**Loan Prediction Model**

**Statistical tool used : R**

**Problem statement** (https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#)

**About Company**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

**Problem**

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

## **Data**

|  |  |
| --- | --- |
| **Variable** | Description |
| **Loan\_ID** | Unique Loan ID |
| **Gender** | Male/ Female |
| **Married** | Applicant married (Y/N) |
| **Dependents** | Number of dependents |
| **Education** | Applicant Education (Graduate/ Under Graduate) |
| **Self\_Employed** | Self-employed (Y/N) |
| **ApplicantIncome** | Applicant income |
| **CoapplicantIncome** | Co-applicant income |
| **LoanAmount** | Loan amount in thousands |
| **Loan\_Amount\_Term** | Term of loan in months |
| **Credit\_History** | credit history meets guidelines |
| **Property\_Area** | Urban/ Semi Urban/ Rural |
| **Loan\_Status** | Loan approved (Y/N) |
|  |  |

**Splitting the data:**

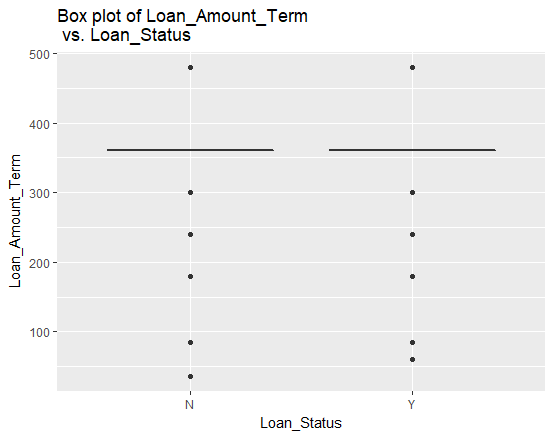
To develop an optimized statistical model for the automated loan eligibility status, the given training data was further divided into training and test subsets. 75 % of the data was used to train the first 2 logistic regression models and the 70 % data was used to train the knn model.

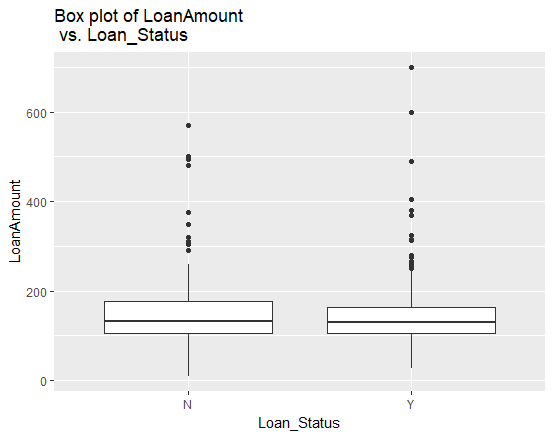
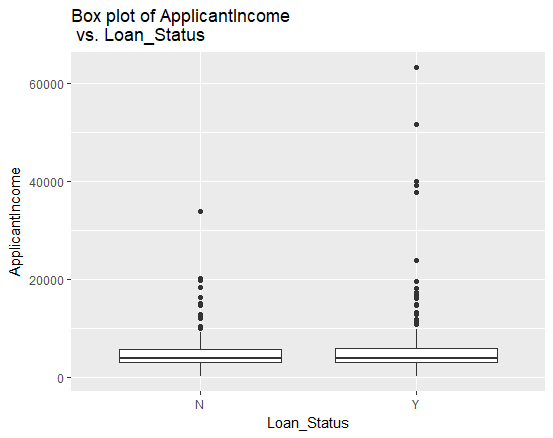
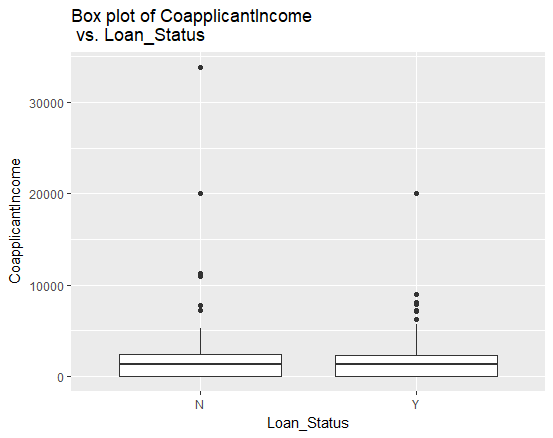
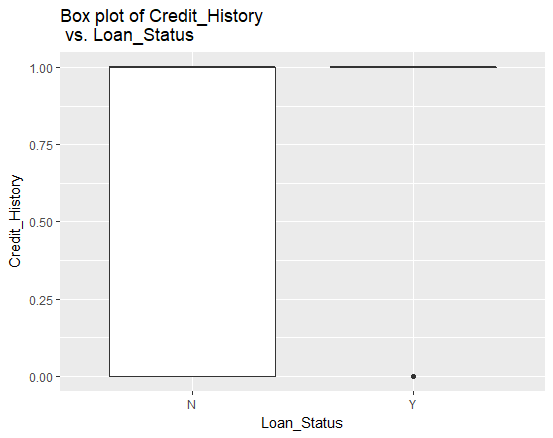
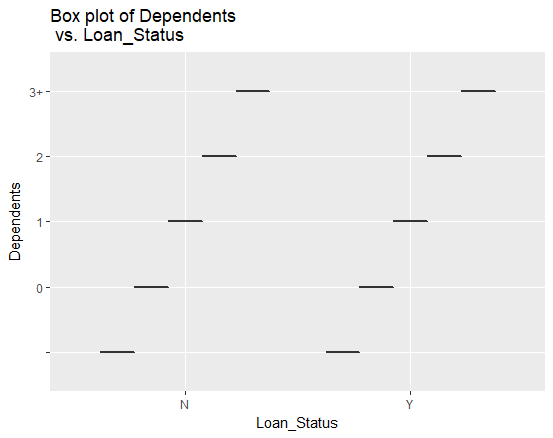
**Exploratory Data Analysis:**

