Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

1) To predict that how many customers from 500 applications of loans are eligible for loan or not on the basis of their previous data available in the Credit-data-training dataset.
2) To build a training set by using credit-data –training dataset.
3) To build a best classification model to check the eligibility of every customer.

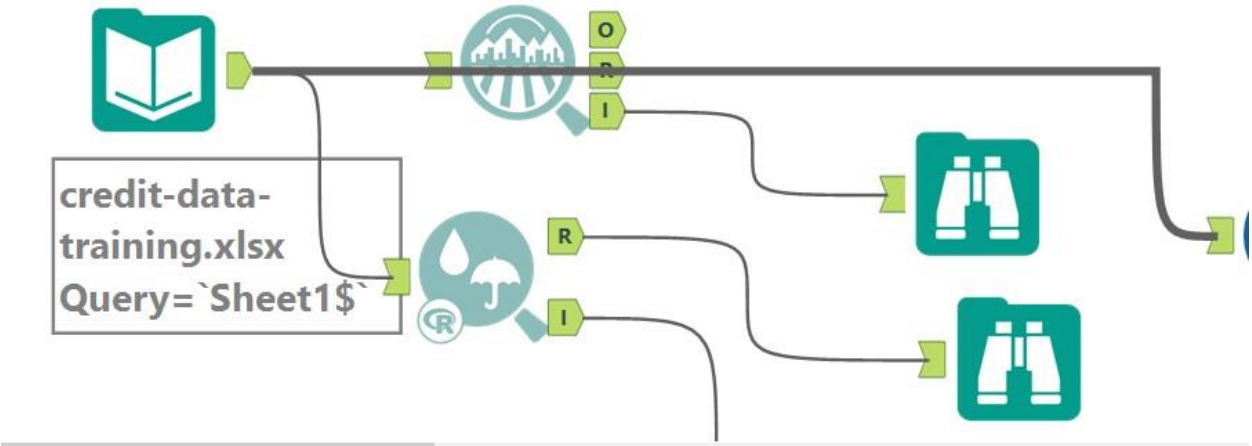
4) To find whether a person is eligible for a loan or not we need to use the previous data of each and every applicant regarding their bank account details.
Here we use Credit-data-training dataset to know about previous credit details.
And then use Customer-Score-data to check the score of creditworthiness of every individual.

5) There are only two options are here that is whether a customer will be creditworthy or not. So, in this case we will build a binary model.

Step 2: Building the Training Set

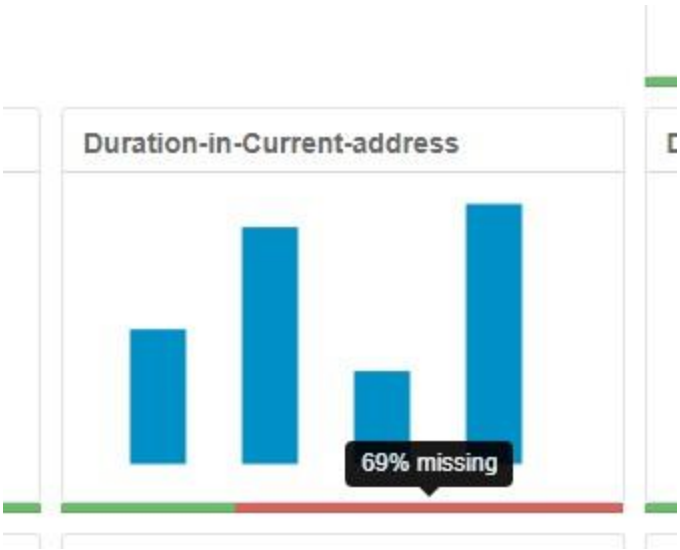
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String

Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

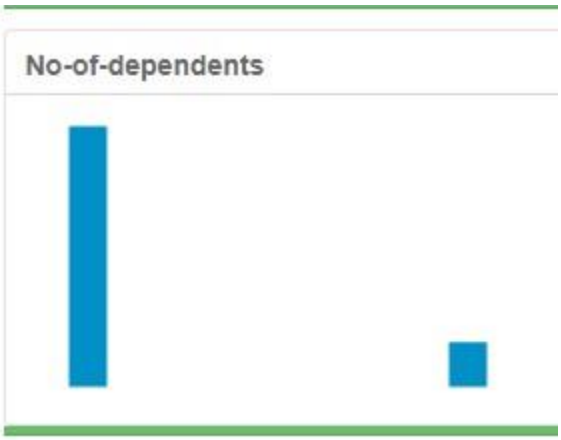


Removed Fields:

1) Duration in Current-address: There is 68% of missing data in this field that will affect the model.



2) Gurantors, Concurrent-Credits, Occupation, Foreign-workers, Type of apartments and No. of dependents have very low variability.



In total seven fields have been deleted due to low variability and missing data.
1) Guarantors
2) Occupation
3) Concurrent-address
4) Foreign-workers
5) No. of dependents
6) Telephone
7) Duration in current address

Field imputed: There were only one field that I decided to impute that was **Age in years**. Age was imputed by its median because there were many null values were present in this field including null values in the field can affect the modelling process so, to avoid this problem I have imputed all the null values with median value.
Median = 33
Age were skewed to one side.

In the final trained dataset there were only 13 fields available to built model.

Step 3: Train your Classification Models

- 1) Logistic Regression Model : In this model predictor variables are Account balance , purpose , Credit amount and payment status of previous credit as all these variables are statistically significant .

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

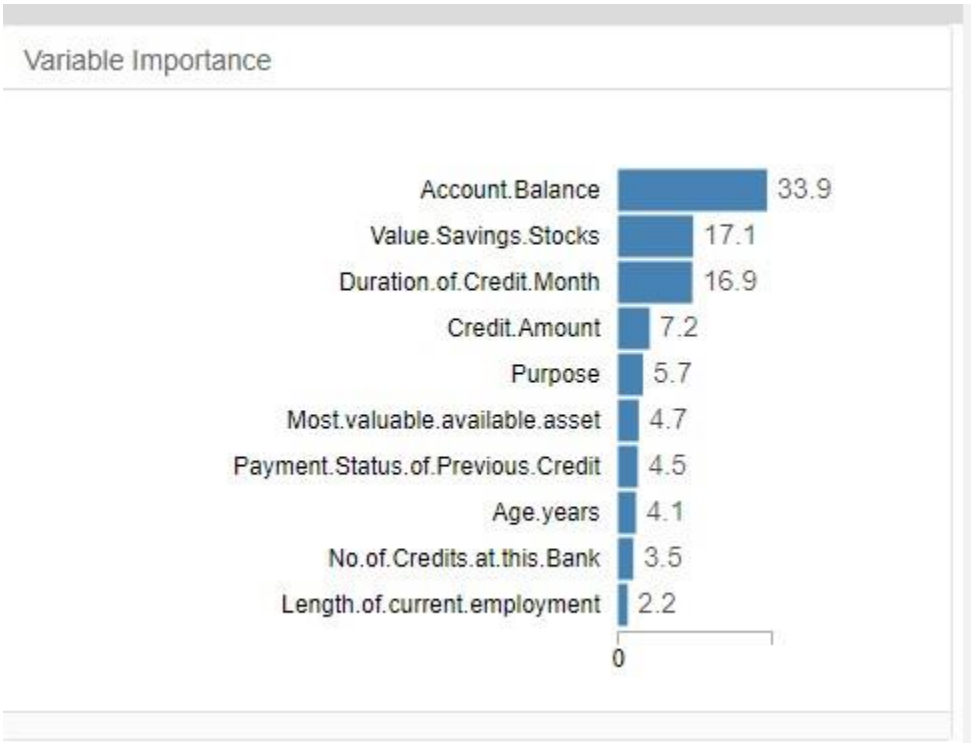
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183	*
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566	**
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042	
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596	*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

