**LOAN ELIGIBILITY PROCESS SOLUTION**

**Introduction**

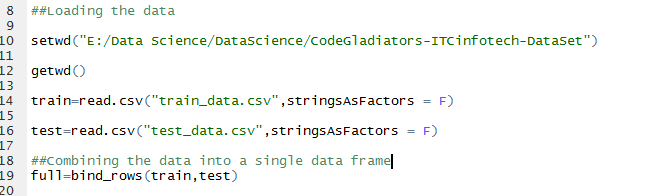
There are following steps included in my script:-

1. Loading the data
2. Preparing the data for prediction
3. Working on validating data to find the best model that suits the prediction
4. Prediction

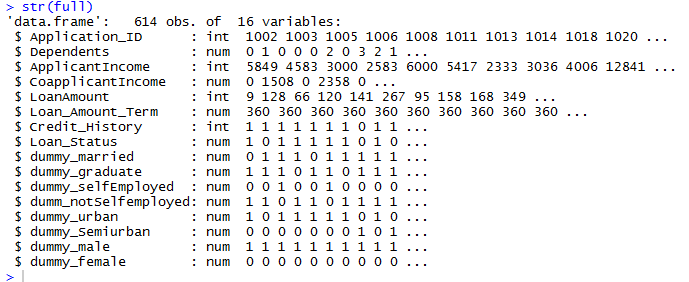
First we load all the libraries that are required to run the prediction



Now that all the packages are loaded let’s look into the data:-

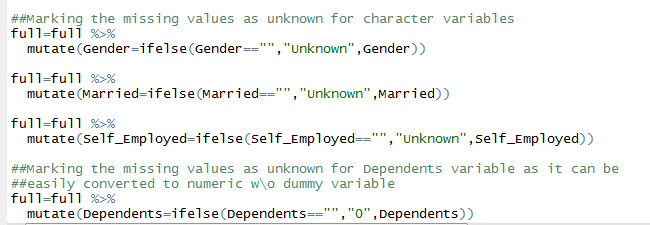


A glimpse of the data to the variable names, data type and some values



**Data Preparation**

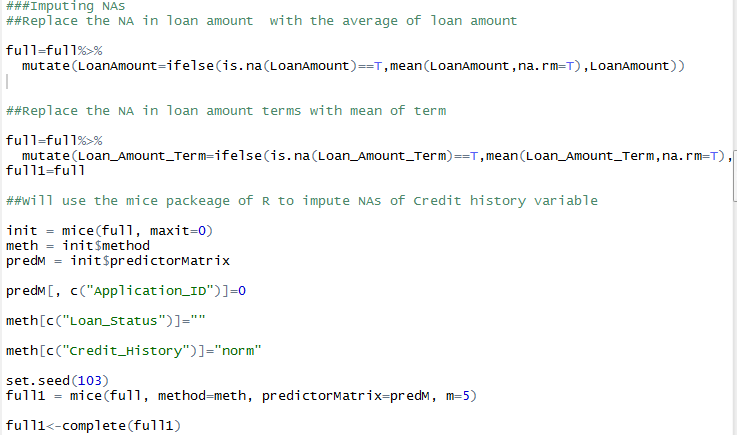
Many of the character variables have missing values , we will convert them to unknown first:-



Next we will start working with character variables and convert them into numeric variables by creating dummy variables for each of them.

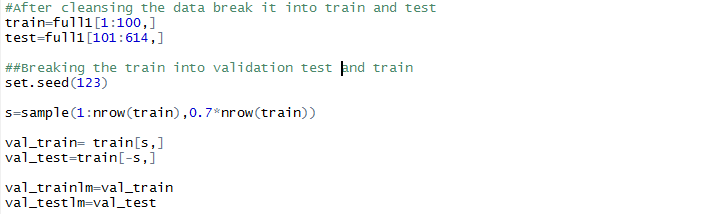


Imputing the NA values of numeric variables:-

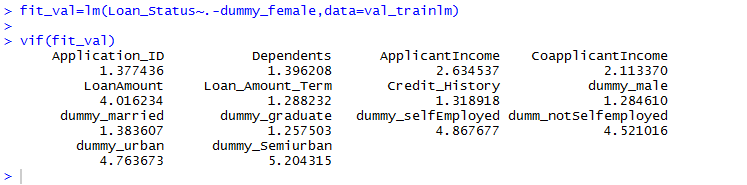


After cleansing the data break it into train and test.