To analyze which details affect the loan eligibility of a customer, exploratory data analysis was done.

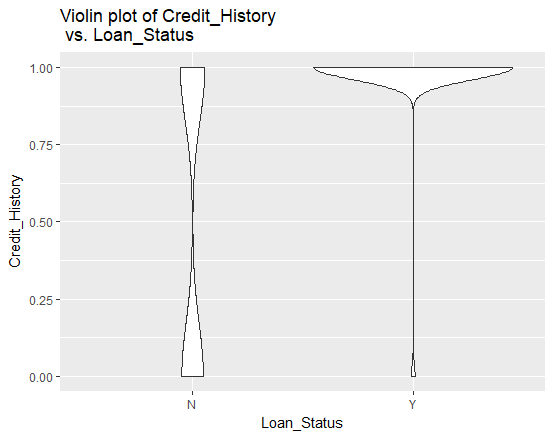
1. For the numeric variables box plots and violin plots were created against the Loan Status as shown below.

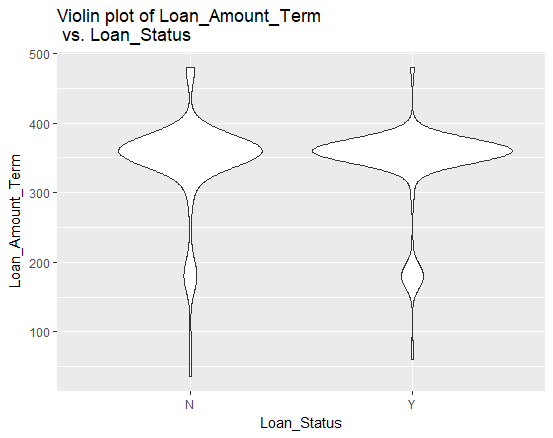
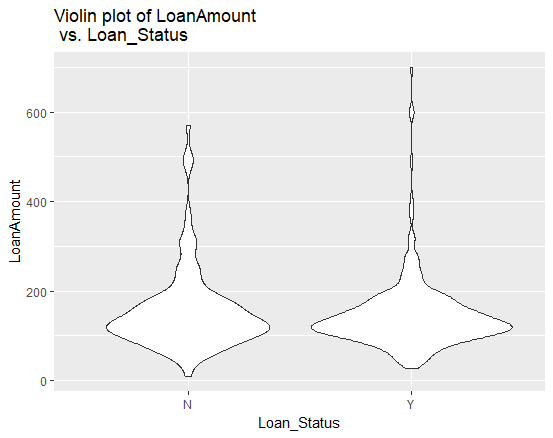
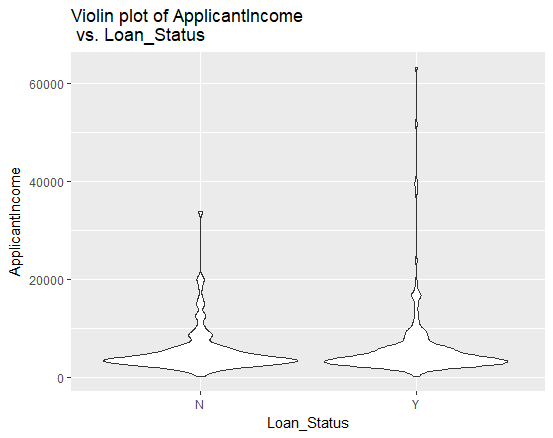
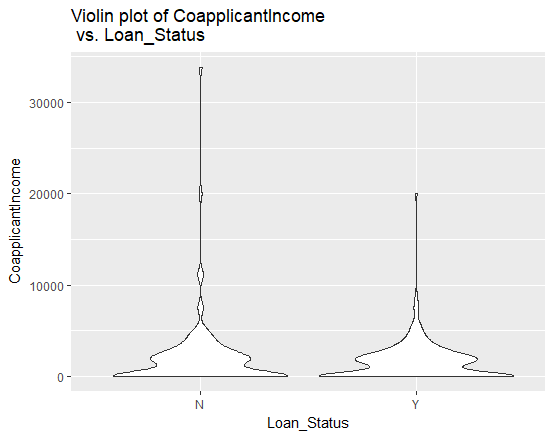
**Box Plots:**





