Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

  The decision needs to be made is "If the customers are creditworthy to give a loan to".

- What data is needed to inform those decisions?

  Data needed to inform those decisions are:

  Data on past applications Credit-Data-Training file) and list of customers (Customers to score file) such as **Credit-application-result, Account-balance, Duration-of-credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Guarantors, Duration-in-Current-address, Most-valuable-available-asset, Age-years, Concurrent-Credits, Type-of-apartment, No-of-Credits-at-this-Bank, Occupation, No-of-dependents, Telephone, Foreign-Worker.**

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

  Since the response of the decisions needs to be made is "yes or no", we are dealing with Binary.

Step 2: Building the Training Set

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

According to the analysis performed on the numerical variables, it appears that there are no variables highly correlated with each other (Higher than 0.70).

```
Out[30]: <function matplotlib.pyplot.show(*args, **kw)>
```

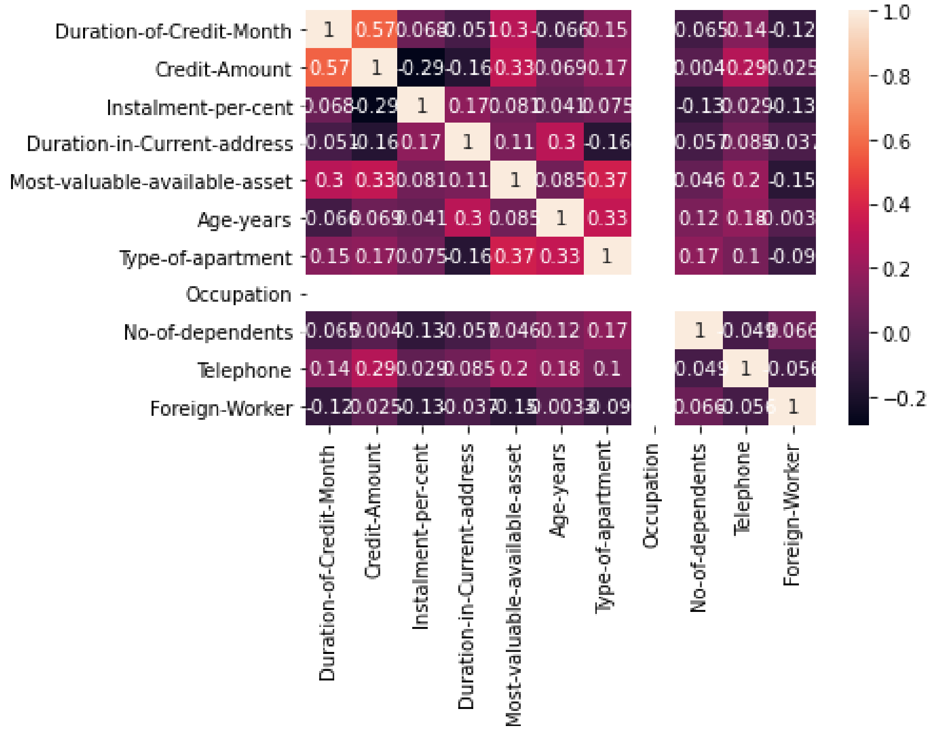


Figure 1: Correlation matrix of variables.

