****

**Loan Prediction**

**ALY 6015 21037 Intermediate analytics**

**Customer Loan Prediction**

**Sec 03 Winter 2018 CPS [[BOS-A-BL]](https://northeastern.blackboard.com/webapps/blackboard/execute/courseMain?course_id=_2515849_1)**

**Submitted by:**

Abhimanyu kumar Singh

Monisha Patel

Prashant Singh

Shrinath Patel

**Submitted to:**

Ani Aghababyan

**Northeastern University**

**February, 2018**

**Abstract**

In this Paper we have described a very effective way for customer loan prediction. Our main interest is to decide whether a customer will get the loan approved or not based on several factors.Basically we are trying to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. We have applied Logistics regression and random forest here. Logistics regression is giving us the probability whether a customer should get loan or not and random forest also performing the same purpose. Depending on the accuracy of these models we will select which one will best fit our data.

**Keywords:- GLM- Generalized Linear Model, RM- Random Forest**

**1. Introduction**

From the past decade, for the extraction and manipulation of the data, data mining has become very efficient in order to make some patterns and to take effective decision. To extract the right information from the data its not only the ability of data mining techniques but it's also depend on the ability of analyst.

In 1997, Berry proposed that the there are six different data mining phase for any human problem and they are : classification, estimation, prediction, affinity grouping and description problems. We can call the whole process as “knowledge discovery”. In 1998, Weiss classified Data mining into two parts: knowledge discovery and prediction. First part includes classification , regression whereas second part defines association rules, classification , summarization etc.

Knowledge Discovery Database (KDD) has three stages.

-> Data Preprocessing

-> Data mining

-> Data Post-processing

For the initial stage, data processing is done which results in data collection, data smoothing, data transformation, data cleaning and data reduction. In the second stage which is called data mining which involves data classification commonly termed as prediction. The final and the third stage which we called data- processing, which shows the conclusion part drawn from the analysis in the previous stage.

Predictive data mining works in the same way how a human do the analysis of any small data set, the only advantage it PDM is that we can use it for large data set with fewer problem than any human analyst has. PDM basically learn from its previous mistakes and will never repeat the same mistake again in future like humans. Nowadays, predictive approach of data mining is in its most developed phase.

**2. Methodology**

We divided the data set into two parts, setting the odd numbered data points as “training set” and the even numbered data set as “test validation set”. The main purpose of using the train data is for model building. To build a model, a predictive data mining technique is used and the various methods of that techniques has also been engaged.The model accuracy is checked by uploading it onto the competition website. In this paper, the study of loan data set has limited to the model validation based on the data set. Finally, all the methods of glm technique is described and compared with random forest result and the best result is shown.

**A. Data Acquisition**

Data set used for the research is Loan Dataset. For this work, Rstudio software is used for all the analyses. All the analyses (glm,rm) have been made before the software is used for the dataset. Before the technique is used on the dataset, an introductory analysis is done on the data set to gain knowledge of dataset.

**B. Data Description and Preprocessing**

The plot between each variable in the data set and the indices is made to see the dispersion between variables, which are different. The plot between the input variables and output variables is made to know the relationship between them.

**C. Generalised logistic model**

In statistics, logistic regression, or logit model is a regression model where the dependent variables or output variable is categorical. It comes under classification techniques. It can be binomial, ordinal or multinomial depending on the outcome of dependent variables. Binary logistic is used where predictor variable has two possible outcomes, “1 / 0”, “Yes / No”, “True / False”. Whereas, multinomial is used with more than two outcomes. Logistic regression can also be thought as a special case of linear regression where the outcome variable is categorical, where we are using log of odds as dependent variable. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function.

**D. Random Forest**

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.The more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. Random forest algorithm or the random forest classifier can use for both classification and the regression task.Random forest classifier will handle the missing values.Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

**E. Procedure**The data set is first introduced, as well as pre-processing on the data set is done, in order to gain insight of the properties of the data set. By plotting the inputs over the output of the raw data set, relationship check is made. To reduce the level of dispersion between the variables in the data set, the data is pre-processed. Pre-processing is done by the scaling or standardizing the data set, it is also known as data preparation.A predictive data mining

**3. Data Introduction**

We took dataset from kaggle which was collected from different customers when they fill out their application while applying the loan. We were having 12 variables and all of them plays an important role in deciding whether a customer should get the loan approval or not. The data set is divided into test and train part which helped us to build our model on train data and then apply it on test data.All the variables in the dataset are independent.