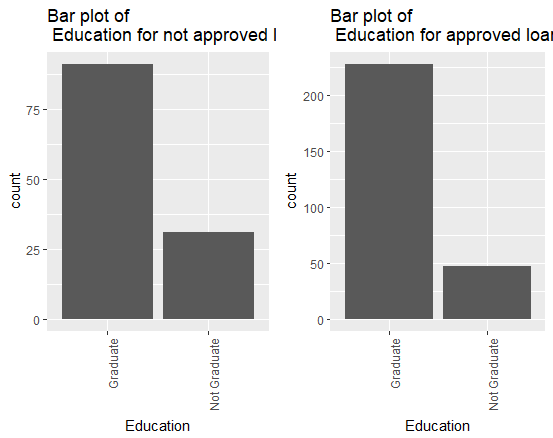
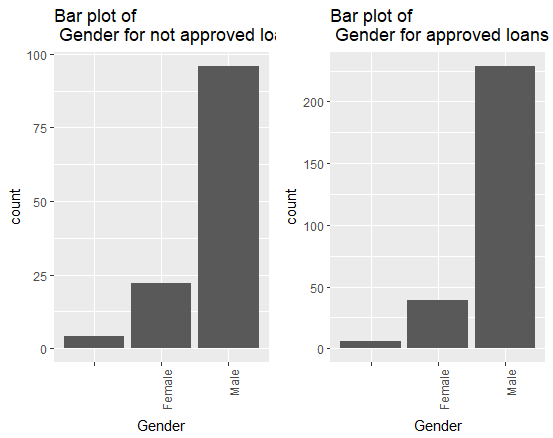
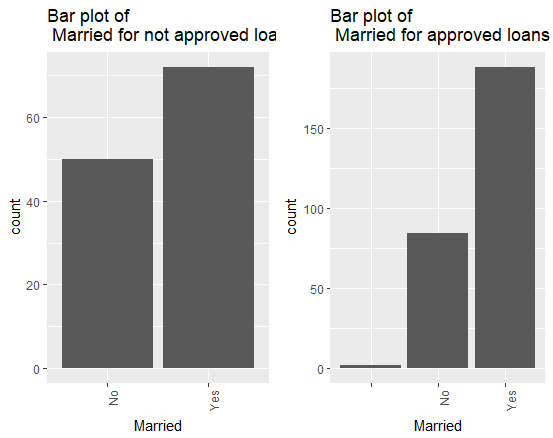
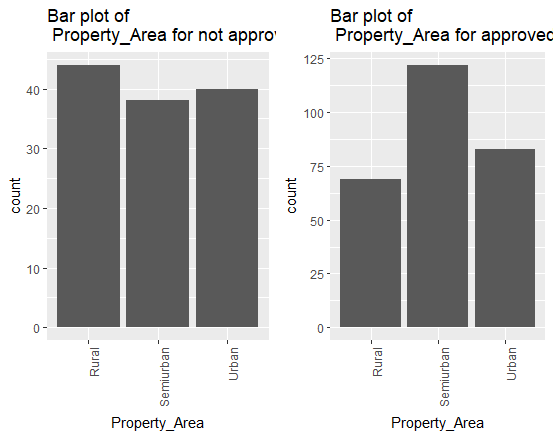
**Violin Plots:**

