DSCI300 Mini Project 7

Nisi Mohan Kuniyil 300321388

05/12/2020

Problem Statement

Logistic Regression analysis and prediction of credit card approvals

Companies have to go through a series of processes to approve a credit card to a person. This process is tedious and mundane. Every time an application is submitted bank has to analyze certain factors that play a vital role in the approval of the credit card, such as the income of the applicant, credit score, employment status, etc. This can be automated using Logistic regression analysis which can be used to understand the factors which have the most effect on the decision-making process of credit card approval. In order to achieve this, a logistic regression model can be trained to predict the probability of credit card approval based on the features from the data set. afterward, the trained logistic regression model can be analyzed to get insights into different features that have the highest impact on the approval rate.

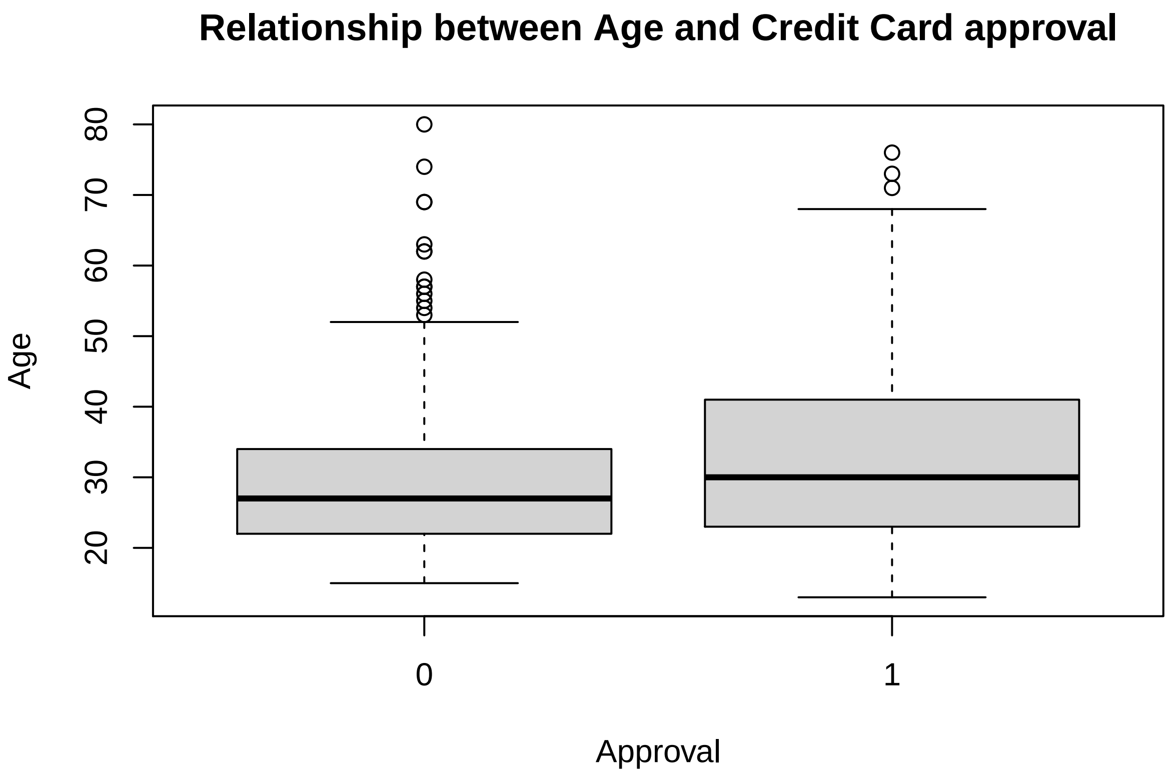
Solution

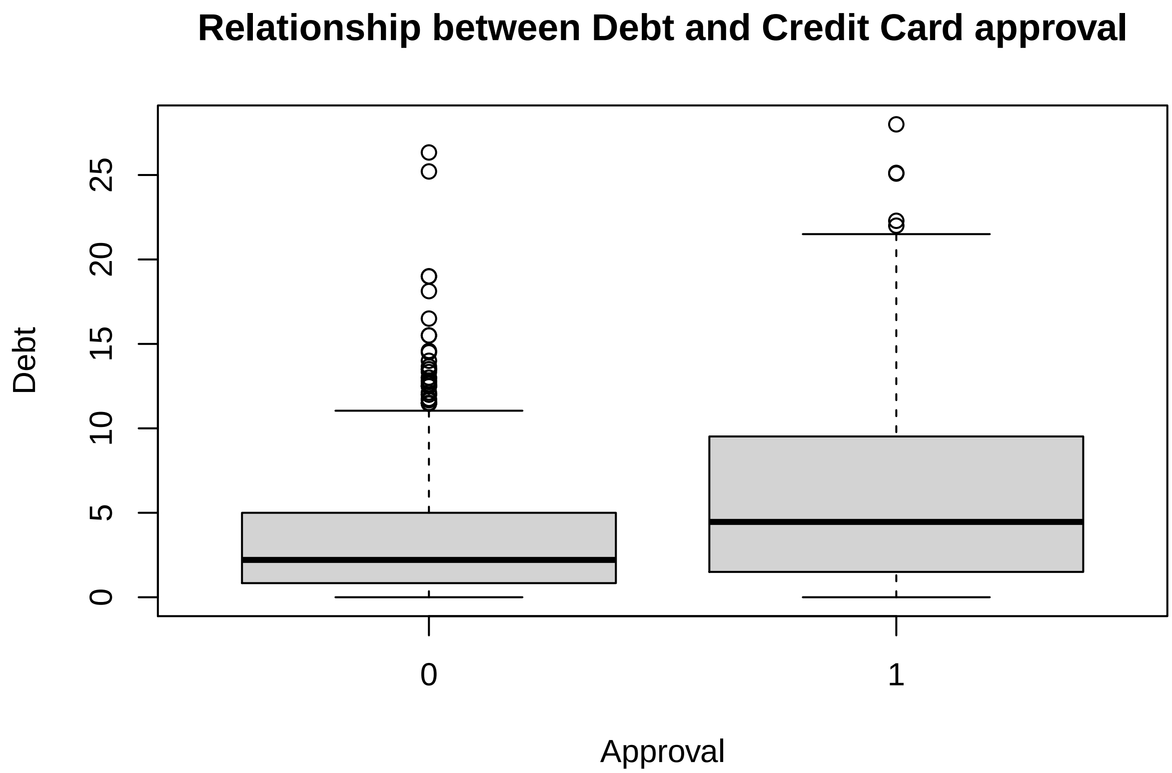
Exploratory Data Analysis

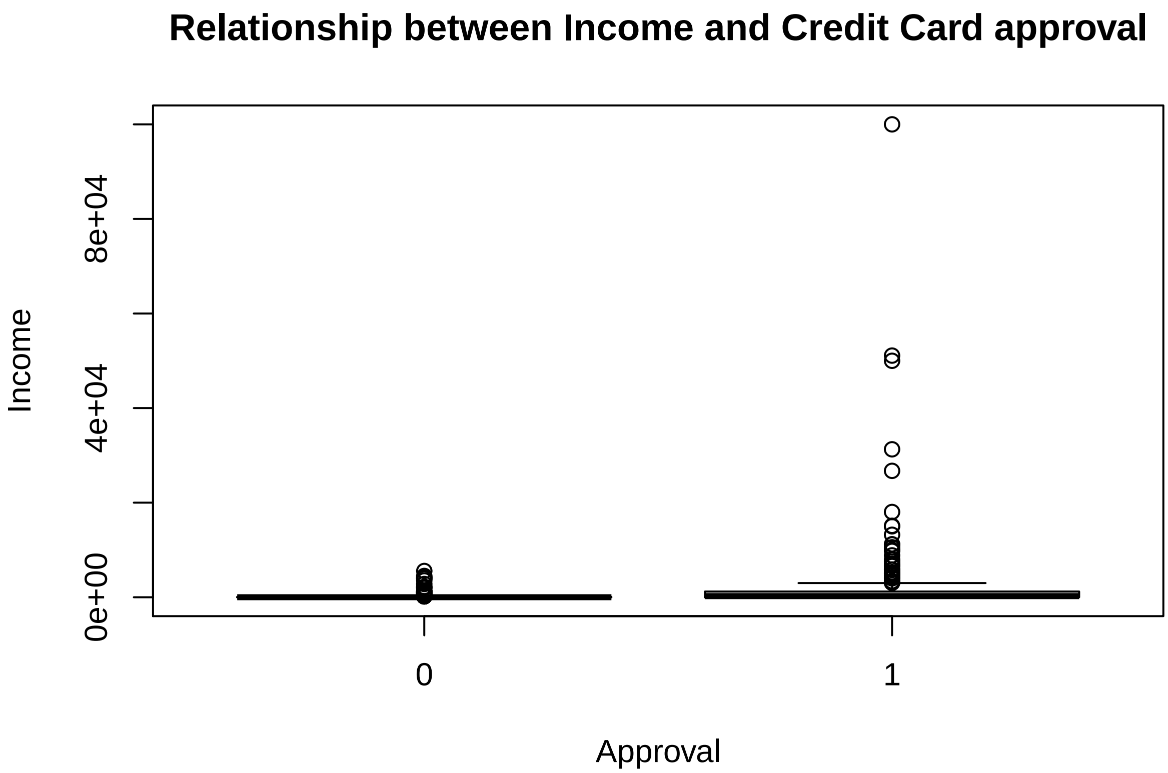
