MA5755: Data Analysis & Visualization

Project Presentation

Loan Default Prediction

Group-9:-
E Revathi Sri AM20S019
Nallam Venkata Sai Narasimha Tarun EE19B101
Nitin Yadav AM20M004

Contents

Introduction:

Problem

In this project, we want to know the chance that some customers will default their loan payment and use that as a parameter to decide whether to approve or disapprove the loan. Also, to identify and specifically target the customers segments, those are eligible for loan amount.

Objective

The goal of this project is to build a model that will classify if a certain customer will default his loan payment or not.

Dataset Description

The dataset used for this study has been taken from Kaggle. The data is from a finance company which deals with home loans.

Data pre-processing:

- The data set consist of 614 observations with 13 variables - Loan ID,Gender,Married,Applicant Income, Co-applicant income,Loan amount, Credit history and so on.
- Sample data:

Loan ID	Gender	Married	Dependent	Education	Self employed	Applicant Income	Co-applicant income	Loan amount	Loan term	Credit history	Property	Loan Status
LP001002	Male	No	0	Graduate	No	5849	0	128	360	1	Urban	Y

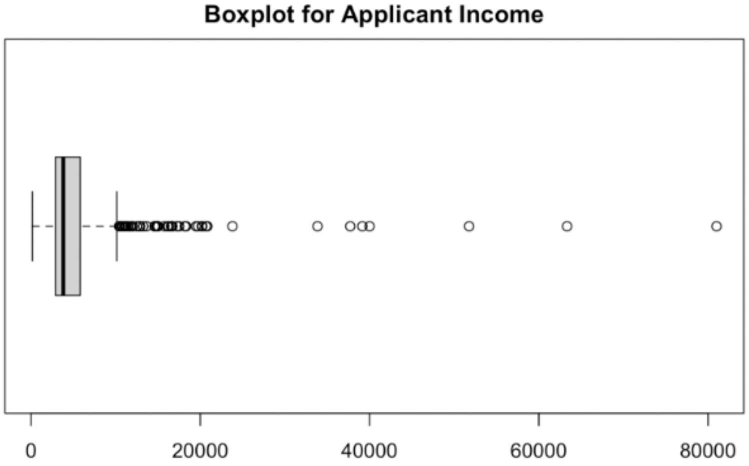
- The data consists of some blank spaces which are replaced by NAs so that R captures them as a missing number.

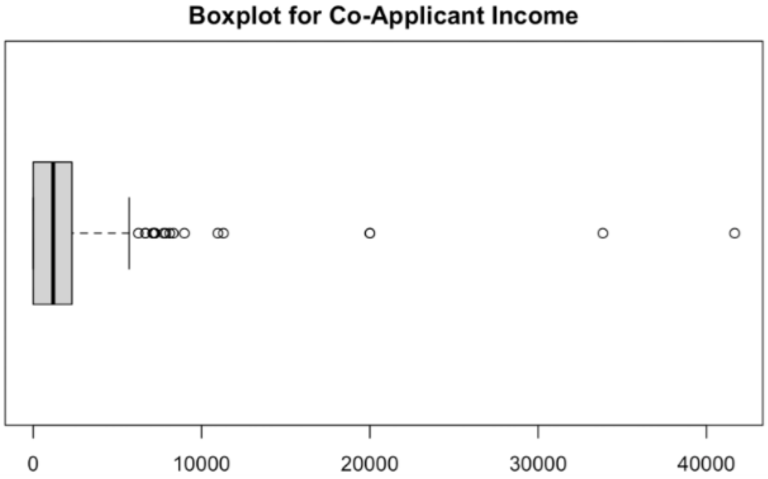
KNN Imputation method:

