**Abstract**

We are predicting loan status for the customers based on previous loan status history using Logistic regression.

**Introduction**

DFH company provides loans to the customers to fulfil the dream of building an own house. This company is located in rural, urban and semi-urban areas. This company validates the application of the customer and check for the loan eligibility of that customer.

I chose this dataset because it would be interesting when dealing with income, loan and credit values.

**Aim and Objective**

The aim of this company is to automate the process of checking loan eligibility of their customers based on the information they provide by the applicants while filling the form. The application forms consist of name, gender, address, marital status, no. of dependants, education status, income of the applicant, income of the partner, requested loan amount, loan term, credit history, property area and other details. We are going to predict the accuracy of the loan approval status based on the information provided.

**Explanation and preparation of datasets**

Description of dataset

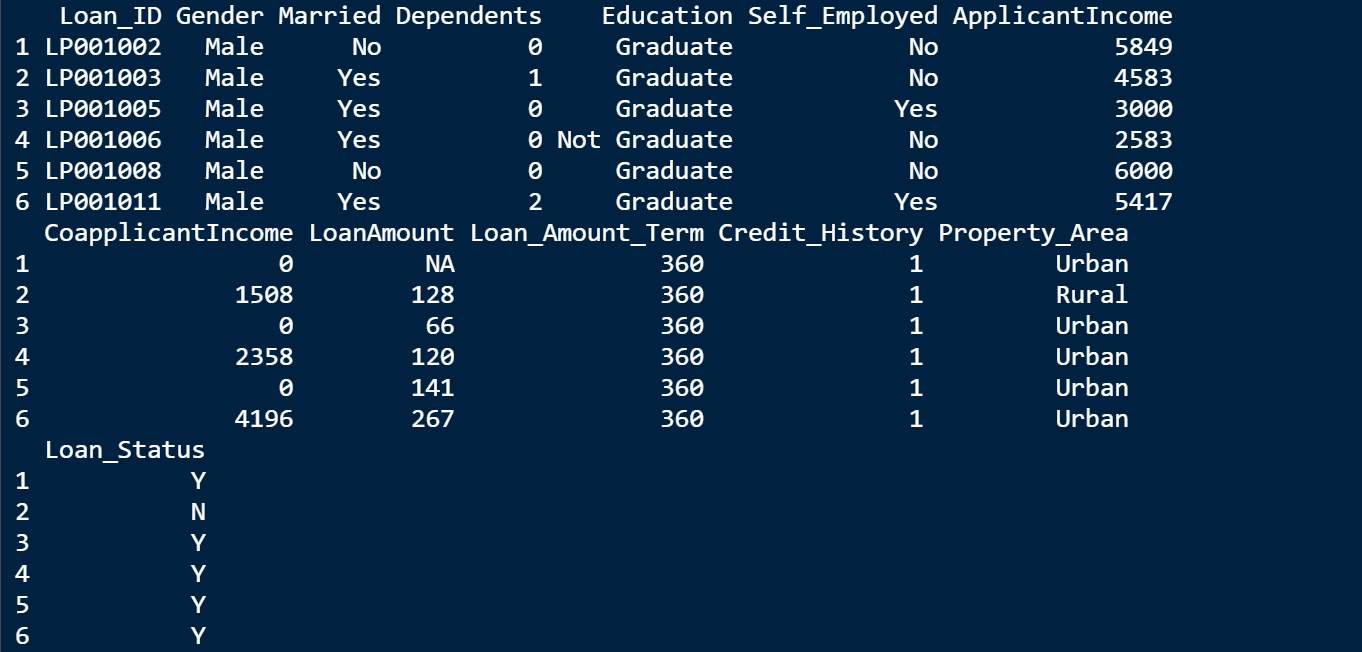
The data set consists of Loan ID, Gender, Married, Dependants, Education status, Self employed, Applicant income, Coapplicant income, Loan amount, loan amount term, credit history, property area and loan status. From this we understand that loan status is the dependant variable and other variables are independent except Loan ID because this is unique to each customer.

**Logistic Regression in R**

Import the required dataset and explore and buld the model.