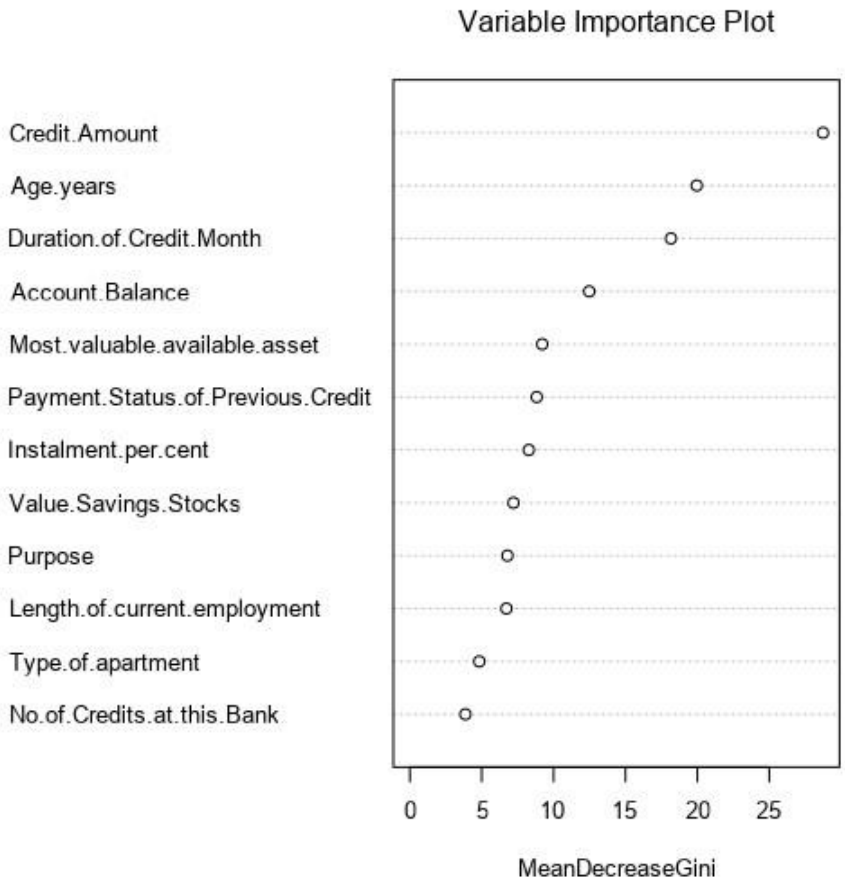
(Dispersion parameter for binomial taken to be 1)

2) Decision Tree model: according to this model best predictable variables are Account Balance, Value saving Stocks and Duration of credit month.