We start off by understanding the type of data in the dataframe. We can see from the below summary that there are 15 variables associated with credit card approval or denial. The outcome values of the last column *approved* are the following symbols, "+" means approved and "-" means denied. These symbols are not meaningful, so we will be transforming that to 1's and 0's for the regression analysis.

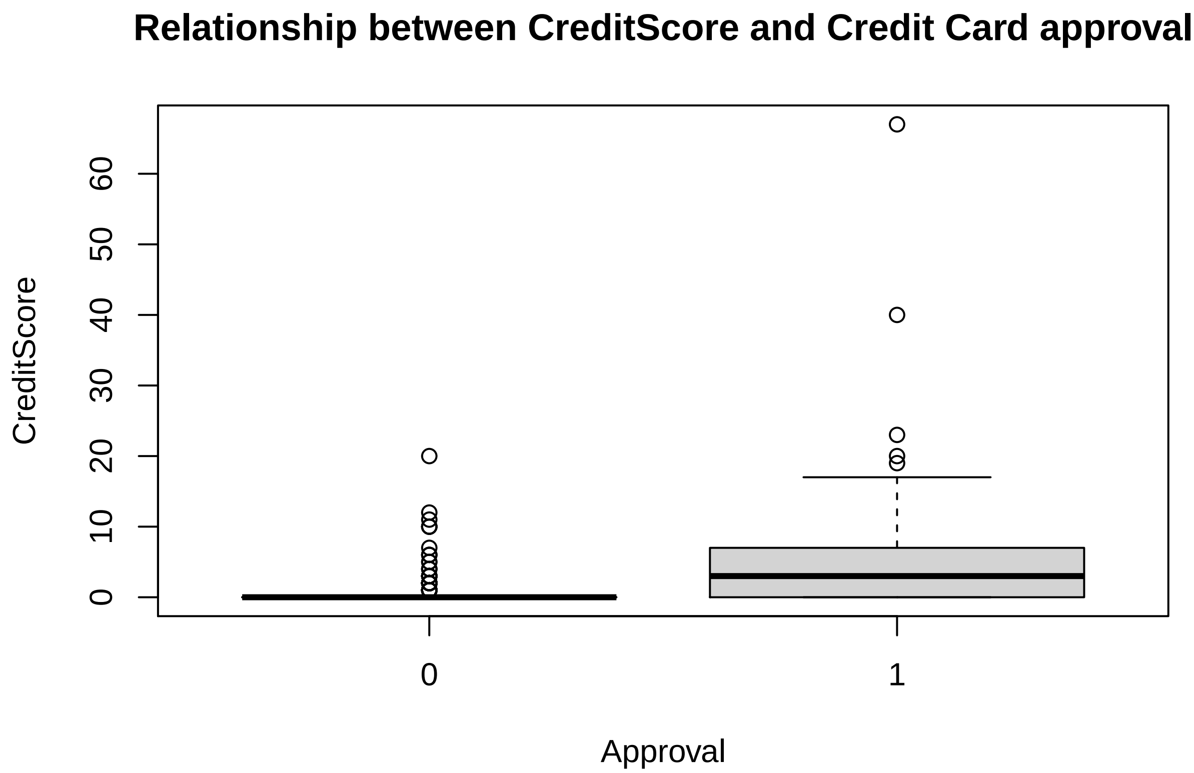
```
## 'data.frame':    689 obs. of  16 variables:
##  $ Male          : chr  "a" "a" "b" "b" ...
##  $ Age           : chr  "58.67" "24.50" "27.83" "20.17" ...
##  $ Debt          : num  4.46 0.5 1.54 5.62 4 ...
##  $ Married       : chr  "u" "u" "u" "u" ...
##  $ BankCustomer  : chr  "g" "g" "g" "g" ...
##  $ EducationLevel: chr  "q" "q" "w" "w" ...
##  $ Ethnicity     : chr  "h" "h" "v" "v" ...
##  $ YearsEmployed : num  3.04 1.5 3.75 1.71 2.5 ...
##  $ PriorDefault  : chr  "t" "t" "t" "t" ...
##  $ Employed      : chr  "t" "f" "t" "f" ...
##  $ CreditScore   : int  6 0 5 0 0 0 0 0 0 0 ...
##  $ DriversLicense: chr  "f" "f" "t" "f" ...
##  $ Citizen       : chr  "g" "g" "g" "s" ...
##  $ ZipCode       : chr  "00043" "00280" "00100" "00120" ...
##  $ Income        : int  560 824 3 0 0 31285 1349 314 1442 0 ...
##  $ Approved      : chr  "+" "+" "+" "+" ...
```