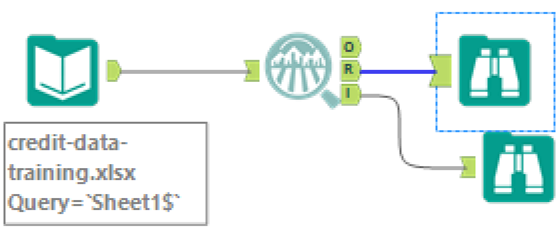


Figure 2: Alteryx workflow for Field summary analysis

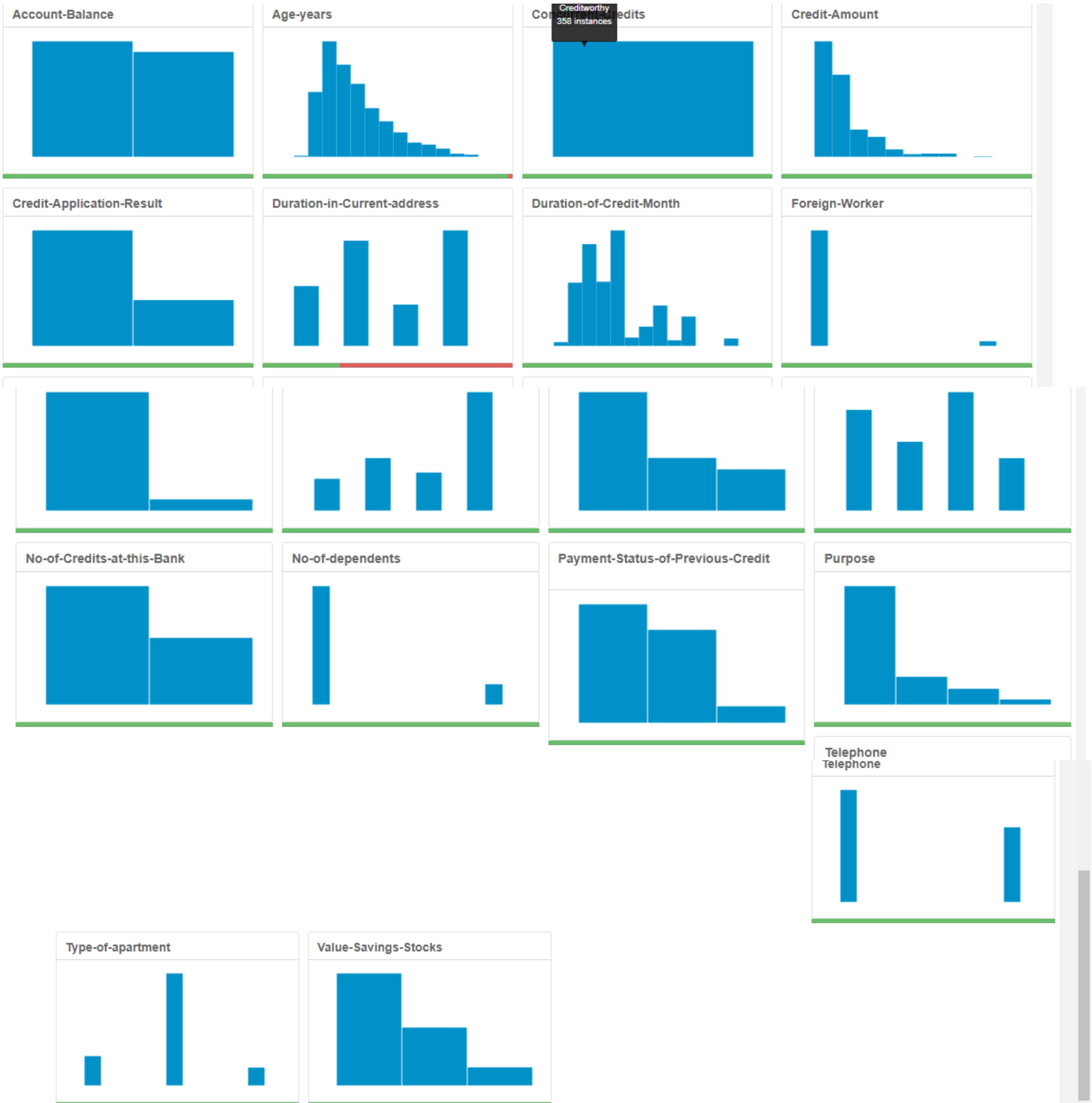
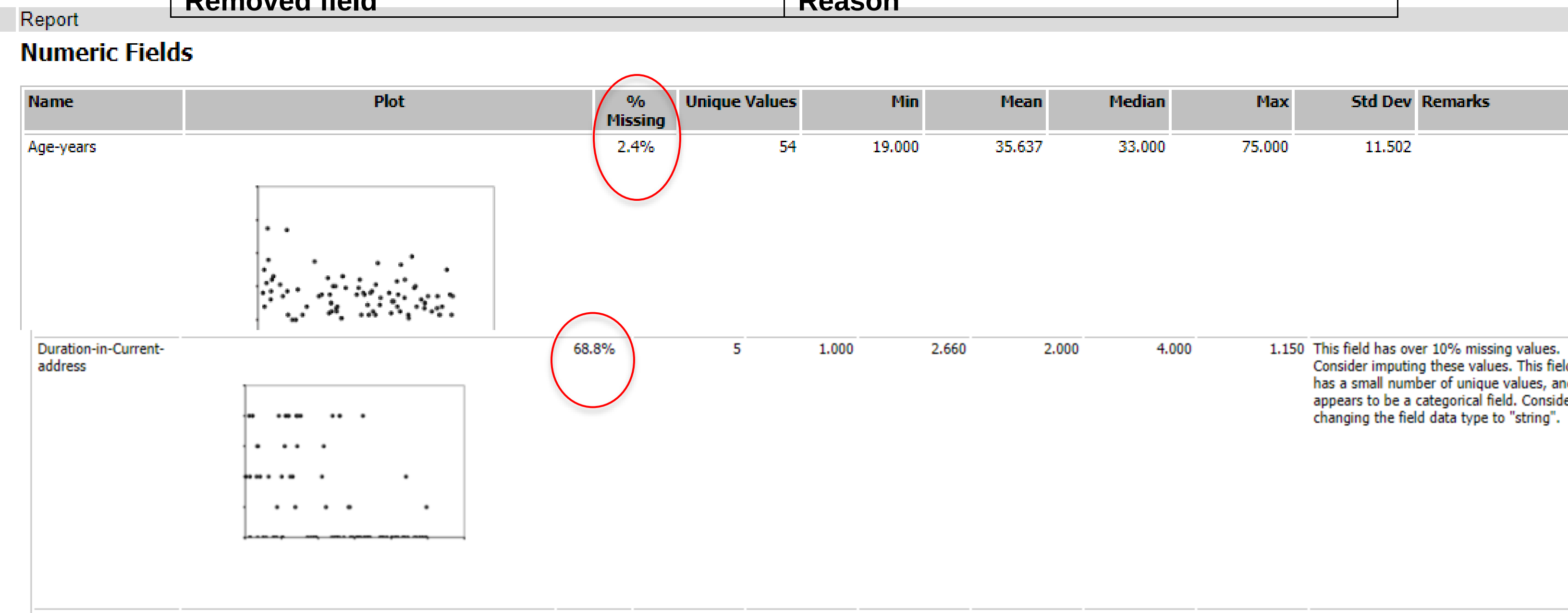


Figure 3 : Summary fiel of all variables.

Figure 4: Field summary showing missing data



Report

Numeric Fields

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This fiel has a small number of unique values, an appears to be a categorical field. Conside changing the field data type to "string".

Removed field	**Reason**
Duration in Current Address	69 % missing data
Concurrent credit	Only one type of data for entire field
Occupation	Only one type of data for entire field
Telephone	Irrelevant data for determining the creditworthiness of customers
No of Dependents	Low variability; heavily skewed
Guarantors	Low variability; heavily skewed
Foreign workers	Low variability; heavily skewed

Age Years has 2.4% missing data so it is appropriate to impute the missing data with the median age.

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

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

- 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?

1. Logistic Regression

Figure 5: Summary Report for stepwise Logit

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

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.

Confusion matrix of Stepwise_log

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

Report for Logistic Regression Model Credit_worthiness

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

Figure 6: Model comparison Report for stepwise Logit

According to the Logistic Regression-Stepwise reports, **the Credit Application Result is the target variable and Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of current employment and Instalment percent** are the 4 most significant predictor variables with p-value of less than 0.05. The R-squared with the value of 0.248 sounds not good at all with a value of 0.2048.

The model comparison report for the logit shows an overall accuracy of 76.0%. Though the accuracy for creditworthiness is high at 87.2%, the accuracy of non-creditworthiness is low at 48.9%. Then, the model is biased towards predicting customers as creditworthy.

2. Decision tree

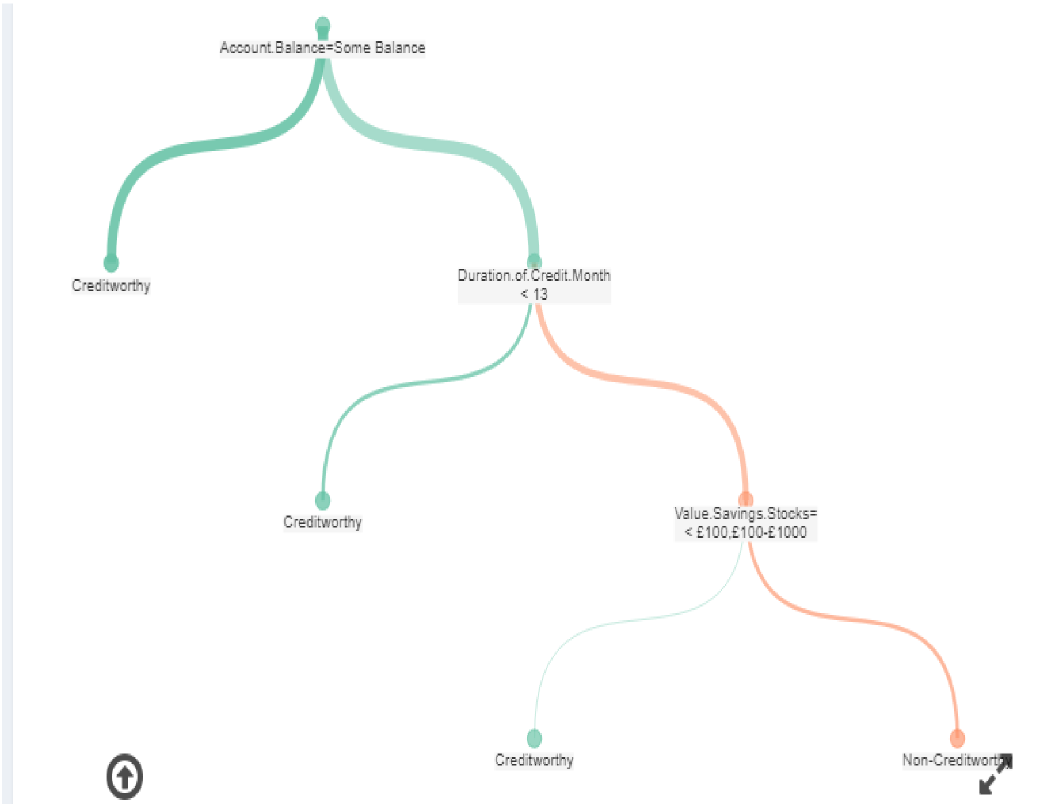


Figure 7: Decision Tree

Figure 8: Summary Report for Decision Tree

Summary Report for Decision Tree Model Credit_worthy_decision_tree

Call:
rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Value.Savings.Stocks, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary

Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.0687285	0	1.00000	1.00000	0.086326
2	0.0051546	3	0.79381	0.83505	0.081342

Leaf Summary

node), split, n, loss, yval, (yprob)
 * denotes terminal node

1) root 350 97 Creditworthy (0.7228571 0.2771429)
 2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
 3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
  6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
  7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
   14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
   15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Credit_worthy_decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667

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.