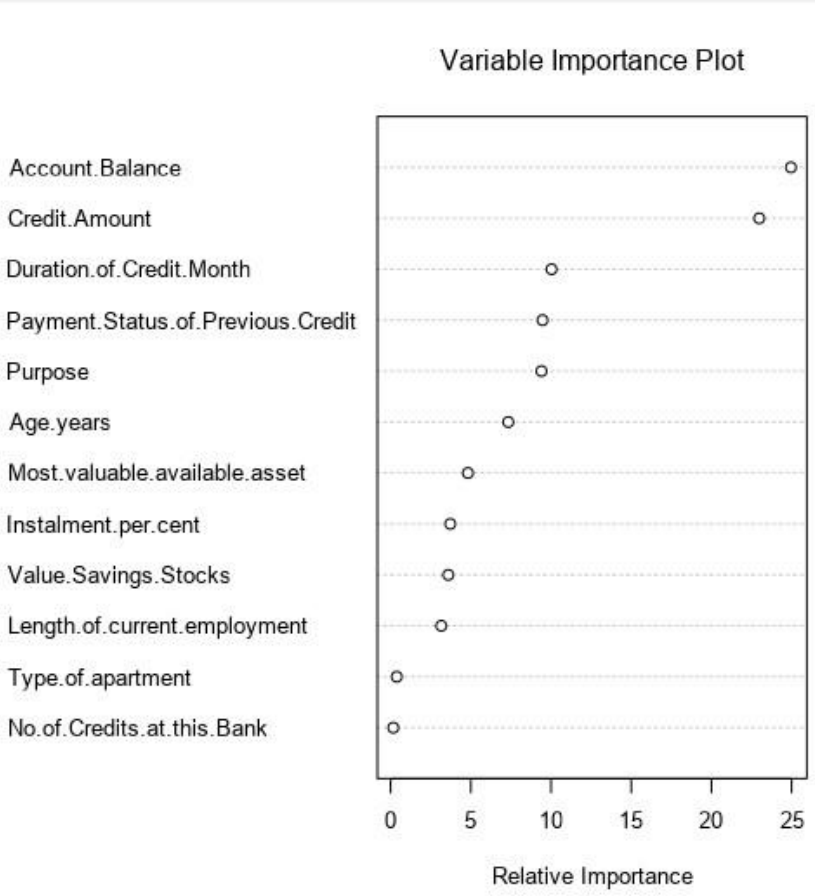


3) Random Forest Model: according to this model best predictor variables are Credit, Age-years, Duration of credit month, Account balance, Most valuable available dataset.