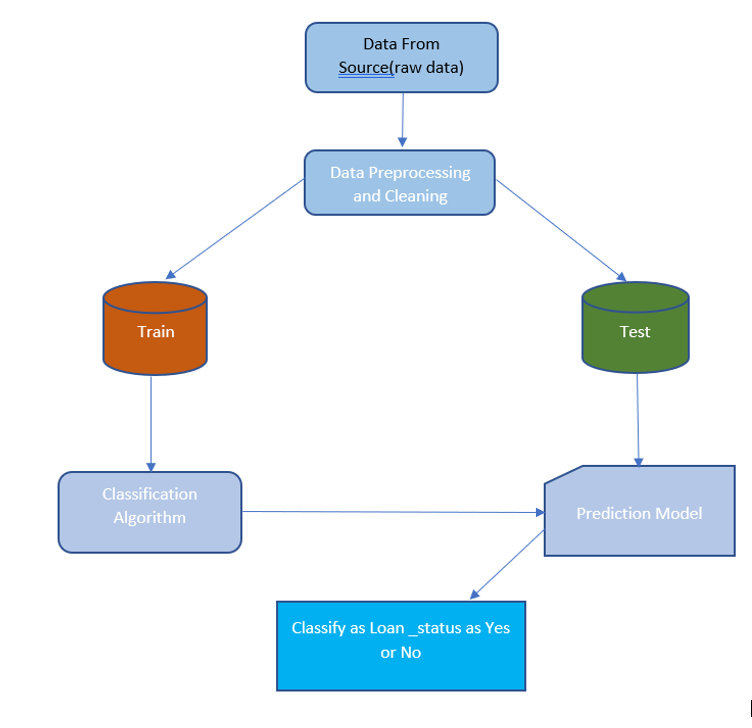
**4. Comparison Criteria**

Mean Square Error, a criterion for model comparison is used in my research.MSE is the most significant criteria for determining and comparing different data mining techniques. MSE measures the difference between actual test outputs and the prediction test output. Smaller MSE is better. Large MSE values show poor prediction. The MSE of the predictions is the mean of the squared difference between the observed and the predicted values.

R-Square, also termed as R-Sq or R2 . It is used to measure the percentage variability in any of the data matrix, which is accounted for by the built model. The value closer to 1 of R-Sq, shows a better prediction.