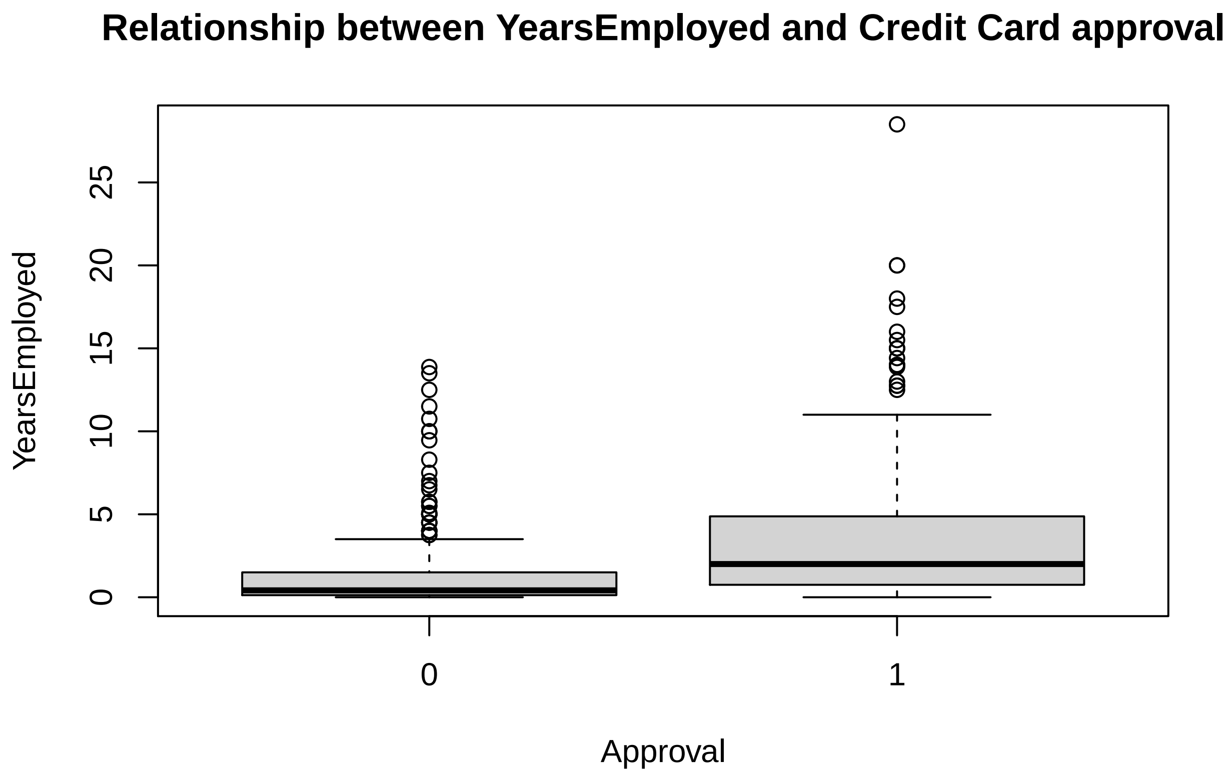
There are five continuous variables in the dataset. We will check the relationship between these variables and credit card approval before jumping to regression analysis. Box plot is used here to understand the correlation between *Age*, *Income*, *Debt*, *CreditScore*, *YearsEmployed*, and the approval rate. The below box plots of these continuous variables show that the means of the features of the approved applications are further distributed from the mean of the denied.




Relationship between Debt and Credit Card approval
Debt
0
5
10
15
20
25
0
1
Approval


Relationship between Income and Credit Card approval
Income
0e+00
4e+04
8e+04
0
1
Approval


Relationship between CreditScore and Credit Card approval
CreditScore
0
10
20
30
40
50
60
0
1
Approval



Modeling

The next step is to perform logistic regression on the five variables identified from the dataset. The Akaike Information Criterion(AIC) value tells us the quality of our model. Summary of the regression can be used to interpret the factors that have a significant influence on the approval of a credit card application.

To effectively evaluate the regression model trained, the dataset needs to be partitioned as train and test data. 75% of the dataset is used for training and the rest is used to predict the credit card application approval. The confusion matrix tells us how accurate the prediction is.

```
##
## Call:
## glm(formula = Approved ~ Age + Debt + YearsEmployed + CreditScore +
##     Income, family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7205  -0.7503  -0.6373   0.7009   1.8657
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.6310224  0.3263787  -4.997 5.81e-07 ***
## Age            0.0046683  0.0097881   0.477 0.633406
## Debt           0.0096415  0.0243884   0.395 0.692598
## YearsEmployed  0.2132374  0.0493099   4.324 1.53e-05 ***
## CreditScore    0.3730668  0.0545419   6.840 7.92e-12 ***
```

```
## Income         0.0005563  0.0001551   3.587 0.000335 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.76  on 507  degrees of freedom
## Residual deviance: 500.26  on 502  degrees of freedom
## AIC: 512.26
##
## Number of Fisher Scoring iterations: 7
```

We develop a multiple regression equation using *Age*, *Debt*, *YearsEmployed*, *CreditScore*, and *Income* to predict credit card approval and check how well the regression model explains the variability in credit card approval.