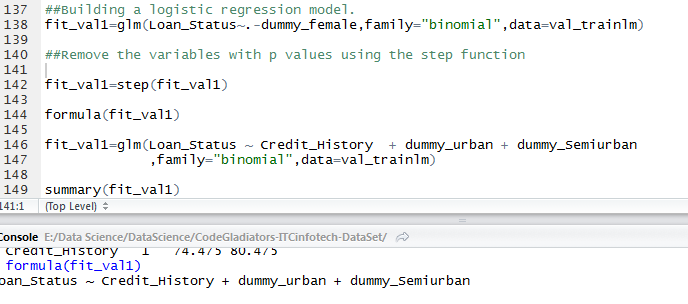
Further break the train data for validation purpose which will help us determine which model will be best for prediction



First we will remove perfect multi collinearity by removing the variables with high vif values one by one

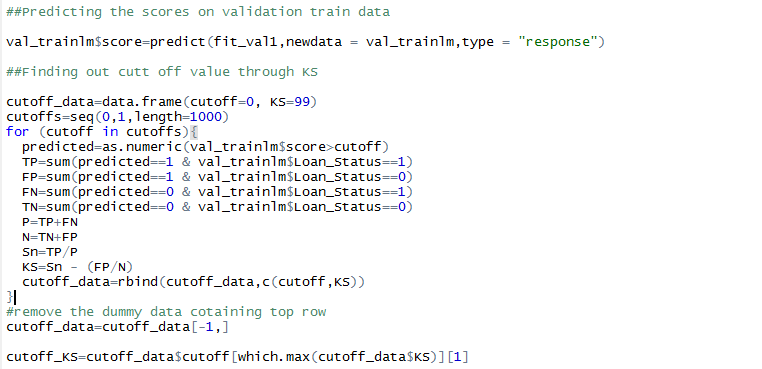


Next we build the logistic regression model on validation data and remove the variable with high p values using step function. Also any variable left with high p value is further removed manually.

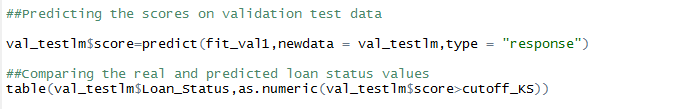
**Final logistic model-**

fit\_val1=glm(Loan\_Status ~ Credit\_History ,family="binomial",data=val\_trainlm)

Predicting the scores on validating train data and finding the cutoff value using KS



Predicting the scores on validation test data and comparing the real and predicted loan status



Predicted Values

1 0

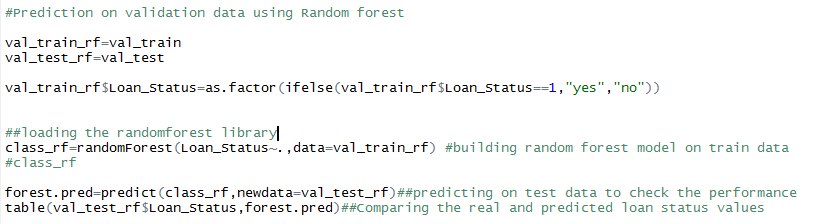
Real values 1 14 2

0 8 6

Following parameters obtained from above data

Misclassification error=10/30=0.33 Sensitivity=14/16=0.875 Specificity =6/14=0.4285714

Prediction on validation data using Random forest:-



Predicted Values

1 0

Real values 1 15 1

0 8 6

Following parameters obtained from above data

Misclassification rate:- 9/30=0.3 Sensitivity:- 15/16=0.93 Specificity :-6/14= 0.4285714

**Random Forest turns out to be more accurate**

From the above predictions the random forest seems to be more accurate. We’ll be using Random forest model on train data to predict on test data.

**Conclusion**

We have used Random forest to predict the loan status on test data. Since the Specificity is not that close to 1 the prediction of ‘N’ for loan status will be less accurate.