#setwd(C:/Users/SAT/Desktop/Sandhya Project\_ASDMTask1R)   
#import data

data = read.csv('Loan Status DataSet.csv')  
head(data);#tail(data)



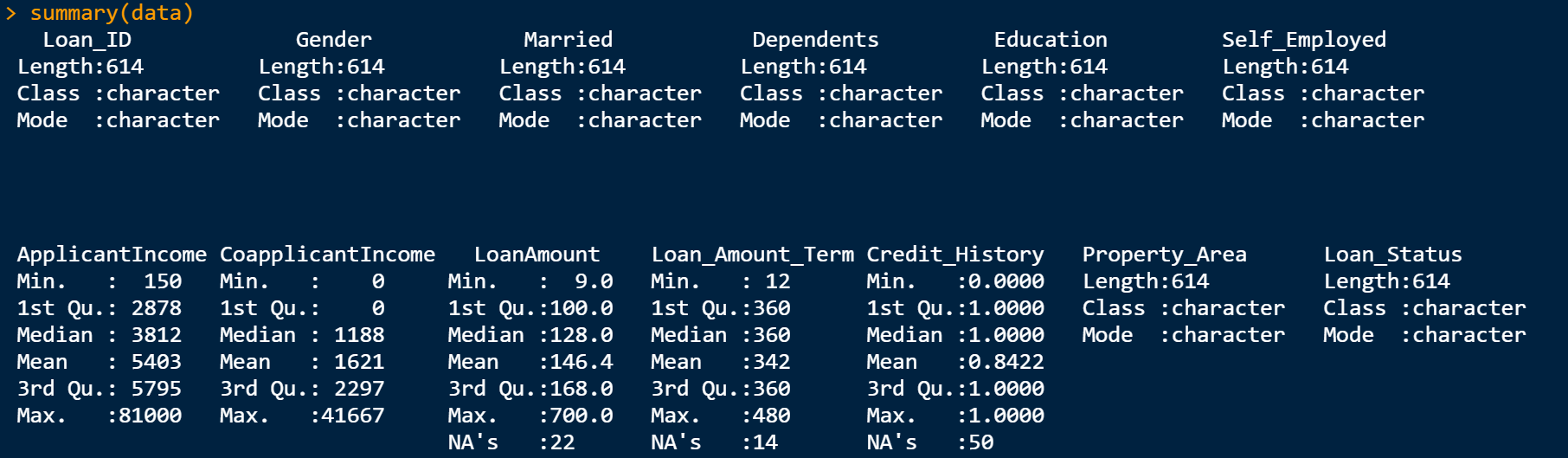
Here the first displays the header of each column (variables in the dataset). Now let’s run the summary and describe the data.

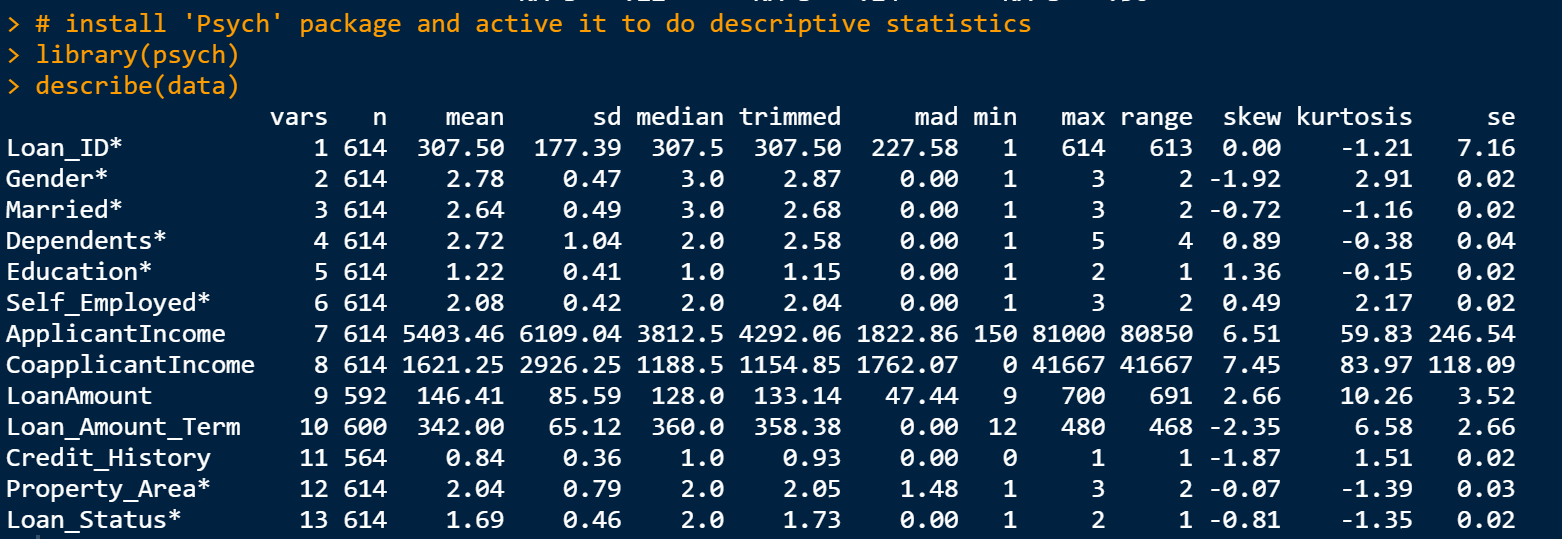
summary(data)