$Logoddsof Approval = 0.0046683Age + 0.0096415Debt + 0.2132374YearsEmployed + 0.3730668CreditScore + 0.0005563Income - 1.6310224$

From the summary of the regression model, we can see that *YearsEmployed*, *CreditScore*, and *Income* has a high significance in predicting the credit card application approval or denial. These factors are significant at α 0.001. Other features like *Age* and *Debt* does not seem to have much significance in predicting the approval of credit card applications.

Furthermore, deviance in the summary is a measure of goodness of fit of the regression model. The null deviance of this model is 700.76 on 507 degrees of freedom. Null deviance includes just the intercept of the model and shows how well the model is predicted, whereas residual deviance includes predictors. For this model, residual deviance is 500.26 on 502 degrees of freedom, which is far less than the null deviance. So we could say that the goodness of fit is higher when we include the predictors in the regression model.

	FALSE	TRUE
0	84	21
1	14	50

The confusion matrix above gives the actual values and predicted values. 84 is the number of credit card applications correctly predicted as denied out of 98 (85.71% accuracy) and 50 is the number of applications correctly predicted as granted out of 71 (71.42% accuracy). Rest are Type 1 and Type 2 errors in the prediction. Approximately, the model is 79% accurate in predicting credit card approval.

From this model, we found that only three factors have a significant influence on predicting approval rate. So in the next step, we will perform another regression model by removing the least significant features such as Age and Debt from the model and analyze the difference in **AIC** and other factors.

```
##
## Call:
## glm(formula = Approved ~ YearsEmployed + CreditScore + Income,
##     family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7273  -0.7479  -0.6456   0.7173   1.8282
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.4628302  0.1552073  -9.425  < 2e-16 ***
## YearsEmployed  0.2210958  0.0477067   4.634 3.58e-06 ***
```

```
## CreditScore    0.3731183  0.0542168   6.882 5.90e-12 ***
## Income         0.0005600  0.0001556   3.600 0.000318 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.76  on 507  degrees of freedom
## Residual deviance: 500.67  on 504  degrees of freedom
## AIC: 508.67
##
## Number of Fisher Scoring iterations: 7
```

Above is the simplified model by removing Age and Debt from our previous model. It can be seen that from the summary of regression the AIC has reduced to 508.67, which was earlier 512.26. The AIC shows the quality of a model, the lower the AIC better the model is. So from this, we can say that the simplified model is much better than the first model. Moreover, there is not much change in the deviance when compared to the previous complex model.

	FALSE	TRUE
0	84	21
1	14	50

The confusion matrix above also shows there is no change in the prediction of credit card approval after simplifying the model. This shows that the predictability of the model stays still 79% even after removing less influencing factors such as *Age* and *Debt*.

Conclusion

From the analysis, we found out that the variables that have a major impact on the variances in credit card approval are *Income*, *YearsEmployed*, and *CreditScore.* The AIC of the model is low, with 508.67. This model is significant at $\alpha = 0.001$ or 99.9% level.

Appendix 1: The Problem

Logistic Regression analysis and prediction of credit card approvals

Companies have to go through a series of processes to approve a credit card to a person. This process is tedious and mundane. Every time an application is submitted bank has to analyze certain factors that play a vital role in the approval of the credit card, such as the income of the applicant, credit score, employment status, etc. This can be automated using Logistic regression analysis which can be used to understand the factors which have the most effect on the decision-making process of credit card approval. In order to achieve this, a logistic regression model can be trained to predict the probability of credit card approval based on the features from the data set. afterward, the trained logistic regression model can be analyzed to get insights into different features that have the highest impact on the approval rate.

Managerial Report

1. The datset contains meaningless variable names in order to protect the privacy of the individuals included in the study. These variable names need to be changed to meaningful names. The article cited in the Reference is used to interpret the meaning of each variables. Also, the outcome values of Approved column is in character symbols, these need to be changed to 1's and 0's in order to use regression.
2. To understand the outcome values in each column, take the structure of the dataset. After finding out the continuous variables, visualize them using box plots to understand the relationship between these features and Approved factor in the dataset.
3. Train a logistic regression model that can be used to predict credit card applications approval given the *Age*, *Debt*, *YearsEmployed*, *CreditScore*, and *Income*. Test for individual significance and discuss your findings and conclusion.
4. From the regression model above, if any variable does not have much significance remove those and predict card approval rate by using the rest of the variables. Based on the results of your analysis, which regression model would you recommend to predict credit card approval? Provide an interpretation of the summary of the logistic regression.

Appendix 2: Analysis

Reading the data.

```
creditcard_df <- read.csv("crx.csv")
head(creditcard_df)
```

```
##   b X30.83    X0 u g w v X1.25 t t.1 X01 f g.1 X00202  X0.1 X.
## 1 a  58.67 4.460 u g q h  3.04 t   t   6 f   g  00043   560  +
## 2 a  24.50 0.500 u g q h  1.50 t   f   0 f   g  00280   824  +
## 3 b  27.83 1.540 u g w v  3.75 t   t   5 t   g  00100     3  +
## 4 b  20.17 5.625 u g w v  1.71 t   f   0 f   s  00120     0  +
## 5 b  32.08 4.000 u g m v  2.50 t   f   0 t   g  00360     0  +
## 6 b  33.17 1.040 u g r h  6.50 t   f   0 t   g  00164 31285  +
```

Question 1:

The datset contains meaningless variable names in order to protect the privacy. These variables names needs to be changed to meaningful names. The article cited in the Reference is used to interpret the meaning of each variables.