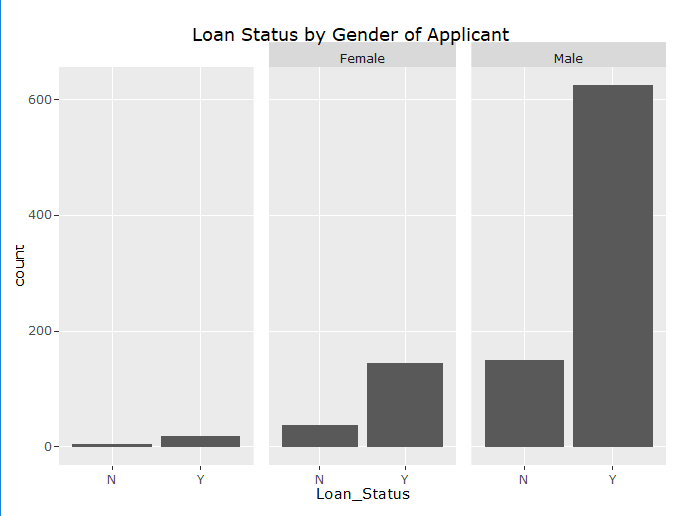
**WORKFLOW OF PROJECT**

**The diagram below shows the workflow of this project.**

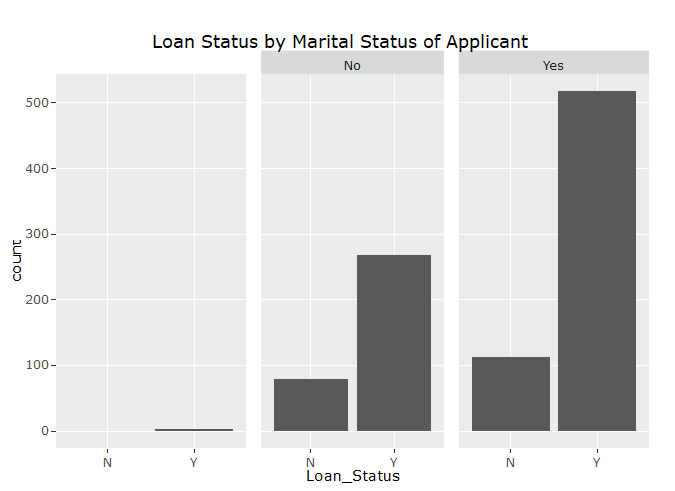
****

**5. Visualization**

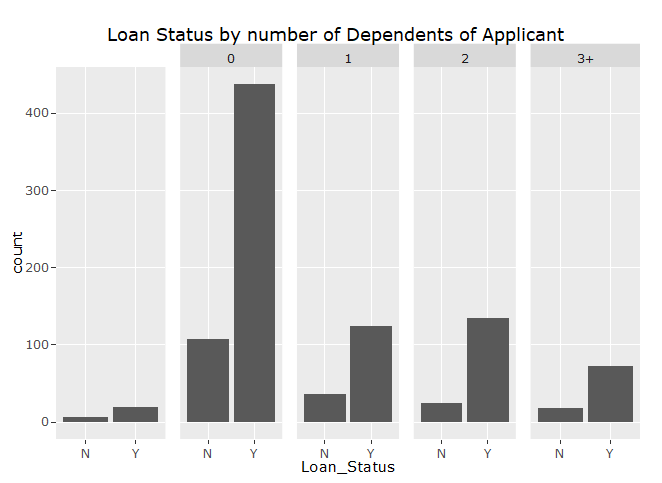
Many factors decide if the bank will give a customer loan or not. From the loan data available we can visualise what factors affects this decision. Histogram of loan status for each gender shows that out of 775 male customer 625 were approved for loan which make 80.6% approval rate for male. While for female customers 145 of them were approved out of 182 which makes approval rate of females to be 79.6%. This shows male customers are more likely to apply for loan than females but the probability of getting loan approved is not dependent on sex as the probabilities are near equal. It is also seen that there are few missing values in this graph. Analysis on missing values will be done in coming steps.

****

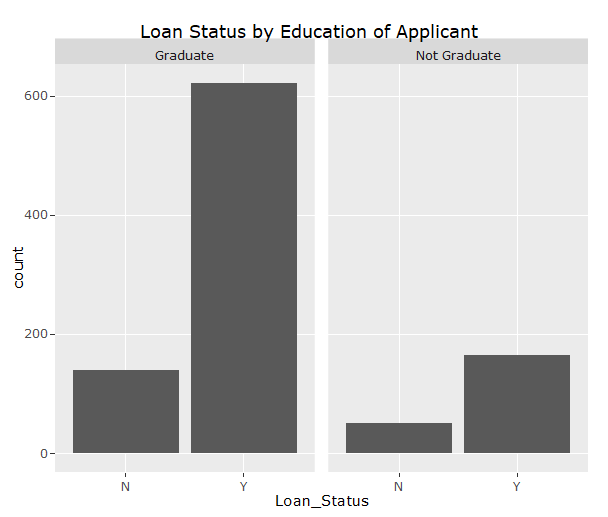
To check if the approval rate is dependent on Marital status or not, plotting a histogram again will help. 631 married customers applied for loan while only 347 unmarried applied for loan. This tells us that married customers are twice as likely to apply for loan than unmarried customers. This makes sense because married customers need to keep pace with children’s education as well as needs of the family. Approval rate for married customers is 82.1% while for unmarried customers is 77.23%. Marital status plays a good role in a successful loan approval. There are few missing values in this column as will which we can see in the first block.

****

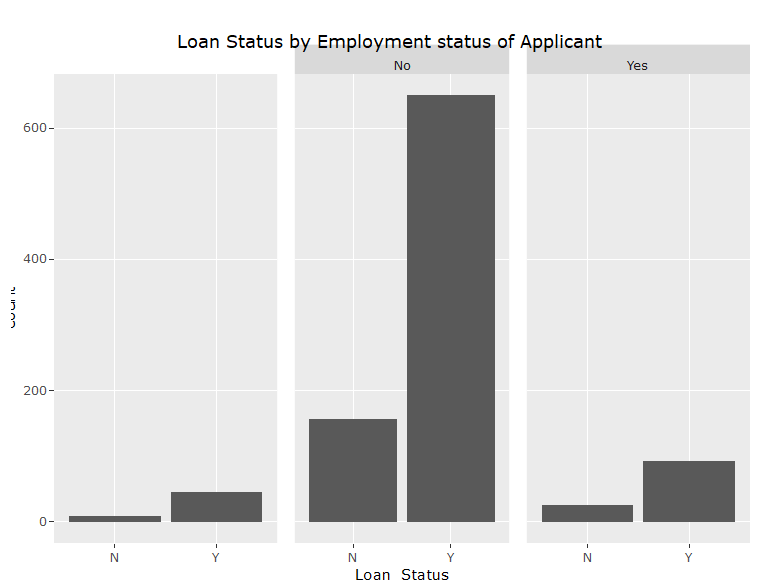
Lets analyse how loan approval is related to number of dependents of the customers. From loan data we see that loan is majorly applied by the customers who are having no dependents which is as high as 545 customers out of 981 customers and approval rate is 80.4%. While the approval rate for customers with one dependent is 77.5%. The first block represent the customers whose data for number of dependents is missing.

****

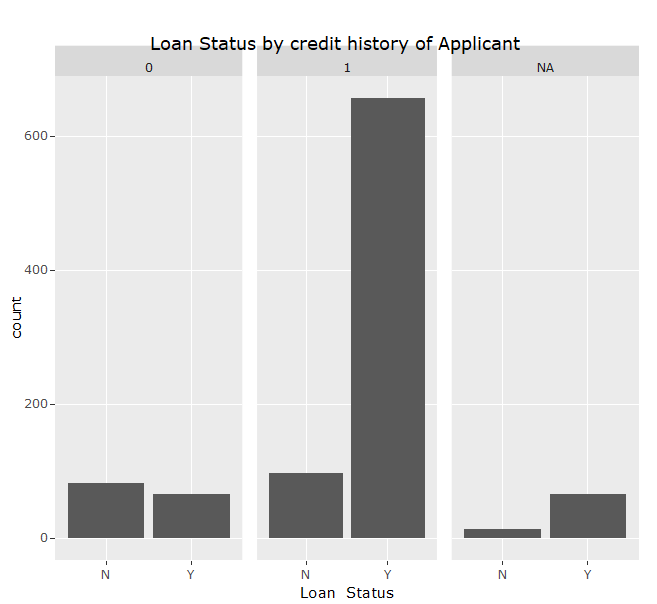
77.78% of customers who are graduated apply for loan verses 22.23% who are not graduated. Therefore we can infer that application of loan is highly dependent on the education level of customer. Graduated customer is more likely to get his/her loan approved as its approval rate is 81.7 % than the one not graduated.

****

Another interesting factor on which loan approval is dependent is Employment status of the Applicant. Self employed customers are very unlikely to take loan. From the total population of 981, 809 customers who took loan were NOT self employed while only 119 were self employed and for 55 customers employment status is missing. There is 80.5% chance to get loan approved if the customer is not self employed. Few values are missing in this column which are shown in 1st block.