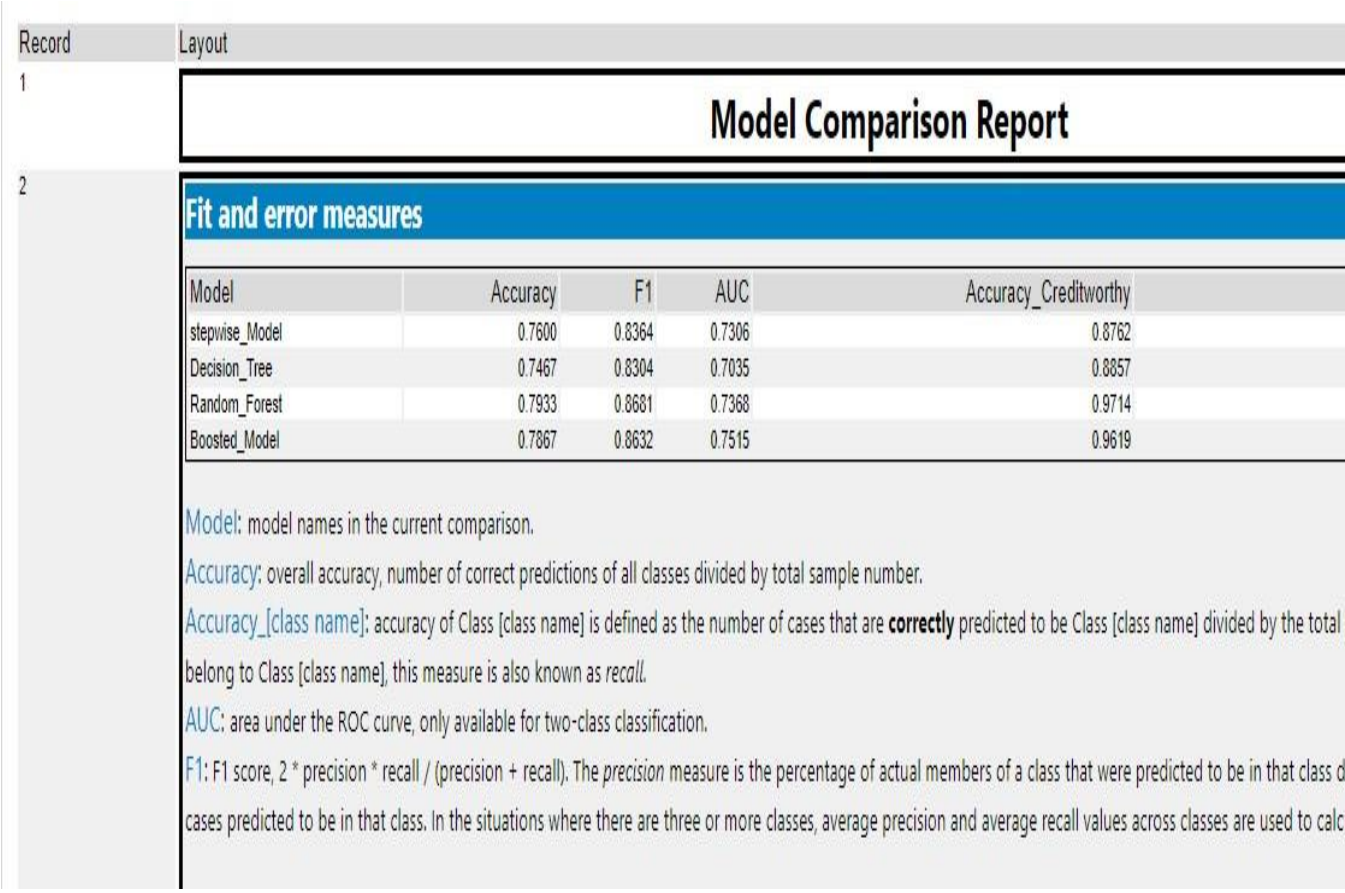
```
9
```



4) Boosted Model: According to this model best predictor variables are account Balance, Credit amount, Duration of Credit month, Payment status of previous credit.


Variable Importance Plot
Account.Balance
Credit.Amount
Duration.of.Credit.Month
Payment.Status.of.Previous.Credit
Purpose
Age.years
Most.valuable.available.asset
Instalment.per.cent
Value.Savings.Stocks
Length.of.current.employment
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
Relative Importance

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?
  According to model comparison report overall accuracy of all the models are as follows:
  1. X – stepwise model: 76%
  2. Decision Tree: 74%
  3. Random forest: 79%
  4. Boosted Model: 78%



Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy
stepwise_Model	0.7600	0.8364	0.7306	0.8762
Decision_Tree	0.7467	0.8304	0.7035	0.8857
Random_Forest	0.7933	0.8681	0.7368	0.9714
Boosted_Model	0.7867	0.8632	0.7515	0.9619

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total

belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class d

cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calc

Confusion Matrix of all the models:

In logistic Regression model, Boosted Model and Decision Tree model there were bias present in predicting actual credit worthy and actual non- credit worthy. But there were no bias present in Random forest Model.

1) The overall accuracy of stepwise model is 76%
   RPV = Actual_Creditwothy\ (Actual_Creditworthy + Actual_Non-Creditworthy)

= 92 / (92+23) =.80
NPV= Actual_Non-Creditworthy \ (Actual_Non-Creditworthy + Actual_Creditworthy)

=22/ (22+13)= .63
So after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

2) The overall accuracy of the Decision Tree Model is 74%

RPV = Actual_Creditworthy\(Actual_Creditworthy + Actual_Non-Creditworthy)

= 93\(93 + 26) = .78

NPV = Actual_Non-Creditworthy\(Actual_Non-Creditworthy + Actual_Creditworthy)

=19\(19 +12) = .61

So after checking the confusion matrix there is bias seen in the model's prediction to Non_Creditworthy.

3) The overall accuracy of the Boosted Model is 78%

RPV = Actual_Creditworthy\(Actual_Creditworthy + Actual_Non-Creditworthy)