Changing the variable names to meaningful names based on the article cited in the appendix.

```
creditCard_df <- rename_(creditcard_df, "Male" = "b", "Age" ="X30.83", "Debt"= "X0", "Married"="u", "Ba
", "CreditScore"= "X01", "DriversLicense"= "f", "Citizen"= "g.1
", "ZipCode"="X00202", "Income" = "X0.1", "Approved"= "X.")


str(creditCard_df)
```

```
## 'data.frame':    689 obs. of  16 variables:
##  $ Male          : chr  "a" "a" "b" "b" ...
##  $ Age           : chr  "58.67" "24.50" "27.83" "20.17" ...
##  $ Debt          : num  4.46 0.5 1.54 5.62 4 ...
##  $ Married       : chr  "u" "u" "u" "u" ...
##  $ BankCustomer  : chr  "g" "g" "g" "g" ...
##  $ EducationLevel: chr  "q" "q" "w" "w" ...
##  $ Ethnicity     : chr  "h" "h" "v" "v" ...
##  $ YearsEmployed : num  3.04 1.5 3.75 1.71 2.5 ...
##  $ PriorDefault  : chr  "t" "t" "t" "t" ...
##  $ Employed      : chr  "t" "f" "t" "f" ...
##  $ CreditScore   : int  6 0 5 0 0 0 0 0 0 0 ...
##  $ DriversLicense: chr  "f" "f" "t" "f" ...
##  $ Citizen       : chr  "g" "g" "g" "s" ...
##  $ ZipCode       : chr  "00043" "00280" "00100" "00120" ...
##  $ Income        : int  560 824 3 0 0 31285 1349 314 1442 0 ...
##  $ Approved      : chr  "+" "+" "+" "+" ...
```

```
creditCard_df$Approved <- as.integer(factor(creditCard_df$Approved))-1
creditCard_df$Age <- as.integer(creditCard_df$Age)
head(creditCard_df)
```

```
##   Male Age  Debt Married BankCustomer EducationLevel Ethnicity YearsEmployed
## 1    a  58 4.460       u            g              q         h          3.04
## 2    a  24 0.500       u            g              q         h          1.50
## 3    b  27 1.540       u            g              w         v          3.75
```

```
## 4     b  20 5.625       u            g             w         v          1.71
## 5     b  32 4.000       u            g             m         v          2.50
## 6     b  33 1.040       u            g             r         h          6.50
##   PriorDefault Employed CreditScore DriversLicense Citizen ZipCode Income
## 1            t        t           6              f       g   00043    560
## 2            t        f           0              f       g   00280    824
## 3            t        t           5              t       g   00100      3
## 4            t        f           0              f       s   00120      0
## 5            t        f           0              t       g   00360      0
## 6            t        f           0              t       g   00164  31285
##   Approved
## 1        1
## 2        1
## 3        1
## 4        1
## 5        1
## 6        1
```

There are only 12 NA's in the dataset. So just removing that.

```r
creditCard_df<-  na.omit(creditCard_df)
which(is.na(creditCard_df$Age))
```

```
## integer(0)
```

Question2:

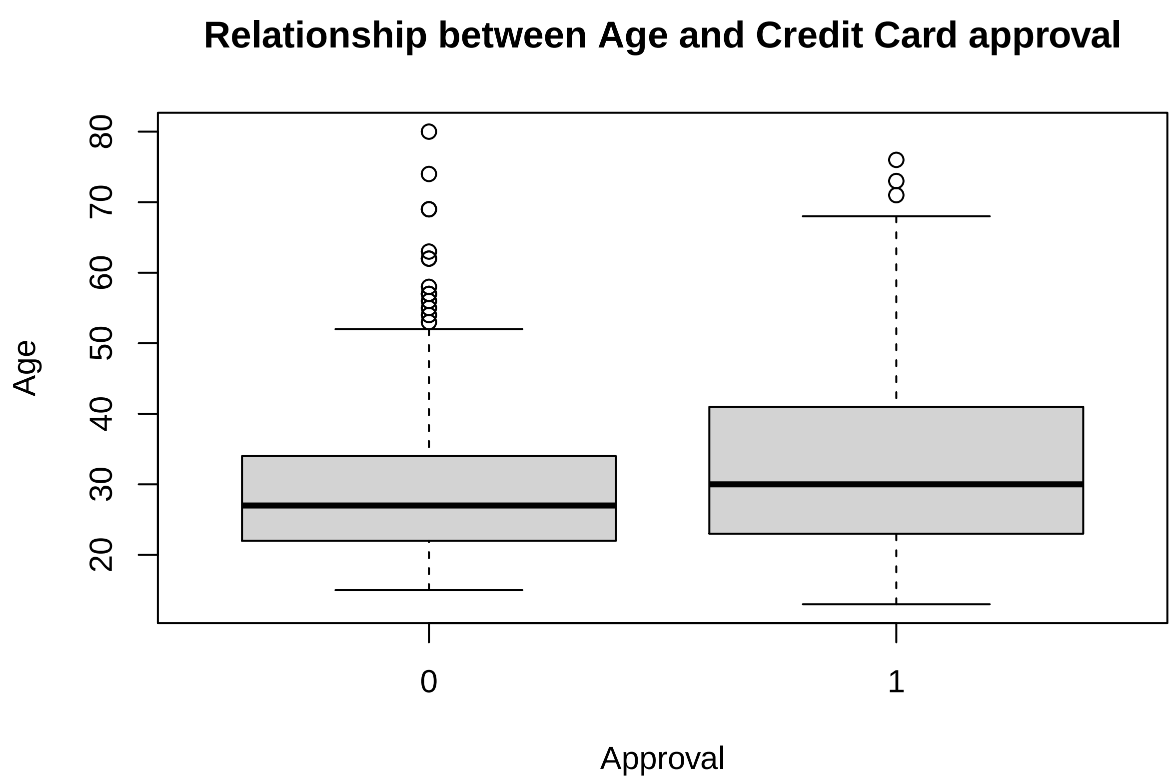
To understand the outcome values in each column, take the structure of the dataset. After finding out the continuous variables, visualize these with box plots to understand the relationship between these features and Approved factor in the dataset.

```r
summary(creditCard_df)
```