# install 'Psych' package and active it to do descriptive statistics

library(psych)

describe(data)





From the summary of the data, we can see Loan\_ID is not necessary for predicting the loan status as it is a unique number to individual customer. Let’s remove the Loan\_ID column.

# Remove unnecessary columns - Loan ID is not unnecessary as because it just unique id

Data = subset (data, select = -1)

Now let’s check on the missing values in the dataset taken

# Check for Missing values - Percentage

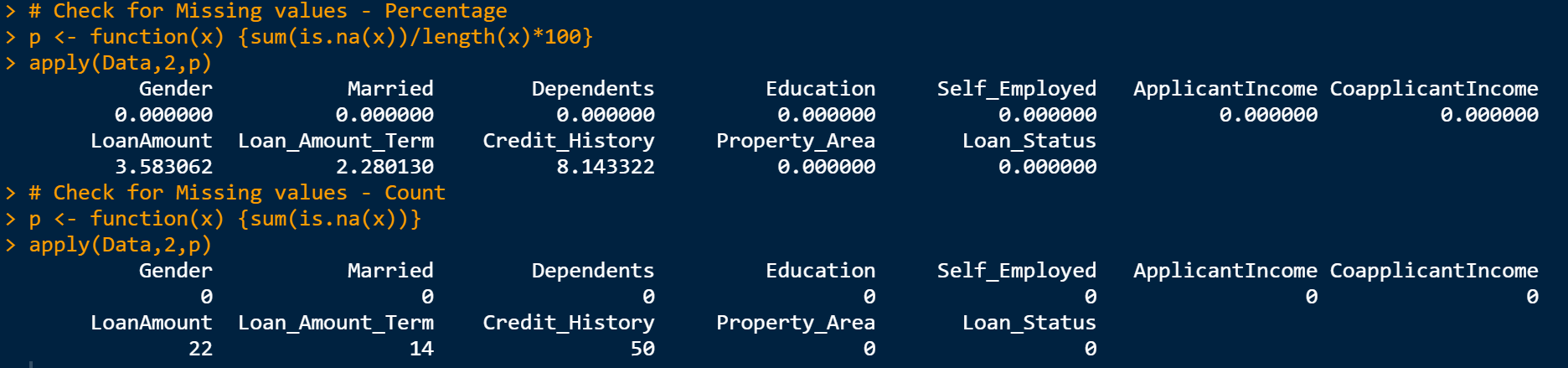
p <- function(x) {sum(is.na(x))/length(x)\*100}

apply(Data,2,p)

# Check for Missing values - Count

p <- function(x) {sum(is.na(x))}

apply(Data,2,p)



From the above information, we understand that there are few missing values. Let’s impute the missing values with mean function

#Impute Missing values With Mean

for(i in 1:ncol(Data)){

Data[is.na(Data[,i]), i] <- mean(Data[,i], na.rm = TRUE)

}

By using factor(), convert all of the categorical and numerical vectors into factors. So that all the variables come to one level now.

Building model in R

Now we are in a stage where we can split the data into train set and test set. Train set is subset of the data set which going to be trained. Lets use test set to validate our prediction.