****

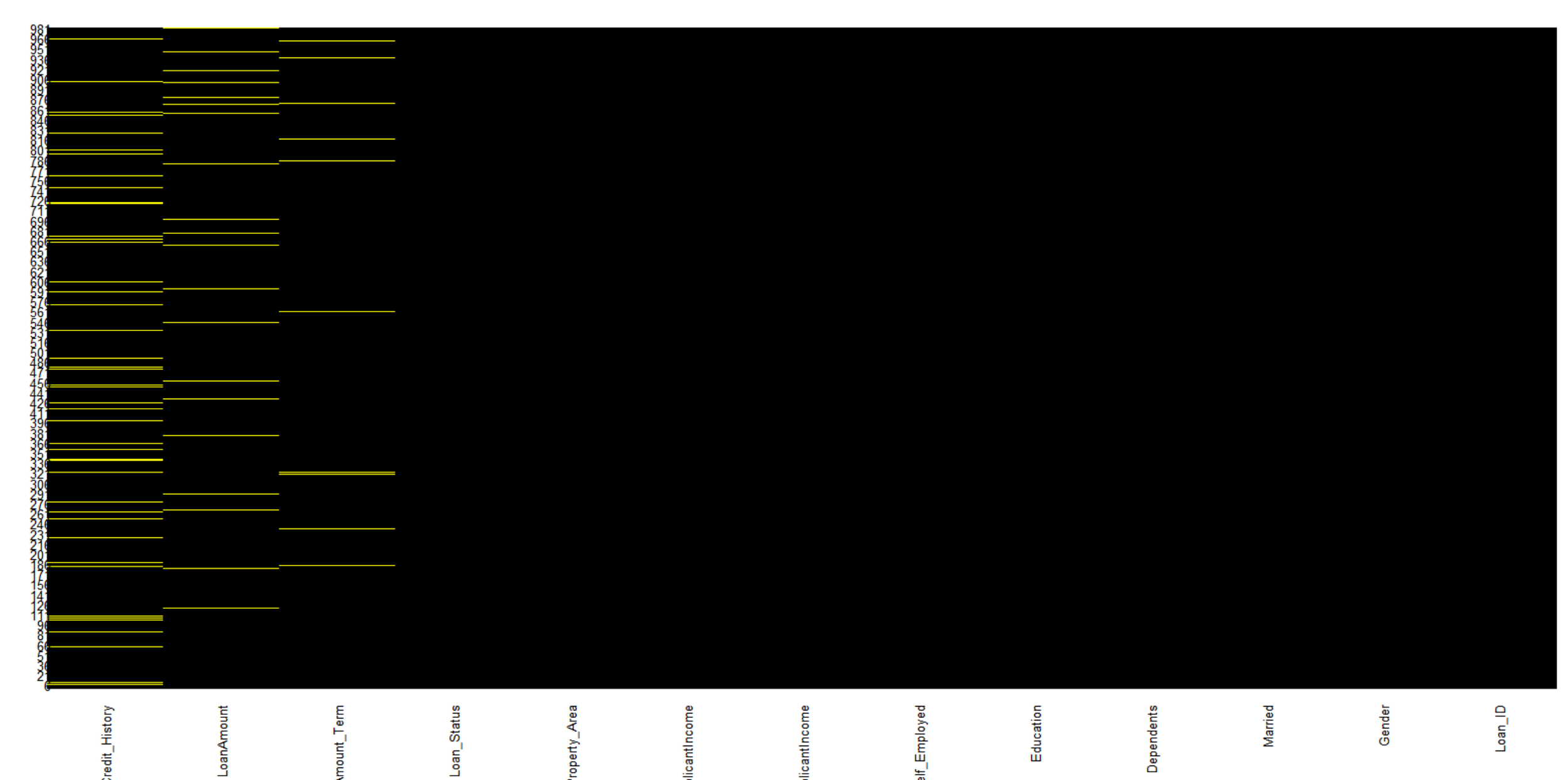
Credit history of a customer plays a very significant role in credit card approval. Our data have 79 missing values in this column which will be cleaned in upcoming steps. Customers who doesn’t have a credit history have on 45% of chance to get the loan approved from the bank while for customers who have a credit history have chances as high as 88%. As the percentage is so high, it proves that credit history do play a significant role in the decision of loan approval. But we can not say what score of credit history will decide this as we don't have this data.

****

**6. Data Cleaning**

Data cleaning is a process of removing data in a database or dataset that is incorrect, incomplete, improperly formatted or duplicated values or missing. Process of removing errors and resolving inconsistencies in source data before loading into targets is called data cleaning.

In this loan data set 13 variables are present in which 3 variables are error free. Loan\_ID, “ApplicantIncome”, “CoApplicantIncome”. Other variables are either having missing values or N/A values. We can visualize this by using Amelia package.



**Before cleaning of dataset**

**library(Amelia)**

**missmap(com,main = 'Main',col= c('yellow','black'),legend = FALSE)**

With the help of amelia package and missmap function one can observe that credit\_history, LoanAmount and LoanAmount\_Term has N/A values. And other variable have missing values which can not be displayed by this function. Removing N/A and missing values and imputing it with the right values is a critical task.

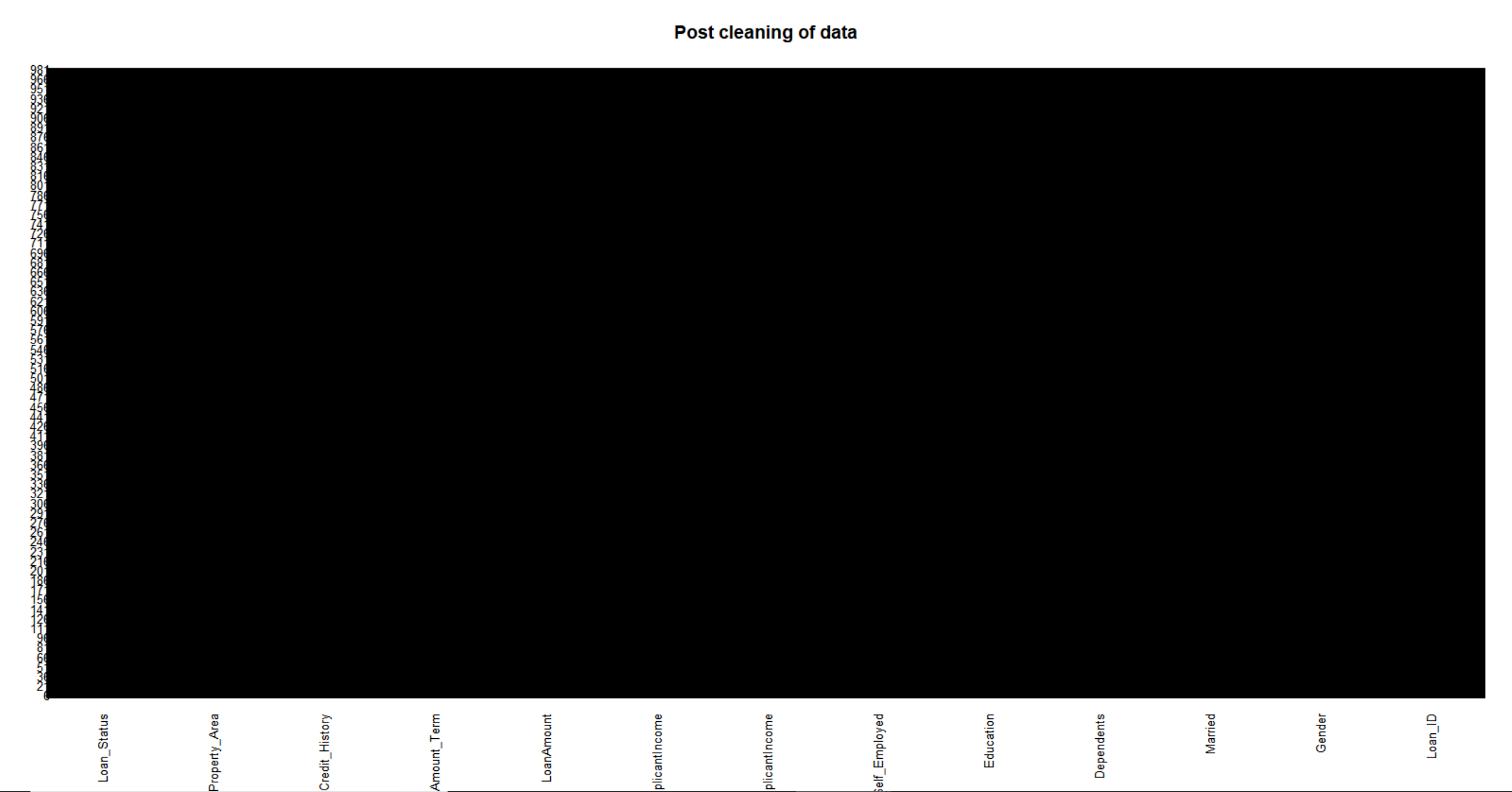
Loan Amount is replaced by using range of Applicant Income and imputed based on the mean taken of the Loan Amount of the specified range. Most of the Applicant range from 0-10000. So range is being taken for the interval of 2000 starting from 0 till 10000 and one more range for 10000 and above. Taking the mean of Loan Amount of the specified range and imputing the mean, and replacing it with N/A values.

Credit History has N/A values and that is being replaced by taking mean of the applicant income for credit history ‘1’ and ‘0’ respectively. Than imputing on the bases of the mean range of Applicant income.

N/A values in Loan term Amount are imputed by using Mode function over Loan Amount term which gave result as 360.

Other variables have missing values and that is not represented by missmap function like Gender. Gender has many missing values and that is replaced by using mean of ApplicantIncome for Male and female separately. Mean of male income was higher as compared to female. Imputation is based on the mean range of ApplicantIncome. Same procedure was followed for Marriage. As Marriage also has missing values and imputation is done accordingly. Employee status have also missing values and imputation of missing values is done by taking mode of the values. Employee statues has 2 variables. ’Yes’ and ‘No’. 85% values were ‘No’ so replace all the missing values it by ‘No’.

Dependent contains lot of missing values. Trimming of mean ApplicantIncome is done based on marital status and number of dependent. Imputation is done based on trimmed mean which is taken for all the possible marital status and number of dependent.



**Post cleaning of data**

Cleaning of data leads to more accuracy and make the algorithms work in better way.

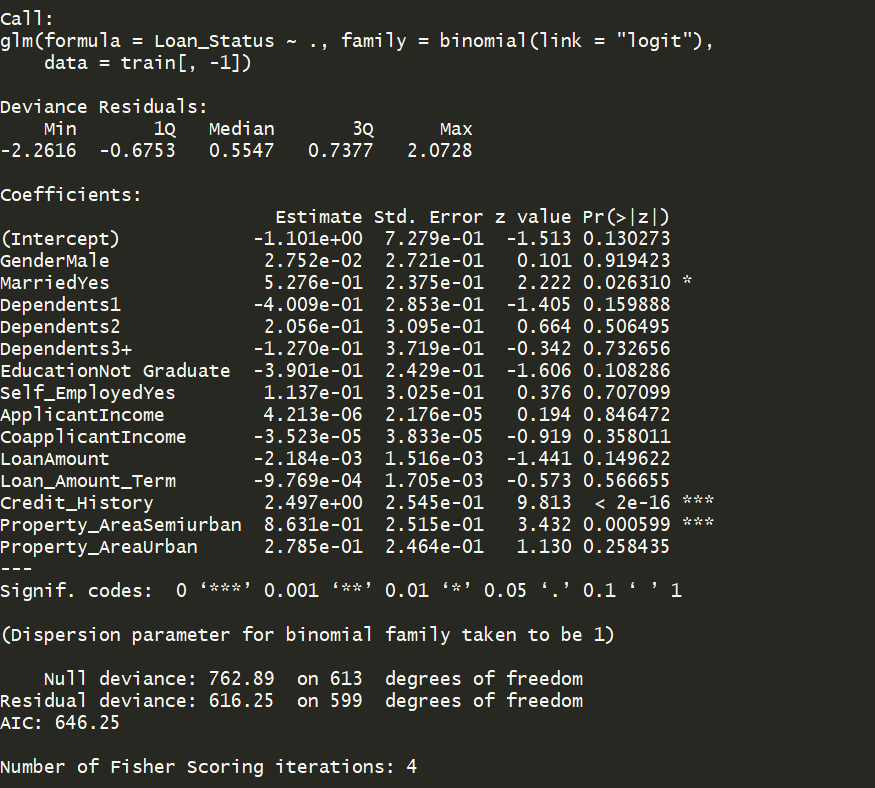
**7. Analysis Using GLM**

We can now fit a logistic regression model to the data using the glm function. We will fit the model on all the independent variables. The code to fit the model is:

log.model<- glm(formula = Loan\_Status ~ .,family = binomial(link = 'logit'),data = train[,-1])

Summary(log.model)

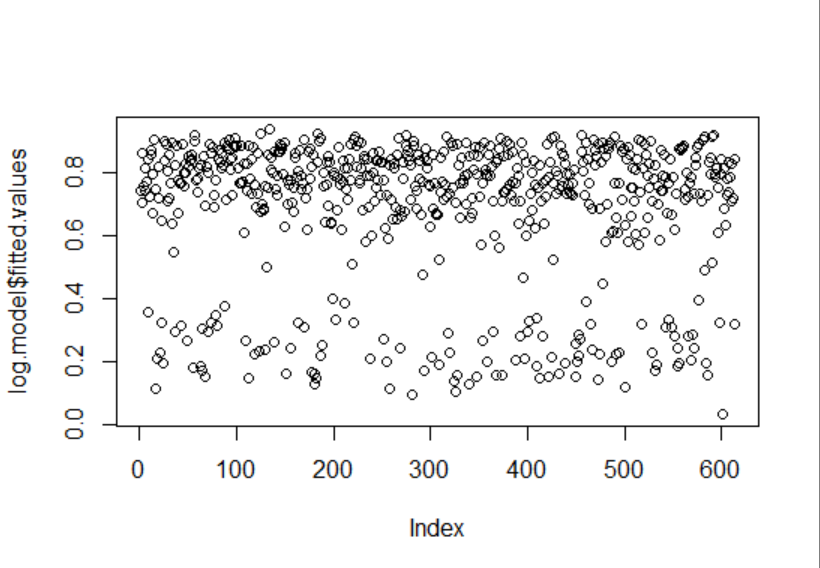
From the results in the below Figure we see that the Credit\_history and Property\_suburban is most significant features as there P-value is less than 0.05.



Now, after this we will predict the values for test data by fitting the GLM model. The code for that is :

df1$Loan\_Status = predict(log.model, type="response",newdata = test)

From the results in the below Figure we see that the Credit\_history and Property\_suburban is most significant features as there P-value is less than 0.05.



It can be seen that we are getting the continuous values of probability but we need a binomial outcomes in form of 0 and 1 so we convert the values like those having probability greater than 0.5 to 1 and rest will be converted to 0.

fitted.results <- ifelse(df1$Loan\_Status > 0.5,1,0)

**8. Analysis Using Random forest**

We can now fit a random forest model to the data using the glm function. We will fit the model on all the independent variables. The code to fit the model is:

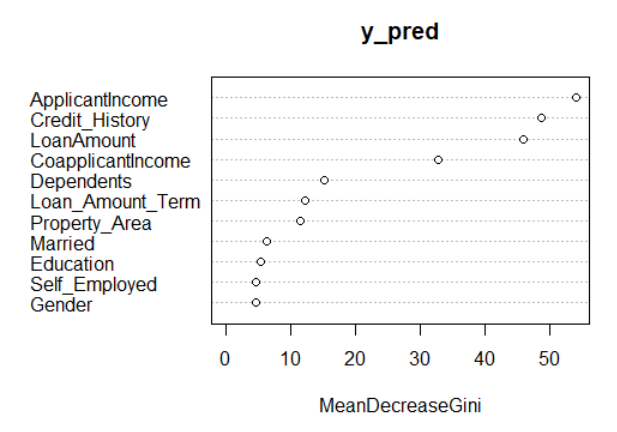
y\_pred <- randomForest(x=train[,c(-1,-13)],y=train$Loan\_Status,ntree = 10)

fitted.results <- predict(y\_pred,newdata = test[-1])

Summary(log.model)

importance(y\_pred)

varImpPlot(y\_pred)



From the results in the Figure we see that the Applicant Income , Credit\_history and Loan Amount Are most significant features as their Mean DecreaseGini value is higher than others.

**9. Result and Comparison**

In this chapter, analysis on the dataset is done. Firstly, glm, predictive data mining technique is applied on the dataset and then random forest, another predictive data mining technique is applied and then both are compared with each other, in order to check the effectiveness of the models. In both the techniques, the resulting probability is then assigned a value 0 and 1. Those probabilities who are having value 0.5 or greater have assigned 1 and rest of them 0. After applying the glm model we got 79.24% accuracy whereas random forest gave us approx 77%. We can see the glm model is more effective here.

**10. Acknowledgment**

We would like to appreciate our fellow classmate for their support and our professor who always have motivated us to complete this project.

**6. References**

**[1] Berry, Michael J. A. et al., Data-Mining Techniques for Marketing, Sales andCustomer Support. U.. S A: John Wiley and Sons (1997).**

**[2] Weiss, Sholom M. et al., Predictive Data-Mining: A Practical Guide. San Francisco, Morgan Kaufmann (1998).**

**[3] Jolliffe, I.T., Principal Component Analysis, New York: Springer-Verlag (1986).**

**[4] Naes, T., and H. Martens, "Principal Component Regression in NIR Analysis: Viewpoints, Background Details and Selection of Components," J.Chemom. 2 (1988).**

**[5] Sun, J., "A correlation Principal Component Regression Analysis of NIR." J.Chemom9 (1995)**

**[6]Practice Problem: Loan Prediction III | Knowledge and Learning. (n.d.). Retrieved February16,2018**