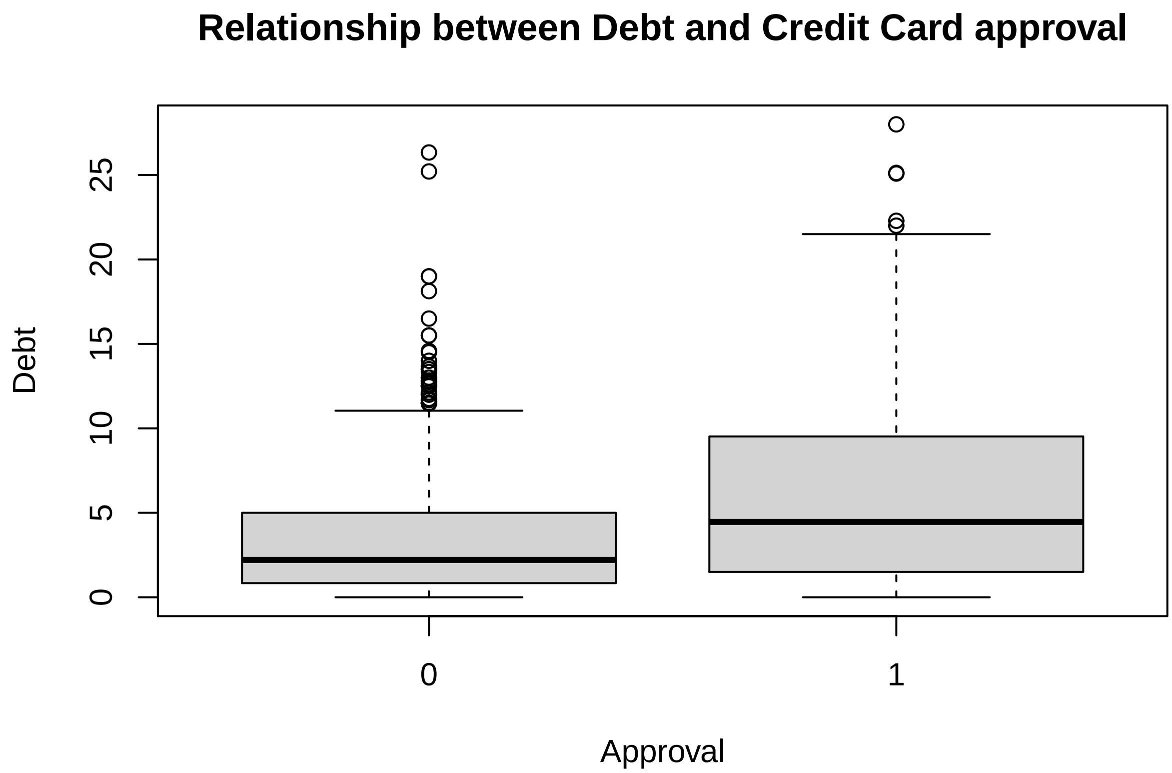
```
##      Male                Age             Debt           Married
##  Length:677         Min.   :13.00   Min.   : 0.000   Length:677
##  Class :character   1st Qu.:22.00   1st Qu.: 1.000   Class :character
##  Mode  :character   Median :28.00   Median : 2.750   Mode  :character
##                     Mean   :31.12   Mean   : 4.785
##                     3rd Qu.:38.00   3rd Qu.: 7.500
##                     Max.   :80.00   Max.   :28.000
##  BankCustomer       EducationLevel      Ethnicity         YearsEmployed
##  Length:677         Length:677         Length:677         Min.   : 0.000
##  Class :character   Class :character   Class :character   1st Qu.: 0.165
##  Mode  :character   Mode  :character   Mode  :character   Median : 1.000
##                                                           Mean   : 2.211
##                                                           3rd Qu.: 2.585
##                                                           Max.   :28.500
##  PriorDefault         Employed          CreditScore     DriversLicense
##  Length:677         Length:677         Min.   : 0.000   Length:677
##  Class :character   Class :character   1st Qu.: 0.000   Class :character
##  Mode  :character   Mode  :character   Median : 0.000   Mode  :character
##                                        Mean   : 2.437
##                                        3rd Qu.: 3.000
##                                        Max.   :67.000
##    Citizen            ZipCode              Income          Approved
```

```
##  Length:677          Length:677          Min.    :      0   Min.    :0.000
##  Class :character    Class :character    1st Qu.:      0   1st Qu.:0.000
##  Mode  :character    Mode  :character    Median :      5   Median :0.000
##                                          Mean   :   1023   Mean    :0.449
##                                          3rd Qu.:    396   3rd Qu.:1.000
##                                          Max.   :100000    Max.    :1.000
```

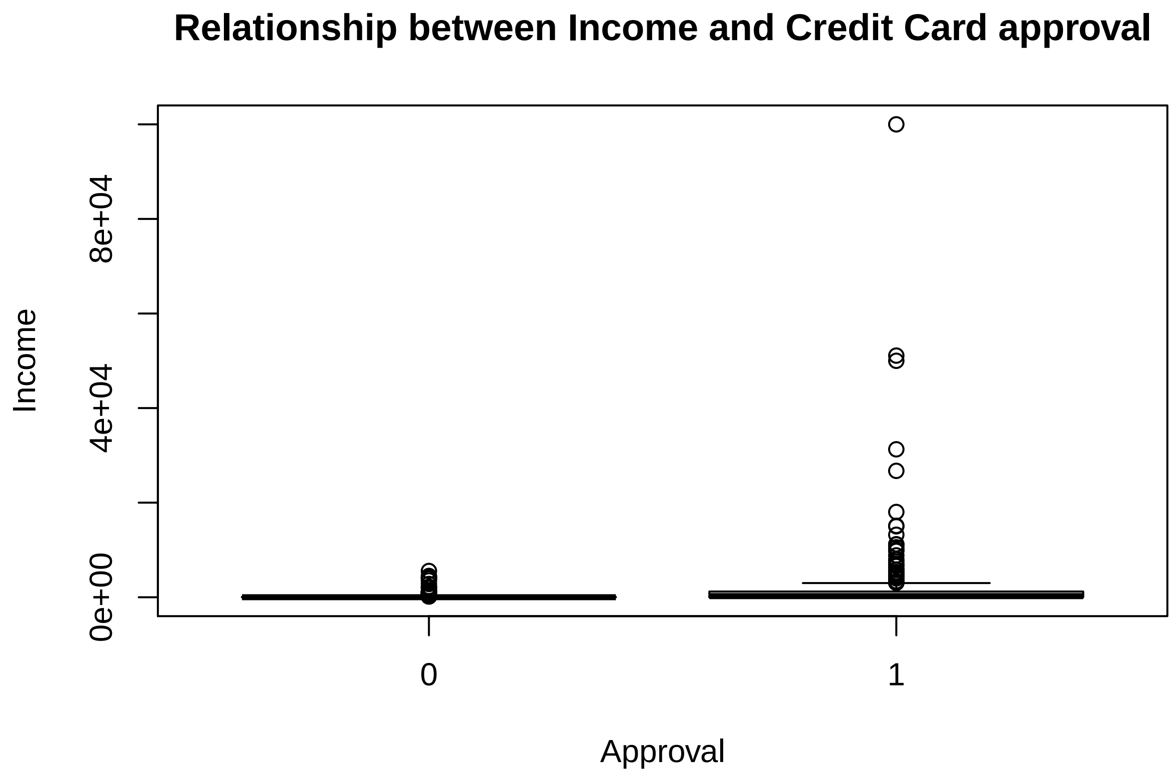
```
boxplot(creditCard_df$Age[creditCard_df$Approved==0],
        creditCard_df$Age[creditCard_df$Approved==1],
        names= c(0,1),
        main = "Relationship between Age and Credit Card approval",
        xlab = "Approval",
        ylab = "Age"
        )
```



