que l'effet d'une variable sur le risque de défaut de paiement dépend de la valeur de l'autre variable. Pour explorer cette interaction, nous devrons effectuer une analyse spécifique afin d'évaluer comment la combinaison de ces deux variables peut influencer le risque de défaut de paiement.

HasDependents et HasMortgage

```
1     modele.HasDependentsHasMortgage <- summary(glm(Default ~
     HasDependents*HasMortgage,data = data_train, family = binomial))
2     modele.HasDependentsHasMortgage
3
```

```
Call:
glm(formula = Default ~ HasDependents * HasMortgage, family = binomial,
    data = data_train)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.86405    0.01963 -94.983  < 2e-16 ***
HasDependentsYes              -0.18091    0.02877  -6.288 3.21e-10 ***
HasMortgageYes                -0.09638    0.02819  -3.418  0.00063 ***
HasDependentsYes:HasMortgageYes -0.09348    0.04182  -2.235  0.02540 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64728  on 89560  degrees of freedom
Residual deviance: 64562  on 89557  degrees of freedom
  (1 observation effacée parce que manquante)
AIC: 64570

Number of Fisher Scoring iterations: 4
```

On constate que le p-value associé à la conjonction des individus ayant un prêt hypothécaire et des individus dépendant est de 0.025, ce qui est significatif au seuil de 5%. De plus, le coefficient est négatif (-0.093). Par conséquent, au sens du test de Wald, nous pouvons conclure que non seulement les emprunteurs ayant un prêt hypothécaire et les emprunteurs ayant des personnes à charge ont moins de chances d'être en défaut de paiement, mais également que la combinaison de ces deux variables diminue encore davantage le risque de défaut de paiement.

3.3 Variables qualitatives non binomial

EmploymentType : Selon les résultats de la Figure 3, par rapport aux emprunteurs travaillant à leur compte (travailleurs indépendants), les emprunteurs travaillant à temps plein, à temps partiel et les sans emploi ont statistiquement plus de chances de se retrouver en défaut de paiement. Par conséquent, en se basant sur ces résultats statistiques, il semblerait qu'entre deux individus ayant des caractéristiques égales pour les autres variables, il serait plus judicieux de prêter de l'argent à un travailleur indépendant plutôt qu'à un individu ayant un autre type de contrat.

Maintenant, regardons quel type de contrat présente le plus de risque de défaut de paiement par rapport aux individus qui travaillent à leur compte. Pour cela, nous allons à nouveau utiliser le « Old Ratio » pour quantifier cette différence de risque.

```
1    modele.EmploymentType <- summary(glm(Default ~ EmploymentType,
    data = data_train, family=binomial))
2    apply(modele.EmploymentType$coefficients,1,function(x){exp(x[
    Estimate])})
```

```
3
```

```
      (Intercept)     EmploymentTypePart-time EmploymentTypeSelf-employed    EmploymentTypeUnemployed
        0.1013325                   1.4010861                   1.2774754                   1.5736294
```

Il est effectivement observé que, comme on pouvait s'y attendre, par rapport aux travailleurs indépendants, les individus ayant plus de chances d'être en défaut de paiement sont dans l'ordre, les individus sans emploi, avec un Old Ratio de 1.574, les individus travaillant à temps partiel, avec un Old Ratio de 1.401 et en fin, les individus travaillant à temps plein.

En résumé, ces résultats suggèrent que lorsque deux individus présentent des caractéristiques équivalentes pour les autres variables, l'individu ayant le plus de chances d'être en défaut de paiement est successivement celui qui n'a pas d'emploi, suivi de l'individu travaillant à temps partiel, puis de l'individu travaillant à temps plein, et enfin de l'individu travaillant à son compte. Cette information est essentielle pour évaluer le risque de défaut de paiement en fonction du type d'emploi de l'emprunteur.

MaritalStatus : D'après la Figure 3, par rapport aux emprunteurs en situation de divorce, les emprunteurs mariés ou célibataires ont statistiquement moins de chances de se retrouver en défaut de paiement. Pour quantifier cette différence de risque, nous allons utiliser les « Old Ratios ». Cela nous permettra de mesurer l'ampleur de la variation du risque de défaut de paiement entre ces différents groupes d'emprunteurs.

```
1     modele.MaritalStatus <- summary(glm(Default ~ MaritalStatus,
    data = data_train, family=binomial))
2     apply(modele.MaritalStatus$coefficients,1,function(x){exp(x[
    Estimate])})
3
```

```
      (Intercept) MaritalStatusMarried  MaritalStatusSingle
        0.1460477           0.7949448           0.9353092
```

On peut constater que par rapport aux individus en situation de divorce, les individus ayant moins de chances d'être en défaut de paiement sont, dans l'ordre décroissant de risque, les personnes mariées, avec un « Old Ratio » de 0.795, les célibataires, avec un « Old Ratio » de 0.957.

En résumé, d'un point de vue statistique, en supposant que toutes les autres variables sont identiques pour plusieurs individus, et en se basant uniquement sur leur statut marital pour prédire leur risque de défaut de paiement, on peut conclure que les emprunteurs ayant le plus de chances d'être en défaut de paiement sont ceux en situation de divorce, suivis des célibataires, et enfin des personnes mariées. Il est toutefois important de noter que ces conclusions sont basées sur des analyses statistiques et qu'il est recommandé de consulter des spécialistes pour valider ou affiner ces conclusions.

4 Conclusion

En conclusion, afin de minimiser le risque de défaut de paiement et garantir que les emprunteurs remboursent leurs prêts conformément aux termes convenus, les services financiers font largement appel à des méthodes d'apprentissage automatique. Comme nous l'avons précédemment souligné, ces méthodes permettent non seulement d'analyser et de prédire le risque de défaut de paiement d'un emprunteur, mais elles permettent également d'étudier les variables pour comprendre leur impact et leur contribution au risque de défaut de paiement. Il est important de noter que ces modèles ne sont pas destinés à remplacer les experts, mais plutôt à les accompagner et à faciliter leurs expériences et leurs prises de décision.

Références

[1] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[2] Ricco Rakotomalala. Pratique de la régression logistique. *Régression logistique binaire et polytomique, Université Lumière Lyon*, 2 :258p, 2011.

[3] Laurent Rouvière. Régression logistique avec r. *Universités Rennes, 2*, 2015.