Here is the dimension for train data and test data

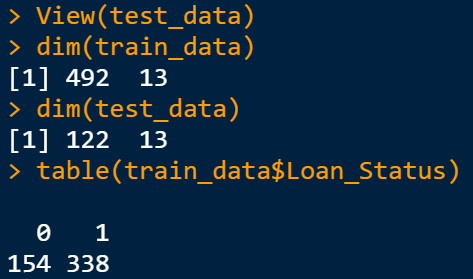
# Train and test split

library(caTools)

sample <- sample.split(Data$Loan\_Status, SplitRatio = 0.8)

train\_data <- subset(Data, sample == TRUE)

test\_data <- subset(Data, sample == FALSE)



Let’s run our model using logistic regression. Logistic regression predicts the probability of occurance of an event by fitting data to a logit function. glm() is used to build the model.

# Build Logistic Model

set.seed(1)

logitmod <- glm(Loan\_Status ~ ., family = "binomial", data=train\_data)

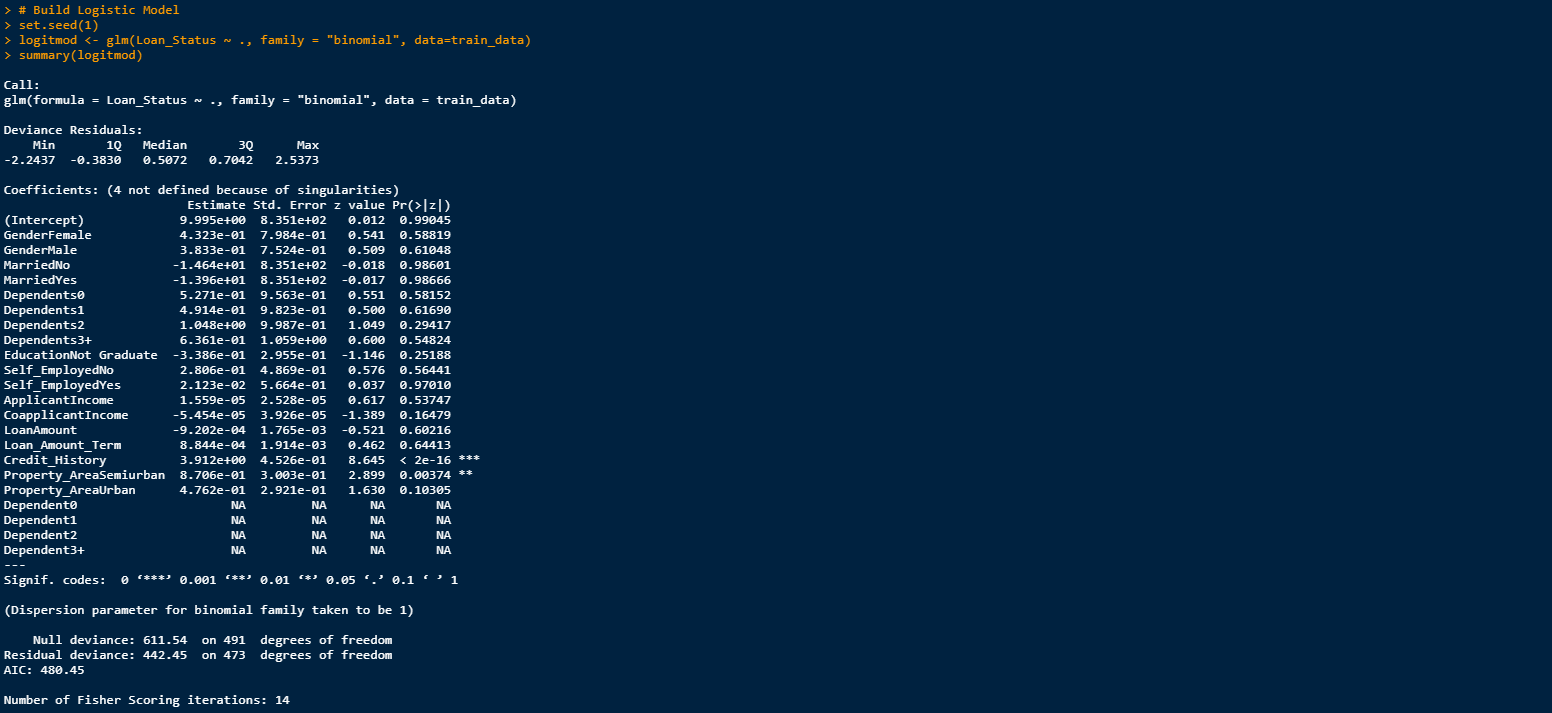
summary(logitmod)