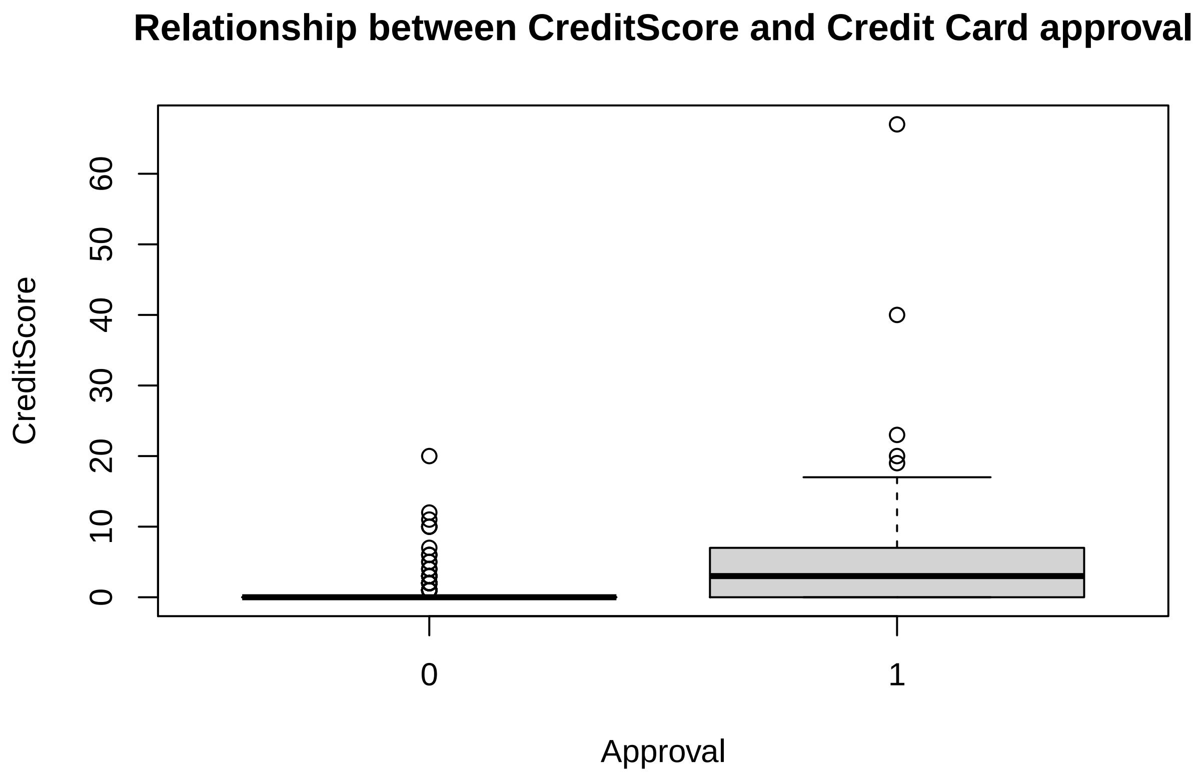

```
boxplot(creditCard_df$Debt[creditCard_df$Approved==0],
        creditCard_df$Debt[creditCard_df$Approved==1],
        names= c(0,1),
        main = "Relationship between Debt and Credit Card approval",
        xlab = "Approval",
        ylab = "Debt"
        )
```



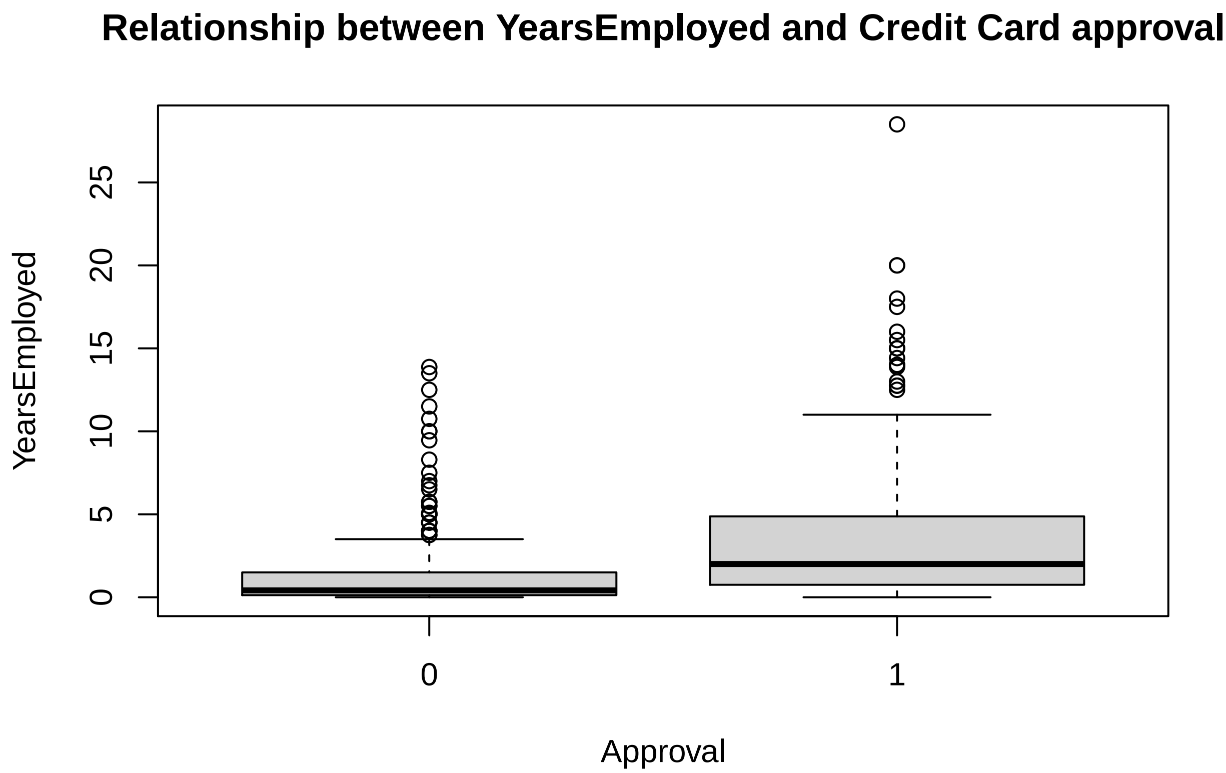

```r
boxplot(creditCard_df$Income[creditCard_df$Approved==0],
        creditCard_df$Income[creditCard_df$Approved==1],
        names= c(0,1),
        main = "Relationship between Income and Credit Card approval",
        xlab = "Approval",
        ylab = "Income"
        )
```




```
boxplot(creditCard_df$CreditScore[creditCard_df$Approved==0],
        creditCard_df$CreditScore[creditCard_df$Approved==1],
        names= c(0,1),
        main = "Relationship between CreditScore and Credit Card approval",
        xlab = "Approval",
        ylab = "CreditScore"
        )
```




```
boxplot(creditCard_df$YearsEmployed[creditCard_df$Approved==0],
        creditCard_df$YearsEmployed[creditCard_df$Approved==1],
        names= c(0,1),
        main = "Relationship between YearsEmployed and Credit Card approval",
        xlab = "Approval",
        ylab = "YearsEmployed"
        )
```