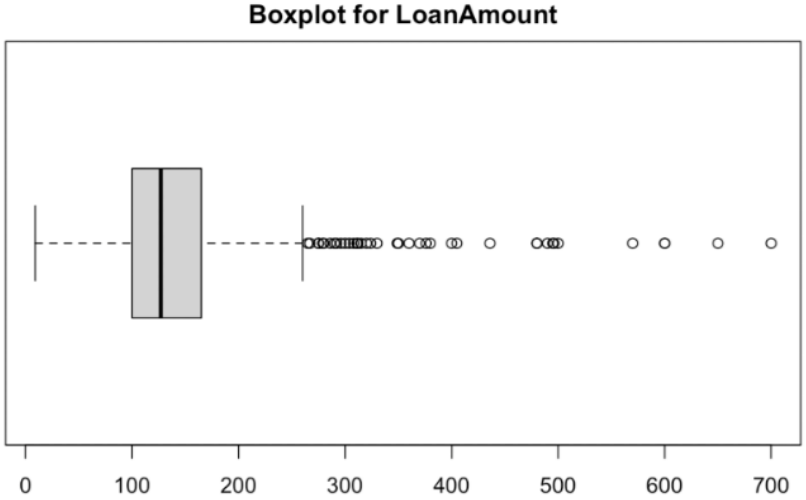
- The missing number has been handled using KNN Imputation method.
- KNN imputation method is a data transform to estimate the missing values.
- The default distance measure is a Euclidean distance and it will not include NAs while calculating distance between members of the training set.
- The columns with missing numbers are picked and KNN is applied to the variables with missing data.
- This creates a copy of the dataset with all missing values for each column replaced by an estimated value.
- Later a subset of data is made from the original dataset.
- The number of NAs in this new dataset are zero which gives a better dataset to train the models with.