= 101\(101 + 28)= .78

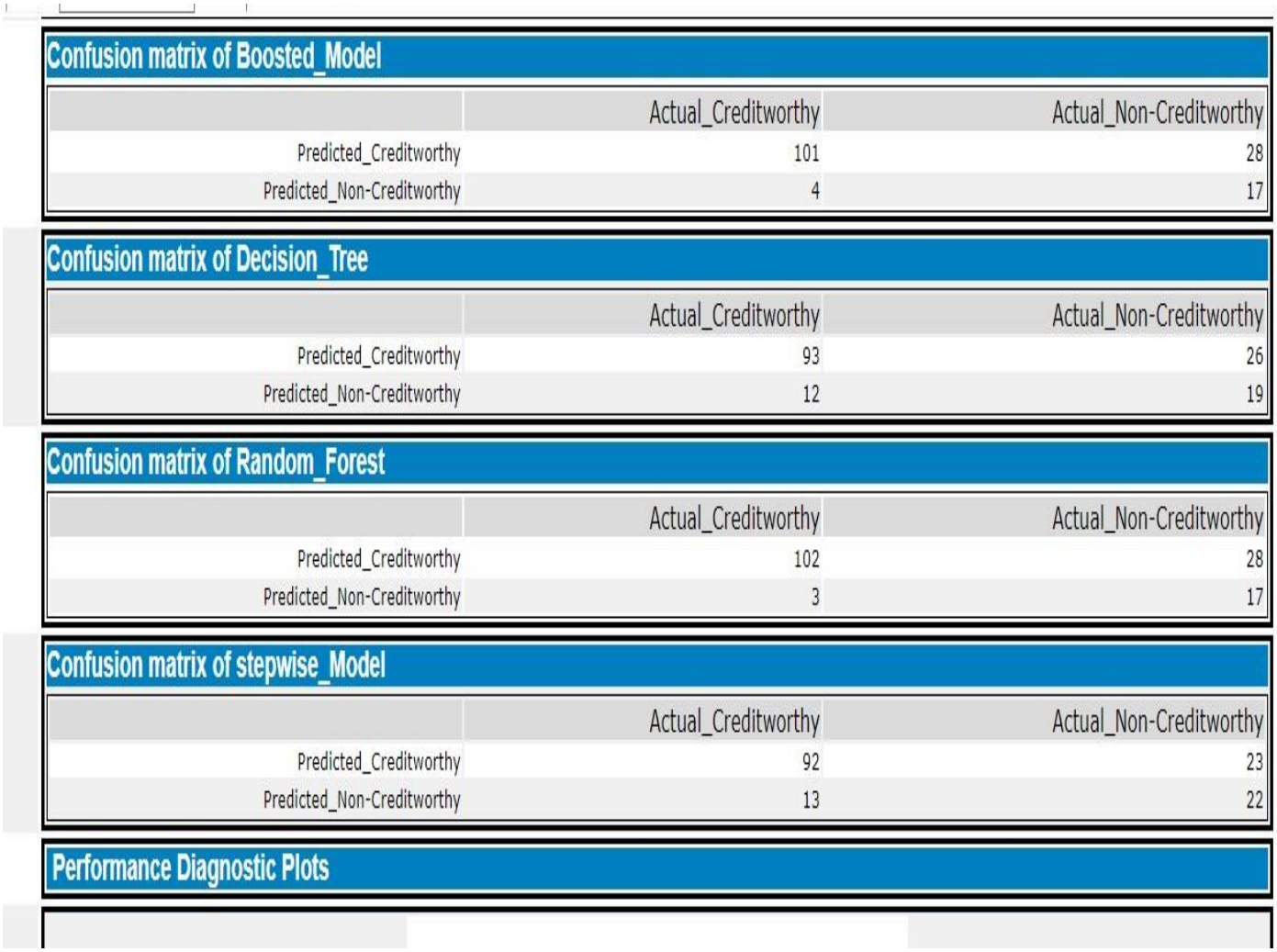
NPV = Actual_Non-Creditworthy\(Actual_Non-Creditworthy = Actual_Creditworthy)

= 17\(17 +4)= .80

So after checking the confusion matrix there is no bias seen in the model.

4) The overall accuracy of the Forest model is 79% which is strong
   PPV= Actual_Creditworthy \ (Actual_Creditworthy + Actual_Non-Creditworthy) = 102 / (102+28) =.78
   NPV= Actual_Non-Creditworthy \ (Actual_Non-Creditworthy + Actual_Creditworthy) = 17/ (17+3) = .85

So after checking the confusion matrix there is no bias seen in the model.



Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Random_Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of stepwise_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Performance Diagnostic Plots

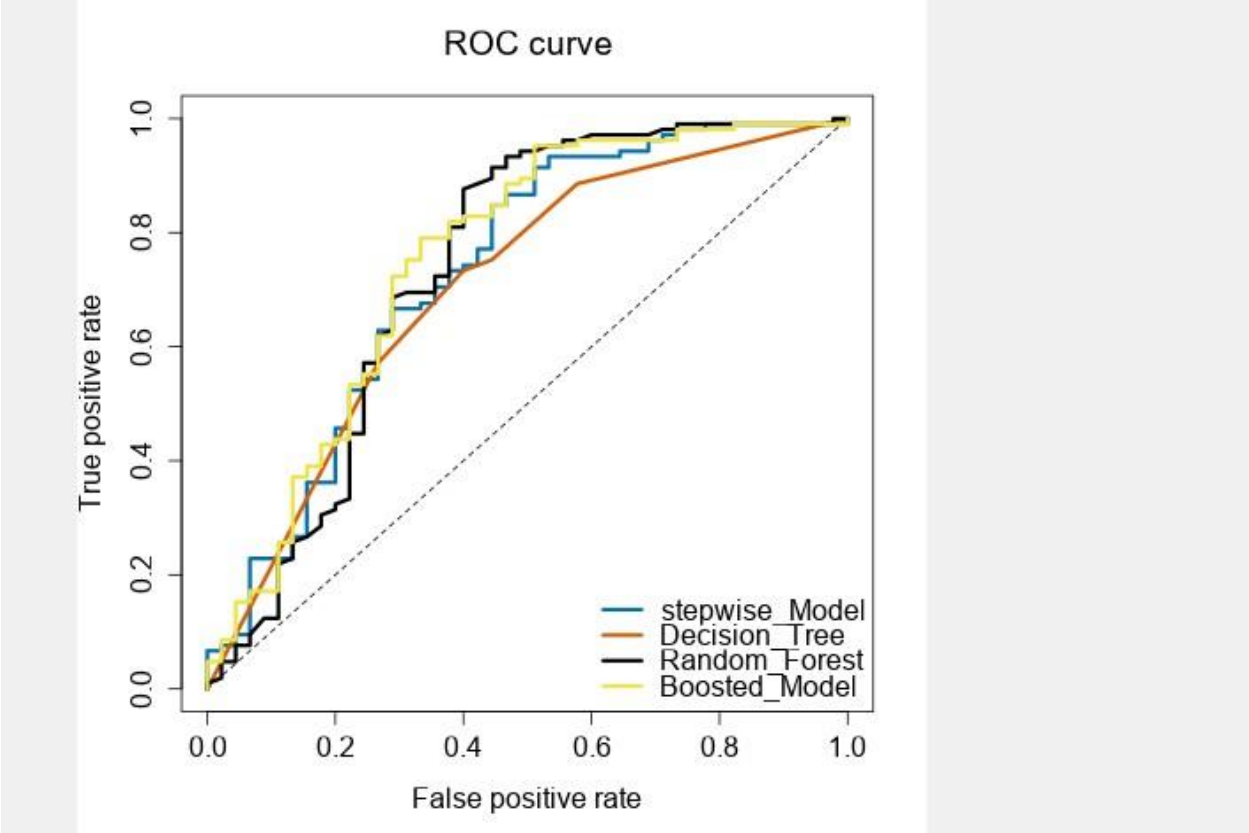
Step 4: Writeup

I decided to choose **Random Forest model.**

- Overall Accuracy against Validation set : In random forest model overall accuracy was 79% that was greater than accuracy rates of all the other models

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments: In Random forest Model accuracy within “Creditworthy” segment is .78 and in “Noncreditworthy” segment is .85. so there is no larger difference in both the segments .

- ROC graph:



- Bias in the Confusion Matrices: In random forest mode there were no bias present in the confusion matrix.
  PPV= Actual_Creditworthy \ (Actual_Creditworthy + Actual_Non-Creditworthy) = 102 / (102+28) =.78
  NPV= Actual_Non-Creditworthy \ (Actual_Non-Creditworthy + Actual_Creditworthy) = 17/ (17+3) = .85

So after checking the confusion matrix there is no bias seen in the model.

- **From 500 applications of loans 408 people were creditworthy.**

Final Workflow: