Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

This particular task involves the process of evaluating 500 loan applicants to determine the credit worthy individuals based on past metrics. Due to the influx of applications received by the bank in this week, physical processing might take a lot of time and resources, ultimately dissuading the applicants to some other banks. Therefore, a more robust approach involving classification using modelling techniques is required for fair evaluation of credit worthy individuals.

- What data is needed to inform those decisions?

The data needed for evaluating the potential credit worthy customers is the past data of applicants and the new data of customers seeking loan.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

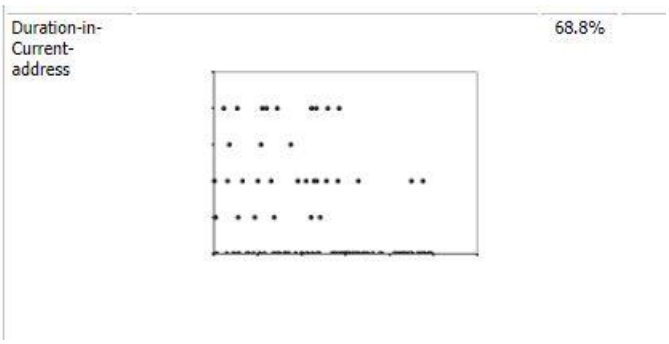
This problem is a clear case of Binary Classification where there are only two outcome:

1. Credit Worthy
2. Non-Credit Worthy

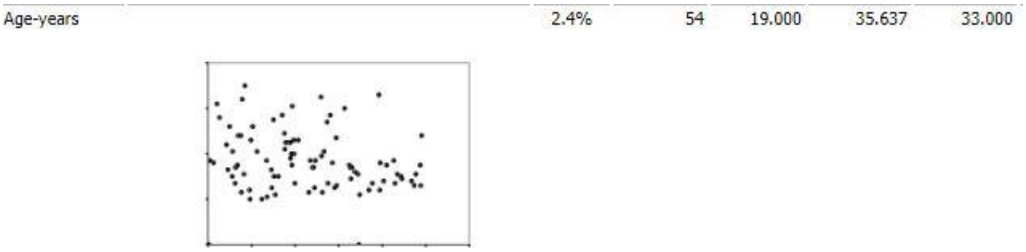
Step 2: Building the Training Set

- In your clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

In the clean-up process which involved drafting a field summary report, it was apparent that the Duration-in-Current-address has to be removed from the analysis. The missing data in this field was equivalent to 68.8% of the entire dataset which could create a lot of unnecessary bias if we decided to fix it with some sort of imputation.

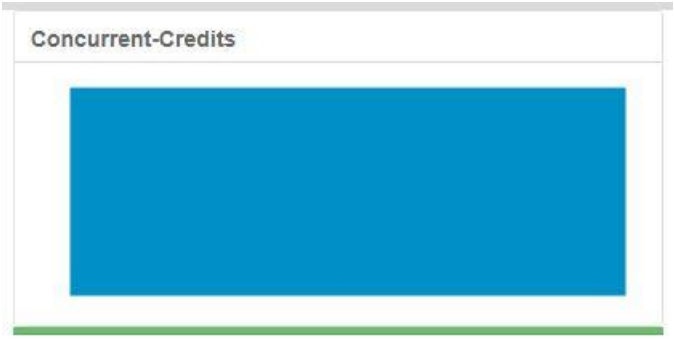


- The second table that needed cleansing was the Age-Years which had a missing value of 2.4%. This missing data was replaced with the median of the dataset (33) because the ages are normally distributed and has some values which tend to be outliers.

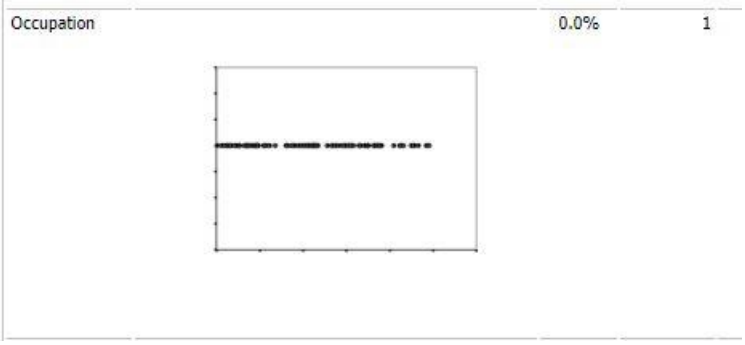


Low Variability:

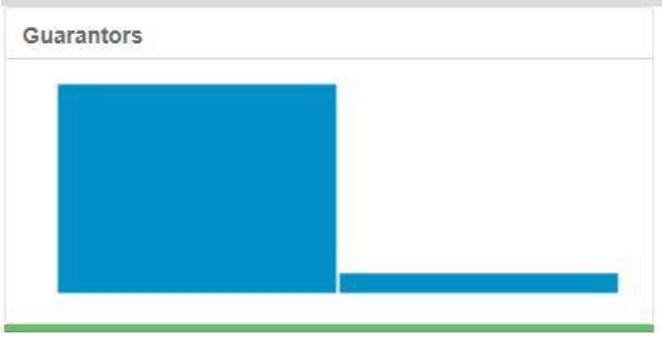
- Concurrent credits table is removed from the consideration due to a low amount of variability



- Occupation field only has one value

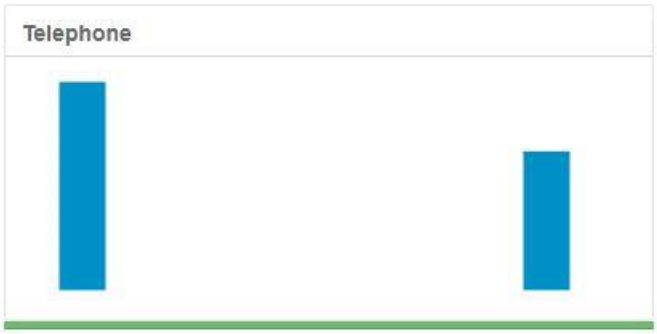


- Guarantors table is entirely skewed towards none

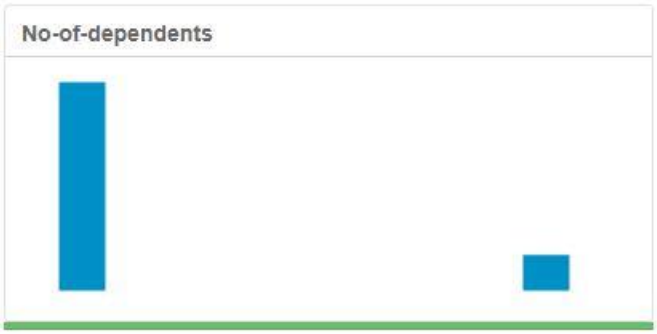


- Foreign workers data is skewed towards 1

- Telephone field doesn’t help to judge the credit worthiness of a applicant



- No-of-dependents is skewed towards the 1



Step 3: Train your Classification Models

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Stepwise Logistic Regression Model:

The predictor variables that are the most significant are the Account_Balance, Payment_Status, Purpose, Credit_Amount, Length_of_Current_Employment and Instalment_Percent.

Report

Report for Logistic Regression Model X

Basic Summary

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)

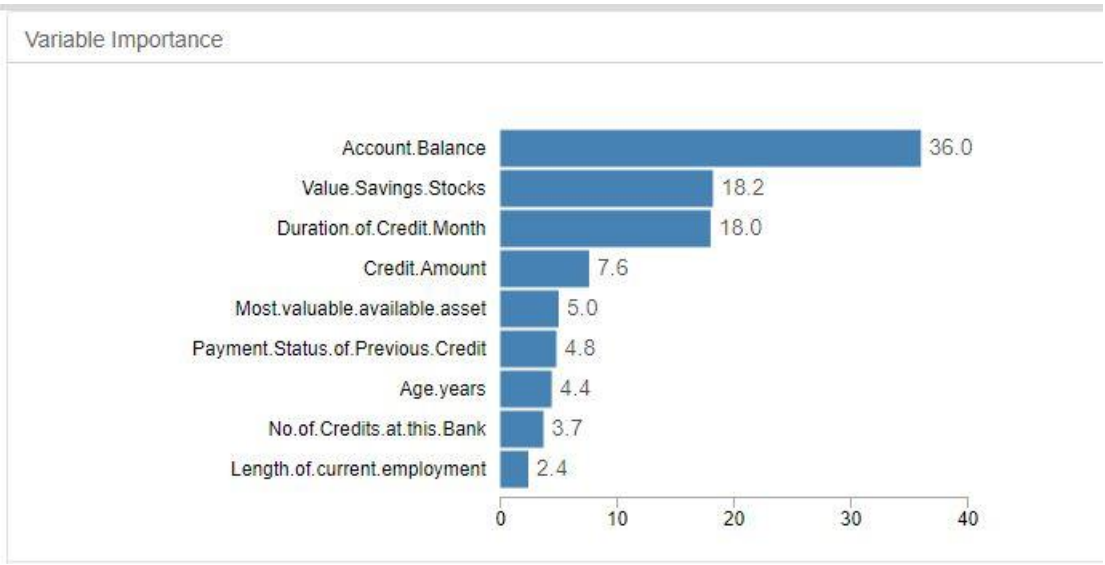
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

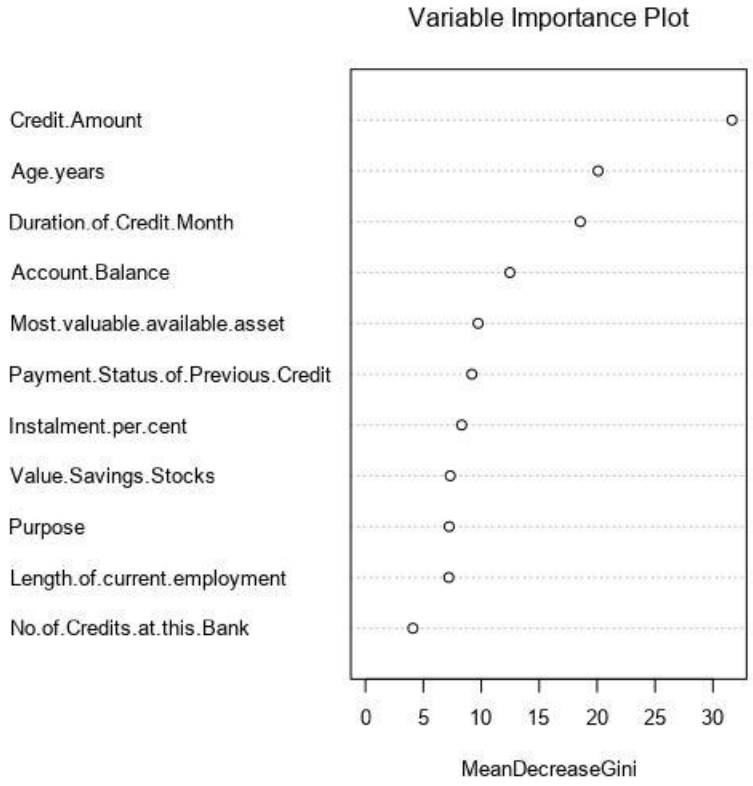
Decision Tree Model:

The top three predictor variables that are the most significant are the Account_balance, Value_Saving_Stocks and Duration_of_Credit_Month.



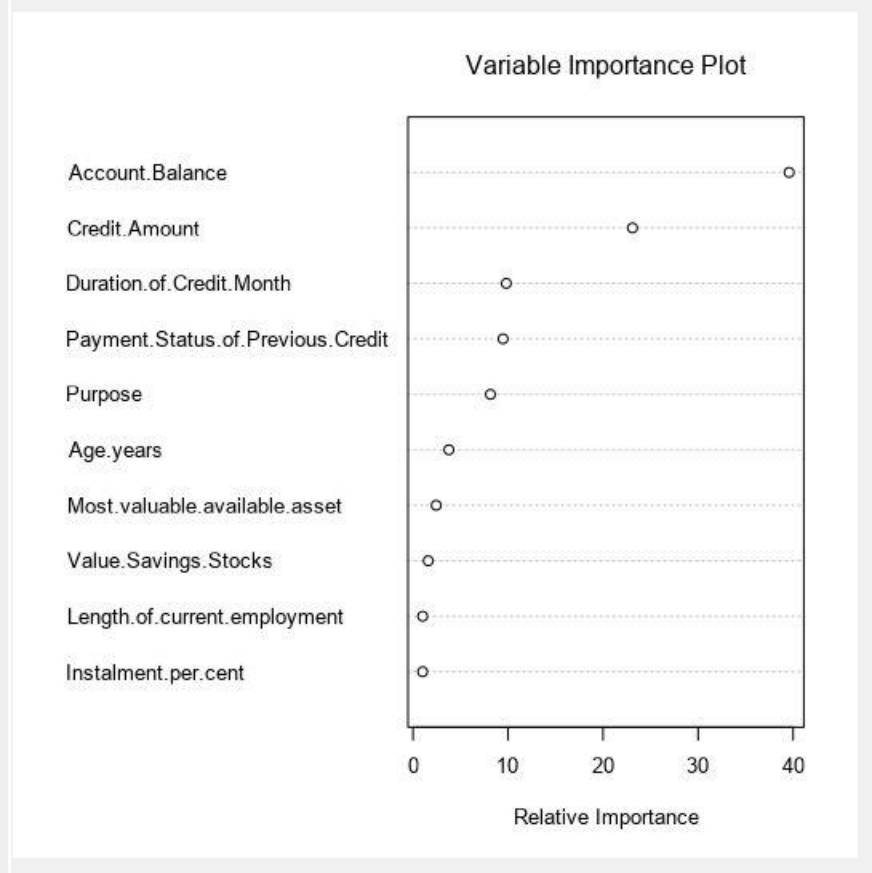
Forest Model:

The top three predictor variables that are the most significant are the Credit_Amount, Age_Yeards and Duration_of_Credit_Month.



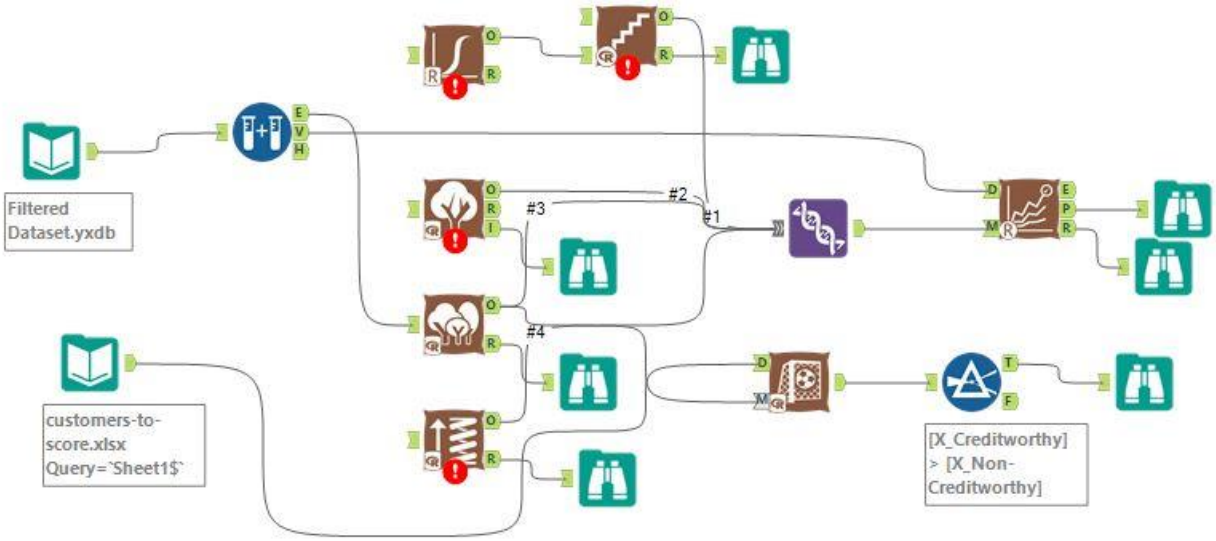
Boosted Model:

The top three predictor variables that are the most significant are the Account_balance, Credit_Amount and Duration_of_Credit_Month.



- 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?

Constructed Model:



Model Comparison Report for Stepwise Logistic Regression, Decision Tree, Forest Model and Boosted Model:

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_log	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Tree	0.8000	0.8707	0.7421	0.9619	0.4222
Boosted_forest_tree	0.7200	0.8306	0.7331	0.9810	0.1111

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 number of cases that actually 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 divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

The most accurate model is the Forest Model with an accuracy of 80%. As the dataset includes a lot more Credit worthy individuals in the past, there is an inherent bias which makes the model predict a lot more Creditworthy individuals than Non-Creditworthy ones.

Confusion Matrix plotted for each model:

Confusion matrix of Boosted_forest_tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	40
Predicted_Non-Creditworthy	2	5

Confusion matrix of Decision_Tree

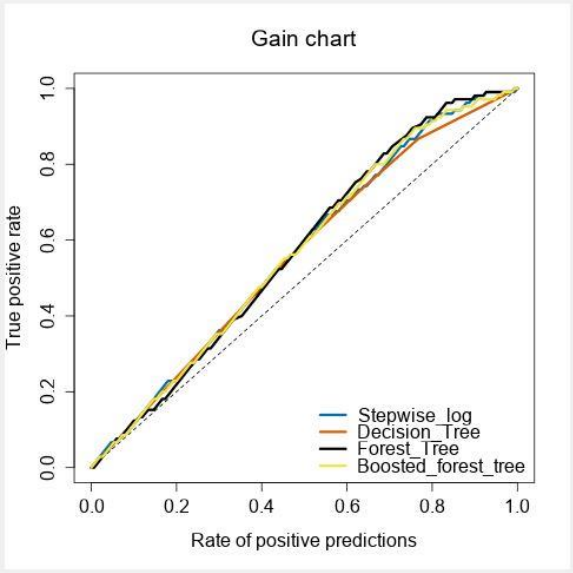
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

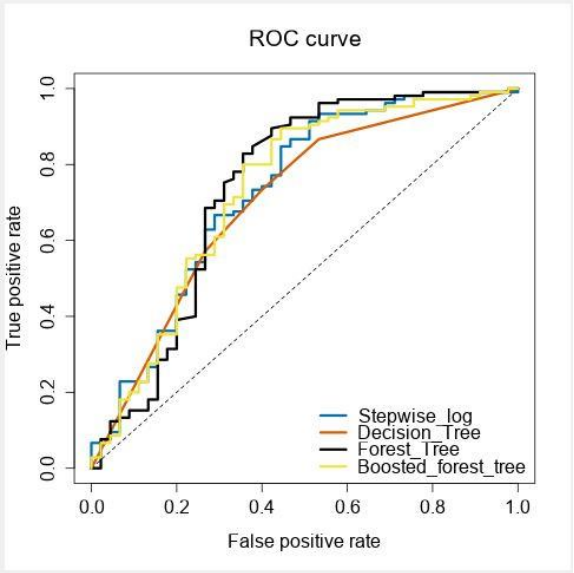
Confusion matrix of Forest_Tree

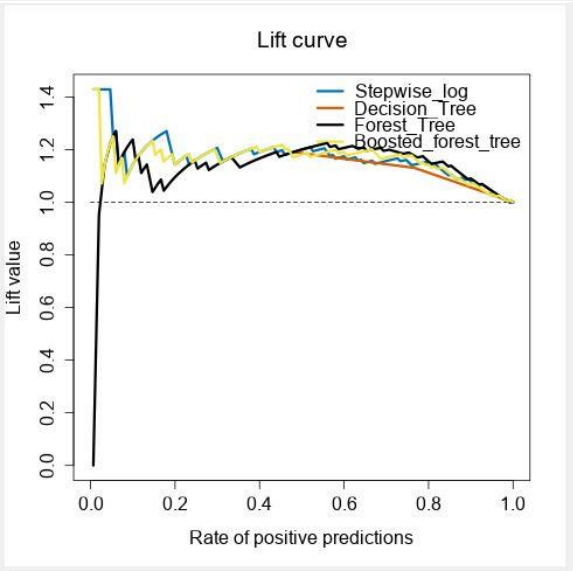
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

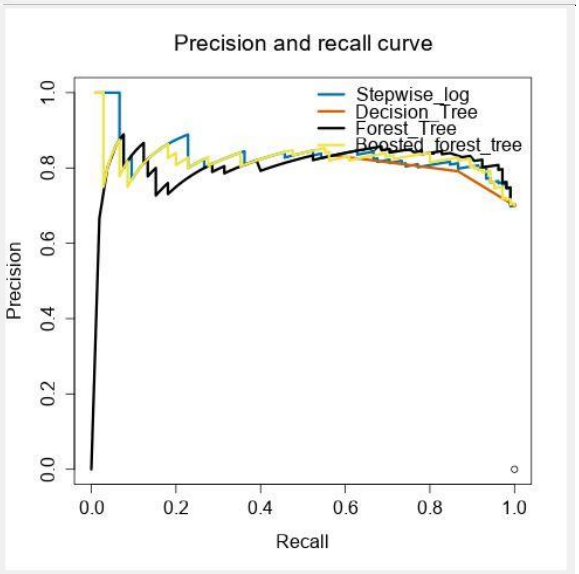
Confusion matrix of Stepwise_log

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









Step 4: Writeup

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    - ROC graph
    - Bias in the Confusion Matrices

After cross-validating the set against the validating set we observed that the accuracy of the forest model returns a value of 0.80, which is the highest amongst all the other models. The area under the curve also attains the max value of 0.7421, while the F1 score comes out to be 0.8707. This values when viewed holistically, gives us an clear understanding of the model’s ability to predict the most creditworthy individuals in a particular dataset.

Despite this accuracy there is an inherent bias, as the model is trained to predict a lot more creditworthy people than non-credit worthy ones. This makes the prediction of non-creditworthy individuals a fairly difficult task as the model isn’t trained on data that is neutral to both the types of individuals.

- How many individuals are creditworthy?

From this evaluation we observe that there are 408 Creditworthy applicants among the 500 loan applications received this week.