Exploratory Data Analysis

Checking for Outliers

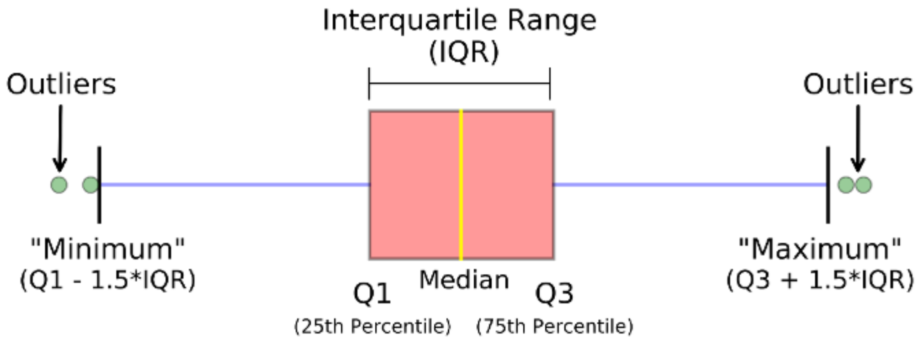




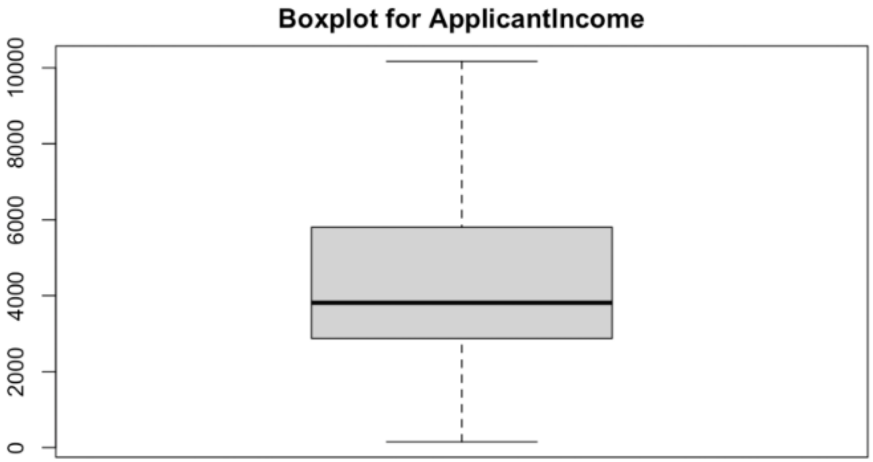


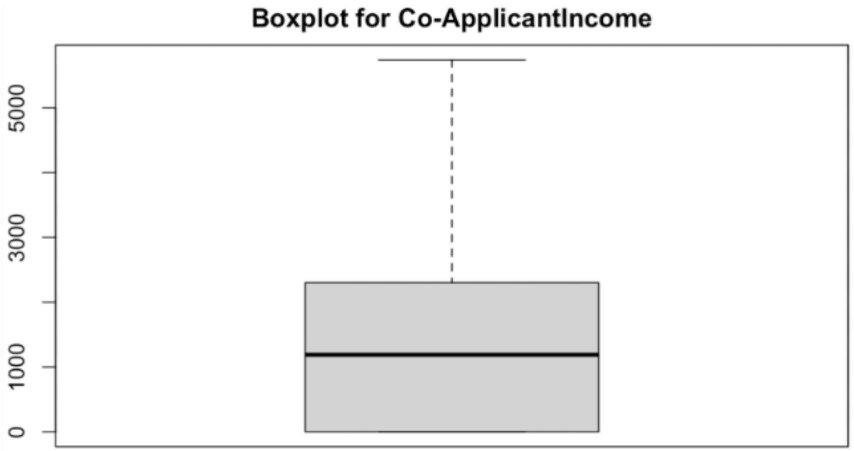
Outlier Treatment by Winsorization

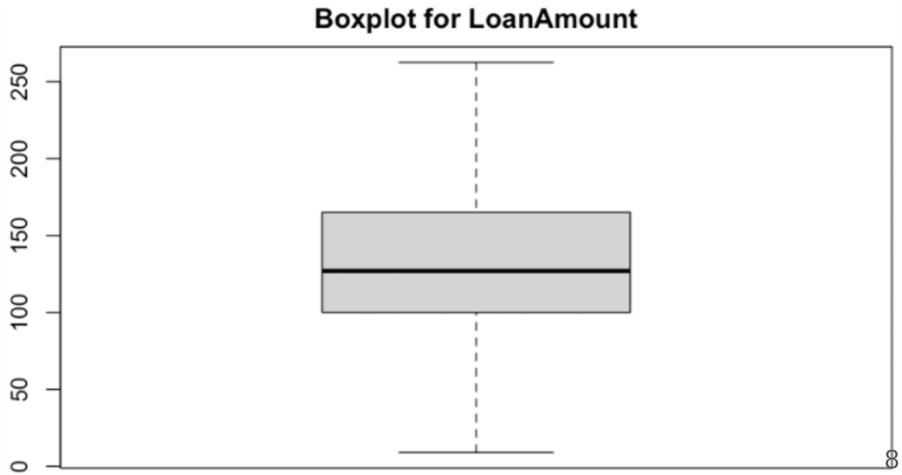
- Winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers.



- We used, $Benchmark = Q3 + 1.5(IQR)$ to bring down outliers within maximum limit


Boxplot for ApplicantIncome
0
2000
4000
6000
8000
10000


Boxplot for Co-ApplicantIncome
0
1000
3000
5000


Boxplot for LoanAmount
0
50
100
150
200
250

Logistic Regression

Confusion Matrix

	Predicted Value: FALSE	Predicted Value: TRUE
Actual Value: N	17	24
Actual Value: Y	4	97