Confusion matrix of Credit_worthy_decision_tree

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

Figure 9: Model comparison Report for Decision Tree

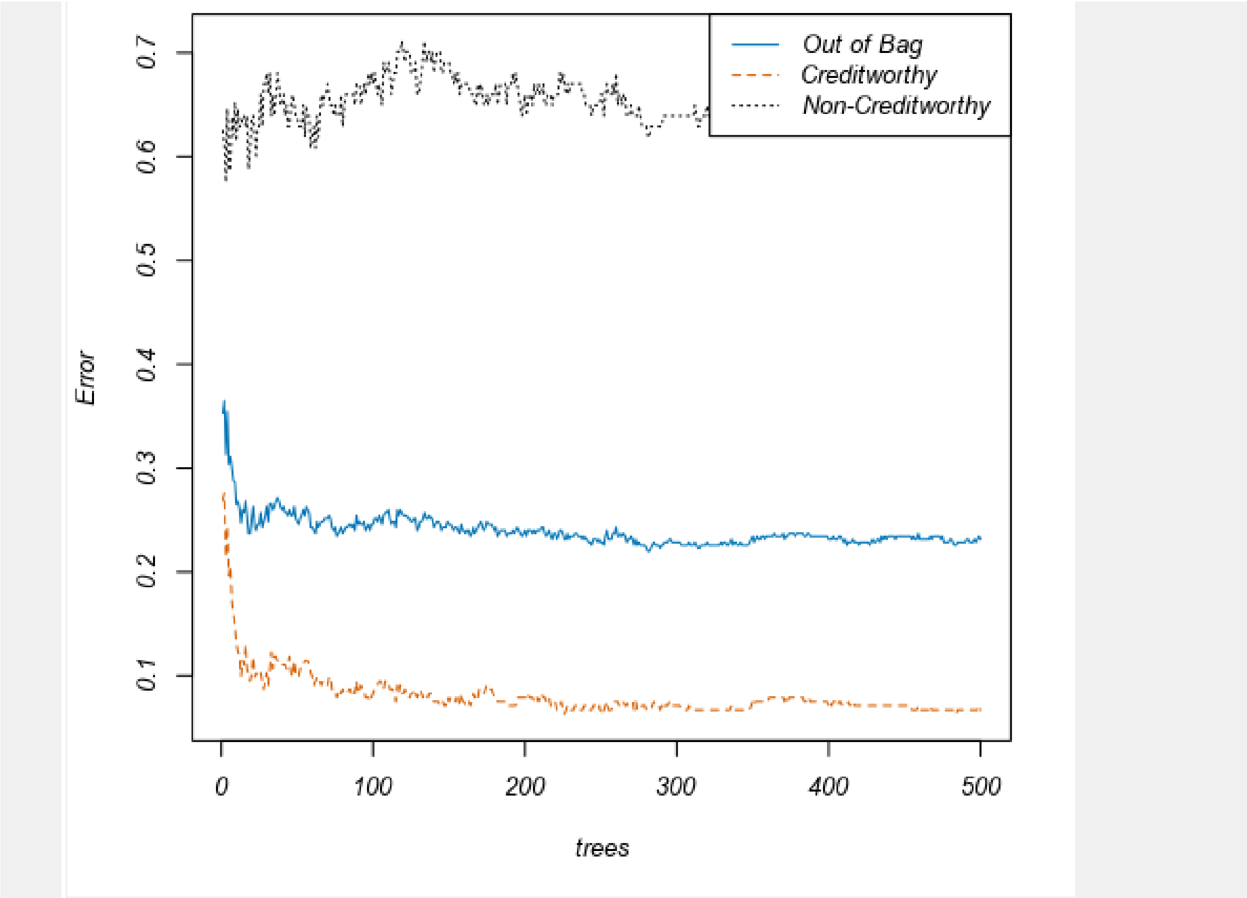
Testing the decision tree model into our dataset, we could see that even if the Root node error is quite high it still under 30%, which is consider as an acceptable error.

When we are validating our model against itself, with the confusion Matrix, we can see that the sum of accuracy is 78%, classifying it as a reliable model.

Using Credit Application Result as the target variables, Account Balance, Value Savings Stocks and Duration of credit Month are the 3 most important variables. The overall accuracy is 74.7%.

Accuracy for creditworthy is 86.7% while accuracy of non-creditworthy is 46.7%. The model seems to be biased towards predicting customers as non-creditworthy.