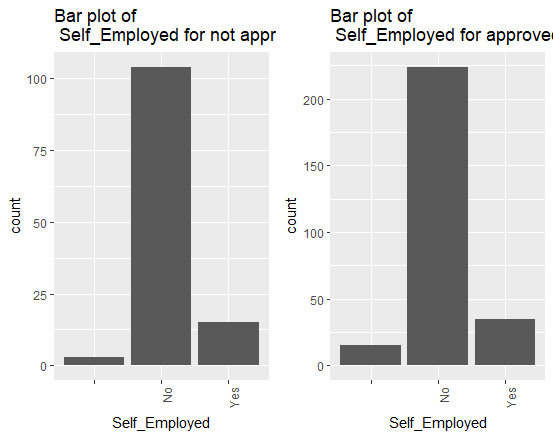




From the above box and violin plots it was concluded that the numeric predictors which affected the loan eligibility were Applicant Income, Coapplicant Income, Loan Amount, Loan Amount term, Credit History.

1. For categorical predictors, bar plot were plotted for approved and not approved loans.





From the above box and violin plots it was concluded that the categorical predictors which affected the loan eligibility were Marital Status, Property Area.\*

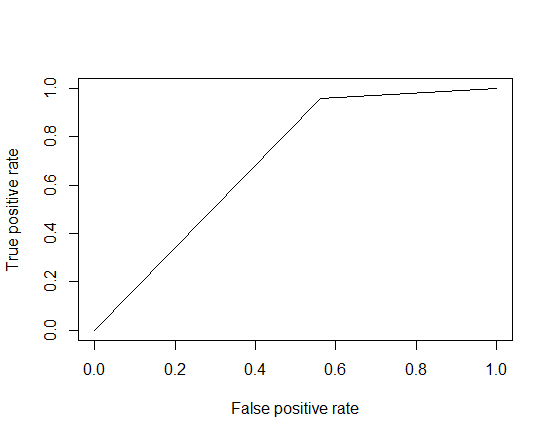
**Fitting the model**

1. From the above exploratory analysis, two logistics regression models were fitted, one with all the predictors (LG1) and one with the predictors which resulted in having a significant affect on the loan eligibility from the above exploratory analysis (LG2).

As mentioned earlier 75 % of the data was used to train these two models with a cutoff probability of 0.6 for LG1 and 0.7 for LG2. The area under the curve (AUC) for each Receiver Operating Characteristic (ROC) curve has also been mentioned.

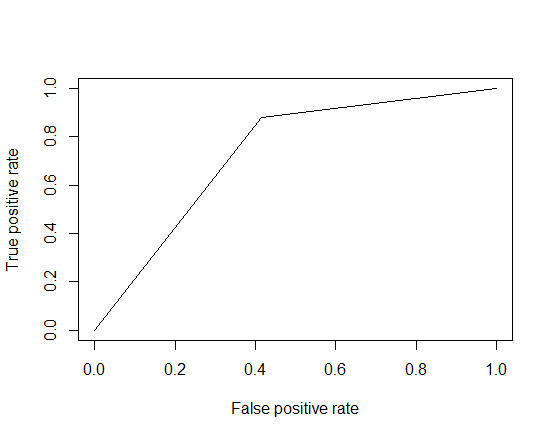
**ROC curve of LG1:**

AUC = .70



**ROC curve of LG2**

AUC = .73

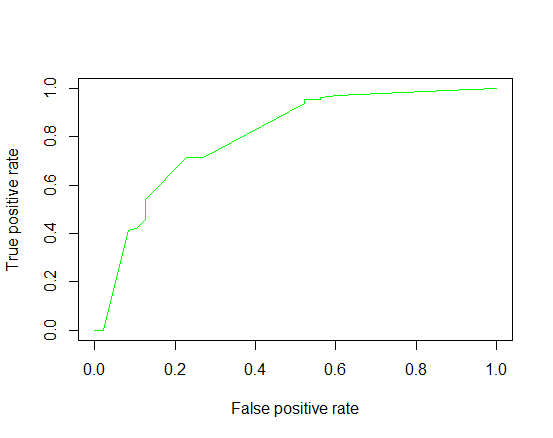


From the above two models LG2 model with an AUC of .73 for the ROC curve resulted in a better model to predict the loan eligibility.

1. A KNN model was also developed using cross validation to choose an optimized value of **k**. The model resulted in AUC of .84 for the ROC curve, with **k = 23** and Credit\_History and Property\_Area as the predictors.

**ROC curve for KNN model with k = 23**

AUC = 0.84



With an AUC of 0.84 KNN was concluded as the best model for predicting the loan eligibility of the given data.