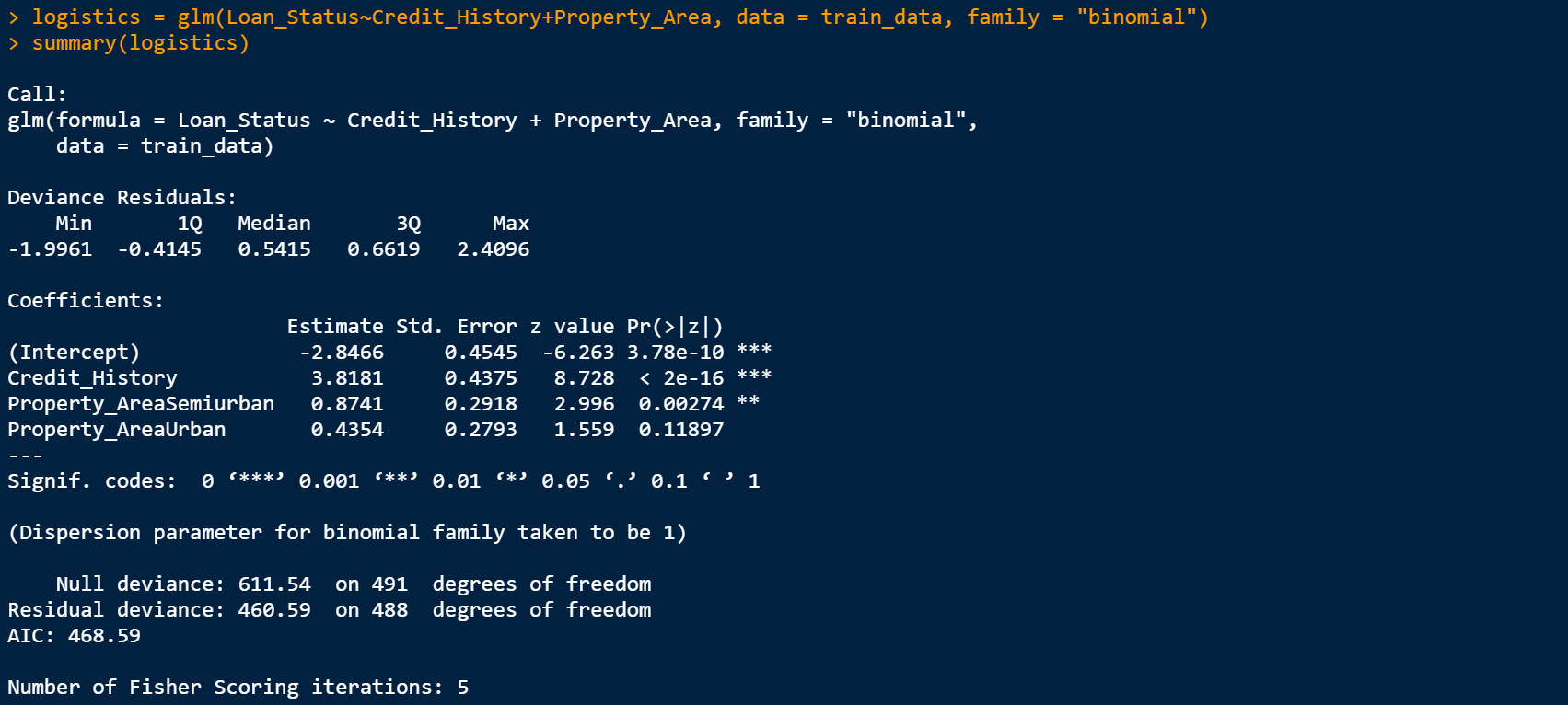
logistics = glm(Loan\_Status~Credit\_History+Property\_Area, data = train\_data, family = "binomial")

summary(logistics)