Trained Model

```
Call:
glm(formula = Loan_Status ~ ., family = binomial, data = train_set[,
    -c(1)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5113  -0.2663   0.4840   0.6735   2.9870

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.263e+00  1.103e+00  -2.958  0.00310 **
GenderMale            -4.743e-01  3.628e-01  -1.307  0.19119
MarriedYes             8.746e-01  3.039e-01   2.878  0.00401 **
Dependents1           -6.002e-01  3.488e-01  -1.721  0.08529 .
Dependents2            3.587e-01  4.224e-01   0.849  0.39573
Dependents3+           1.747e-01  5.210e-01   0.335  0.73739
EducationNot Graduate -3.602e-01  3.124e-01  -1.153  0.24887
Self_EmployedYes       3.116e-01  4.024e-01   0.774  0.43869
ApplicantIncome        3.763e-05  7.908e-05   0.476  0.63414
CoapplicantIncome      1.344e-04  1.020e-04   1.318  0.18747
LoanAmount            -5.838e-03  3.418e-03  -1.708  0.08760 .
Loan_Amount_Term      -3.832e-04  2.173e-03  -0.176  0.86002
Credit_History         4.773e+00  6.185e-01   7.717 1.19e-14 ***
Property_AreaSemiurban 8.257e-01  3.198e-01   2.582  0.00981 **
Property_AreaUrban     3.806e-01  3.156e-01   1.206  0.22784
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 591.70  on 471  degrees of freedom
Residual deviance: 396.18  on 457  degrees of freedom
AIC: 426.18

Number of Fisher Scoring iterations: 5
```

Decision Tree Model

It is used to classify customers based upon the loan status.It also helps us to find the variables which are affecting the target variable.

Summary of decision tree from training data:

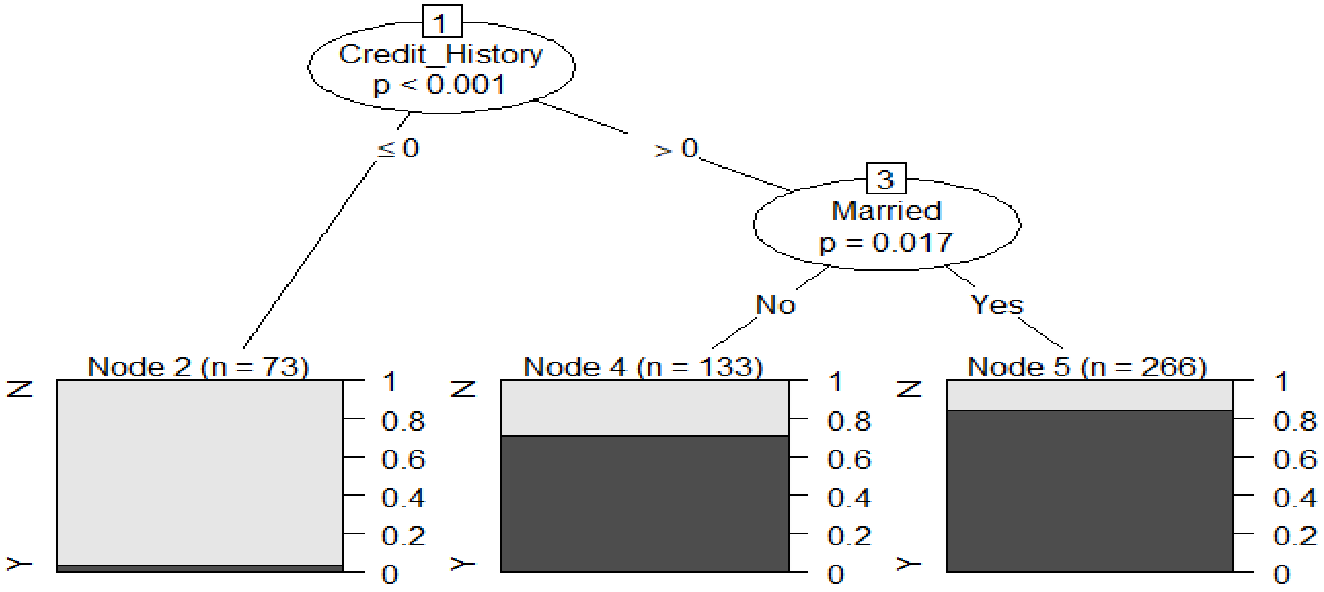
```
> Tree_Classifer

         Conditional inference tree with 3 terminal nodes

Response:  Loan_Status
Inputs:  Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantInco
 LoanAmount, Loan_Amount_Term, Credit_History, Property_Area
Number of observations:  472

1) Credit_History <= 0; criterion = 1, statistic = 161.718
  2)*  weights = 73
1) Credit_History > 0
  3) Married == {No}; criterion = 0.983, statistic = 10.498
    4)*  weights = 133
  3) Married == {Yes}
    5)*  weights = 266
```

- Decision Tree of Training data



- Applying the above decision tree to the test data and comparing it with the original data we get an accuracy of 80.28% for the model.

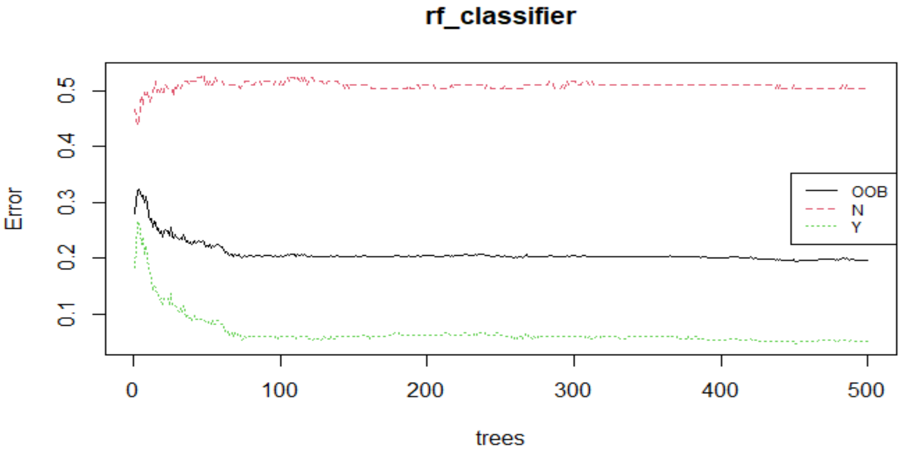
Random Forest Model

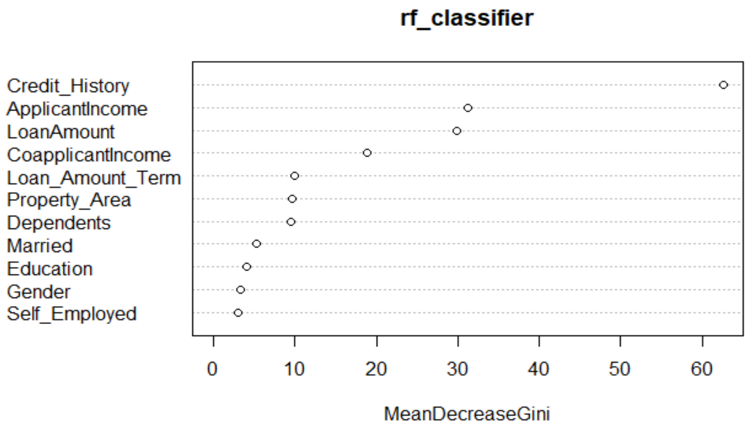
It combines multiple decision trees with flexibility resulting in a vast improvement in accuracy.The model chooses predictors randomly at the time of training.

Summary of Random Forest Model from training data

```
Call:
 randomForest(formula = Loan_Status ~ ., data = train_set[, -c(1)])
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 19.49%
Confusion matrix:
   N   Y class.error
N 75  76  0.50331126
Y 16 305  0.04984424
```





- The above fig represents the training error of the model

- The above fig represents the importance of the variables in the classifier model

- Applying the random forest model to the test data and comparing it with original data we get an accuracy of 80.28%.

Conclusion

Model	Logistic Regression	Decision Tree	Random Forest
Accuracy (%)	80.28	80.28	80.28
Most Important variable	Credit_History	Credit_History	Credit_History

Based on the performance of Logistics Regression, Decision Tree and Random Forest models we can conclude that if adequate pre-processing methods were carefully observed, then these models can perform extremely well on classification problem.

Appendix

1. MA5755 Project html link
2. MA5755 Project.nb.html

Thank you