Question 3:

Develop a logistic regression model that can be used to predict credit card applications approval given the *Age*, *Debt*, *YearsEmployed*, *CreditScore*, and *Income*. Test for individual significance and discuss your findings and conclusion.

```
set.seed(42)
TrainingIndex <- createDataPartition(y=creditCard_df$Approved, p=0.75, list=FALSE)
training <- creditCard_df[TrainingIndex,]
testing <- creditCard_df[-TrainingIndex,]
```

```
cred_Lg<- glm(formula =  Approved ~ Age + Debt + YearsEmployed +
    CreditScore + Income, family = binomial, data = training)


summary(cred_Lg)
```

```
##
## Call:
## glm(formula = Approved ~ Age + Debt + YearsEmployed + CreditScore +
##     Income, family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7205  -0.7503  -0.6373   0.7009   1.8657
##
```

```
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.6310224  0.3263787  -4.997 5.81e-07 ***
## Age            0.0046683  0.0097881   0.477 0.633406
## Debt           0.0096415  0.0243884   0.395 0.692598
## YearsEmployed  0.2132374  0.0493099   4.324 1.53e-05 ***
## CreditScore    0.3730668  0.0545419   6.840 7.92e-12 ***
## Income         0.0005563  0.0001551   3.587 0.000335 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.76  on 507  degrees of freedom
## Residual deviance: 500.26  on 502  degrees of freedom
## AIC: 512.26
##
## Number of Fisher Scoring iterations: 7
```

```
CredCardPredict <- predict(cred_Lg, newdata = testing, type="response")
glm.pred <- ifelse(CredCardPredict > 0.5, 1, 0)

confusionMatrix(as.factor(glm.pred), as.factor(testing$Approved))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 84 21
##          1 14 50
##
##                Accuracy : 0.7929
##                  95% CI : (0.7239, 0.8513)
##     No Information Rate : 0.5799
##     P-Value [Acc > NIR] : 4.073e-09
##
##                   Kappa : 0.5691
##
##  Mcnemar's Test P-Value : 0.3105
##
##             Sensitivity : 0.8571
##             Specificity : 0.7042
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.7812
##              Prevalence : 0.5799
##          Detection Rate : 0.4970
##    Detection Prevalence : 0.6213
##       Balanced Accuracy : 0.7807
##
##        'Positive' Class : 0
##
```

Question4:

From the regression model above, if any variable does not have much significance remove those and predict card approval rate by using the rest of the variables. Based on the results of your analysis, which regression model would you recommend to predict credit card approval? Provide an interpretation of the summary of the logistic regression.

```
set.seed(42)
cred_Lg1<- glm(formula =  Approved ~ YearsEmployed +
    CreditScore + Income, family = binomial, data = training)


summary(cred_Lg1)
```

```
##
## Call:
## glm(formula = Approved ~ YearsEmployed + CreditScore + Income,
##     family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7273  -0.7479  -0.6456   0.7173   1.8282
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.4628302  0.1552073  -9.425  < 2e-16 ***
## YearsEmployed  0.2210958  0.0477067   4.634 3.58e-06 ***
## CreditScore    0.3731183  0.0542168   6.882 5.90e-12 ***
## Income         0.0005600  0.0001556   3.600 0.000318 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.76  on 507  degrees of freedom
## Residual deviance: 500.67  on 504  degrees of freedom
## AIC: 508.67
##
## Number of Fisher Scoring iterations: 7
```

```
CredCardPredict_new <- predict(cred_Lg1, newdata = testing, type="response")
glm.predNew <- ifelse(CredCardPredict_new > 0.5, 1, 0)

confusionMatrix(as.factor(glm.predNew), as.factor(testing$Approved))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 84 21
##          1 14 50
##
##                Accuracy : 0.7929
##                  95% CI : (0.7239, 0.8513)
##     No Information Rate : 0.5799
##     P-Value [Acc > NIR] : 4.073e-09
```

```
##
##                Kappa : 0.5691
##
##  Mcnemar's Test P-Value : 0.3105
##
##            Sensitivity : 0.8571
##            Specificity : 0.7042
##         Pos Pred Value : 0.8000
##         Neg Pred Value : 0.7812
##             Prevalence : 0.5799
##         Detection Rate : 0.4970
##   Detection Prevalence : 0.6213
##      Balanced Accuracy : 0.7807
##
##       'Positive' Class : 0
##
```

Appendix 3: Data Source and References

Dataset is taken from the http://archive.ics.uci.edu/ml/datasets/credit+approval site. The data description of the dataset is taken from the article https://nycdatascience.com/blog/student-works/credit-card-approval-analysis/.

Reference

http://archive.ics.uci.edu/ml/datasets/credit+approval

https://nycdatascience.com/blog/student-works/credit-card-approval-analysis/.