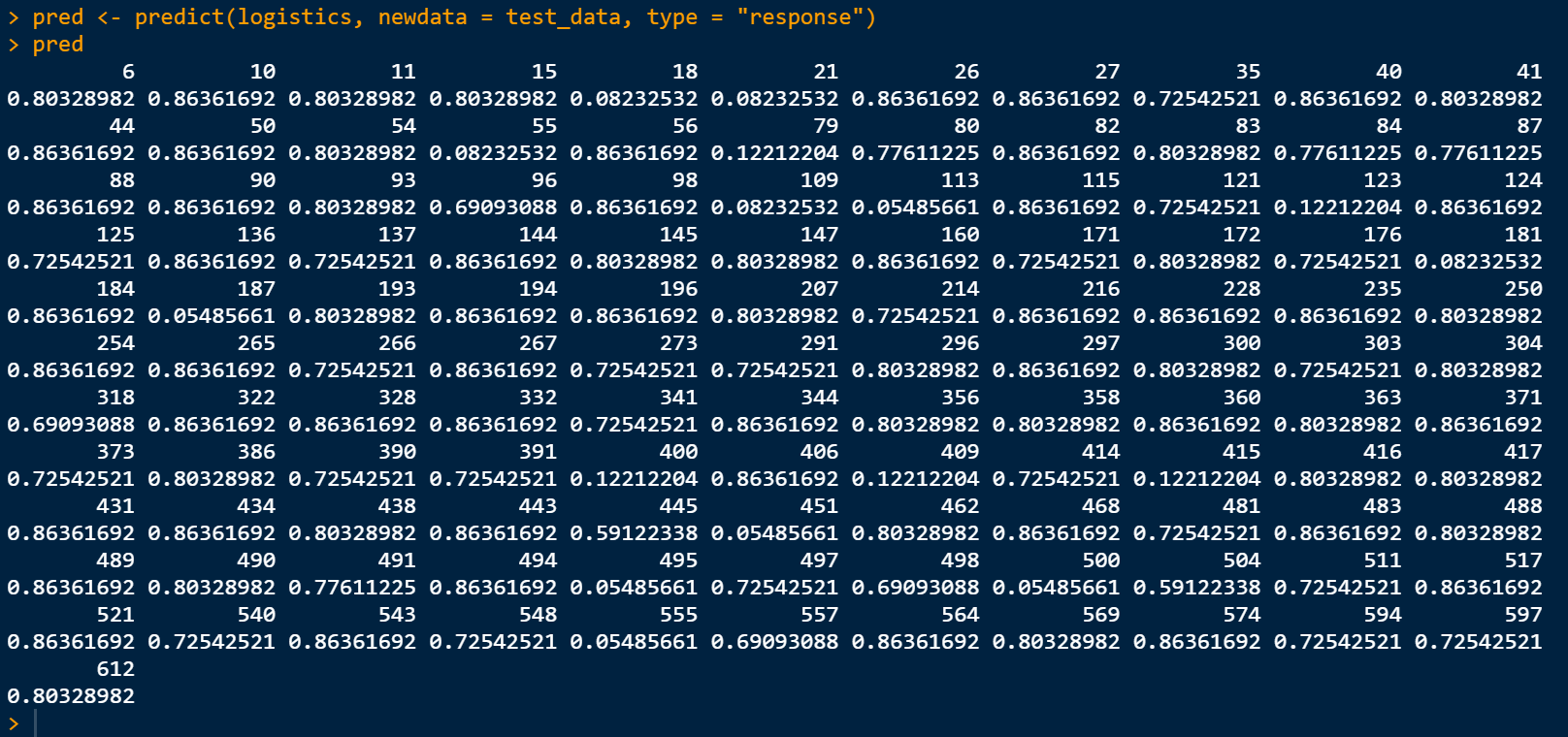
We are not going to take all the variables because it may lead to overfitting of the data. Lets choose the variables logically by examining the importance of each variable. For example the chance of approving loan is higher if the applicant has take a loan before. This mean credit history is the variable. Loan can be approved if the applicant has high salary. This mean Apllicant income is next variable. We can consider some other factors like education, about the job stability and here it is a home loan so we can consider the location of the area.

From the below observations, we understand that Credit\_history and Property\_AreaSemiurban are important variable to be validated to approve the loan. AIC is taken into consideration when we run more models and we choose the least AIC score can be considered. It shows call in the first which means the model we run for the data. We also see some distribution statistics in the result. It also give the p value(<0.05). The least the p value we can associate the changes with dependant variable and it is significant. Here the p value is least for Credit\_history and Property\_AreaSemiurban. Hence these variables are taken onto consideration to run the model.





The below screenshot displays the predictions for all the values in test\_data



**Results in R**

A confusion matrix table is also generated to check the accuracy of both the train data and the test data. From the confusion matrix we can say that we have predicted well. And over all accuracy of the model is 80.33% which can be considered.