3. Forest Model



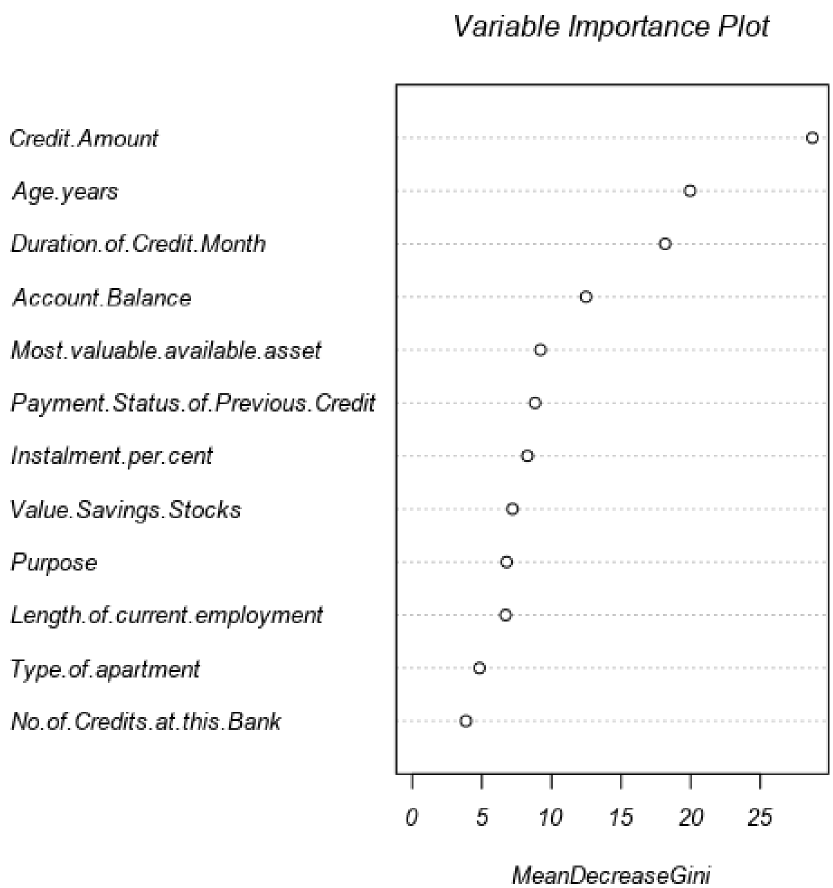


Figure 10: Percentage Error for Different Number of Trees and Variables Importance plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_model	0.7933	0.8681	0.7368	0.9714	0.3778

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.

Confusion matrix of Forest_model

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

Figure 11: Model comparison report for Forest Model

Credit Application Result is the target variable and Credit Amount, Age years, and Duration of credit Month are the 3 most significant variables.

The model comparison report for the forest Model shows that this model has an overall accuracy of 80.0%. The accuracy for creditworthiness is 79.5% whereas the credit accuracy of non-creditworthiness is 82.6%. The accuracies are comparable and the model is not biased in its predictions for creditworthiness of customers

4. Boosted Model

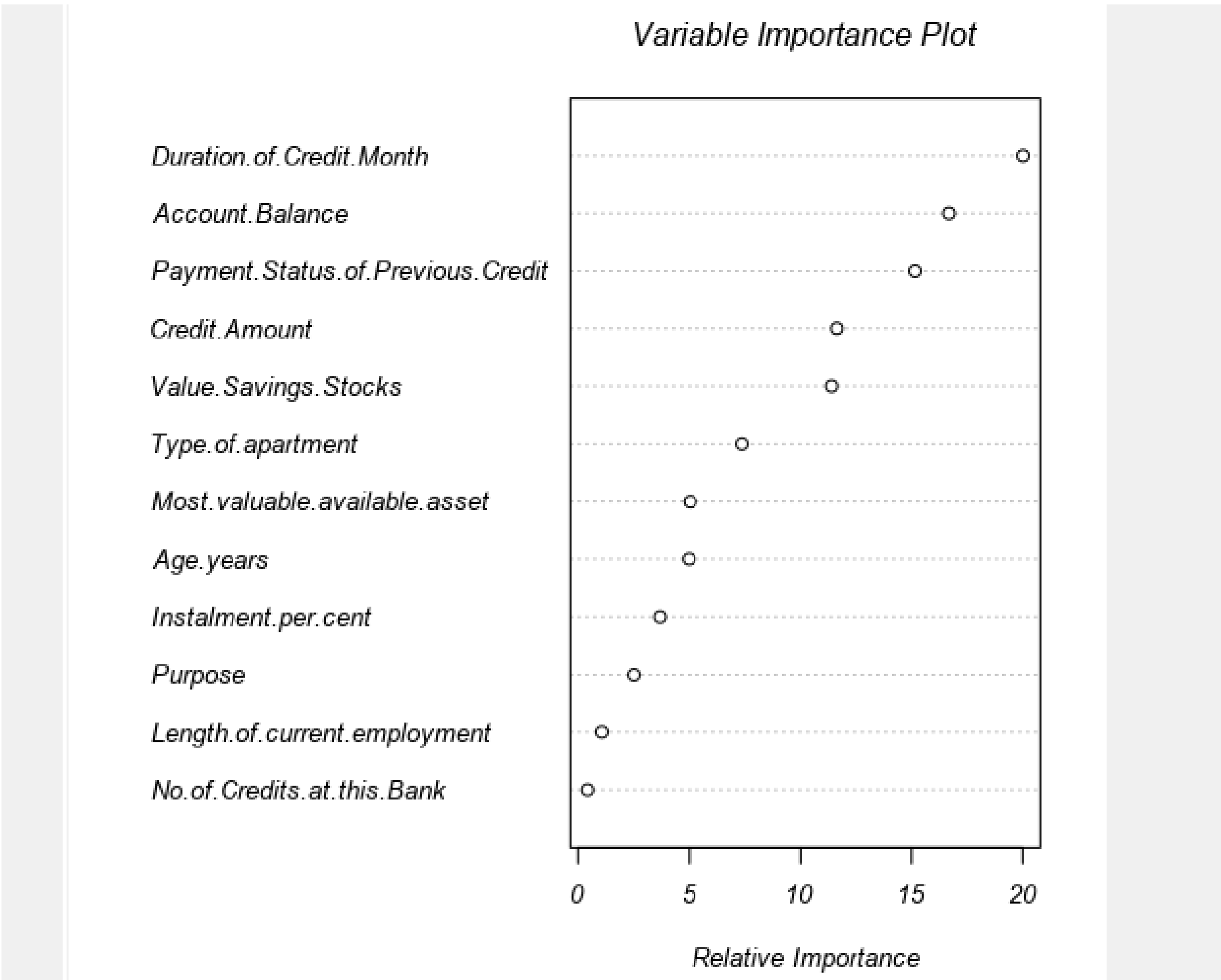


Figure 12: Variable Importance Plot for Boosted Model

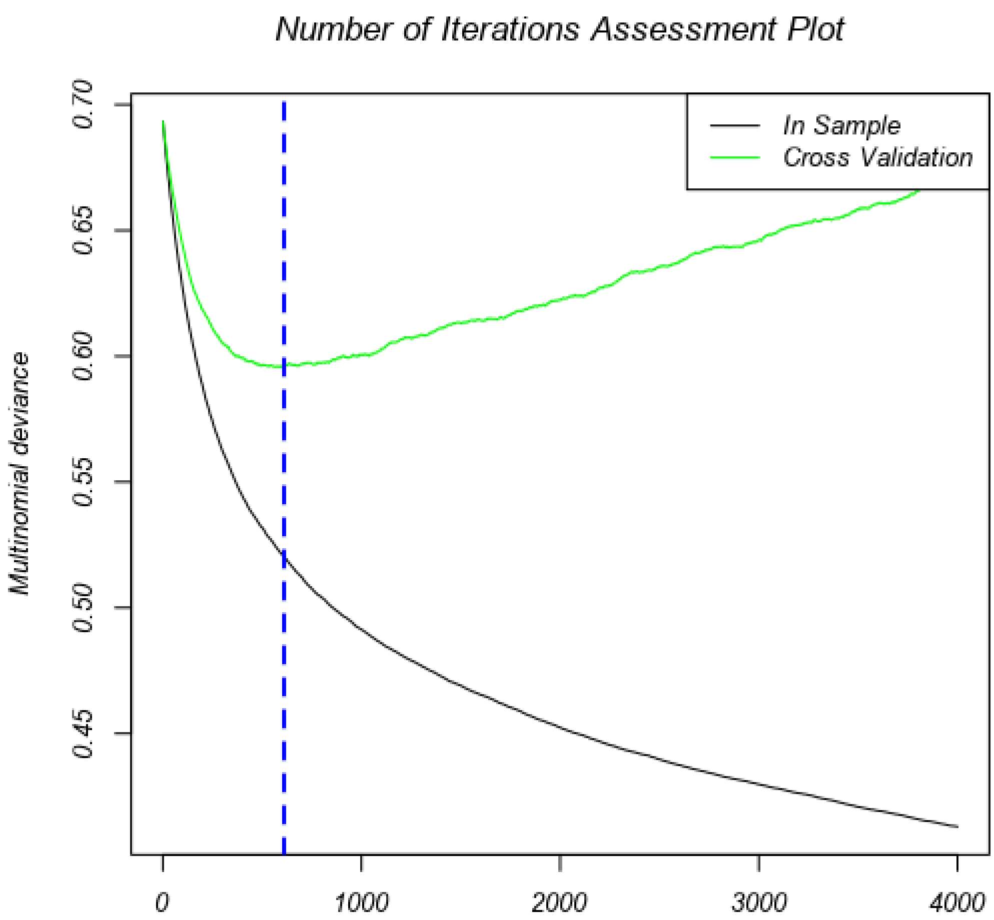


Figure 13: Number of Iteration Assessment Plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Test_Model	0.7667	0.8548	0.8080	0.9810	0.2667

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.

Confusion matrix of Test_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	33
Predicted_Non-Creditworthy	2	12

Figure 14: Model comparison Report of Boosted Model.

Account Balance and Credit Amount are the most significant variables from figure 14. Overall accuracy is 78.7%. This time the accuracy of the creditworthiness and non-creditworthiness are 78.3% and 82.0% respectively and they both share close percentages, indicating a lack of biais in predicting credit eligibility.

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

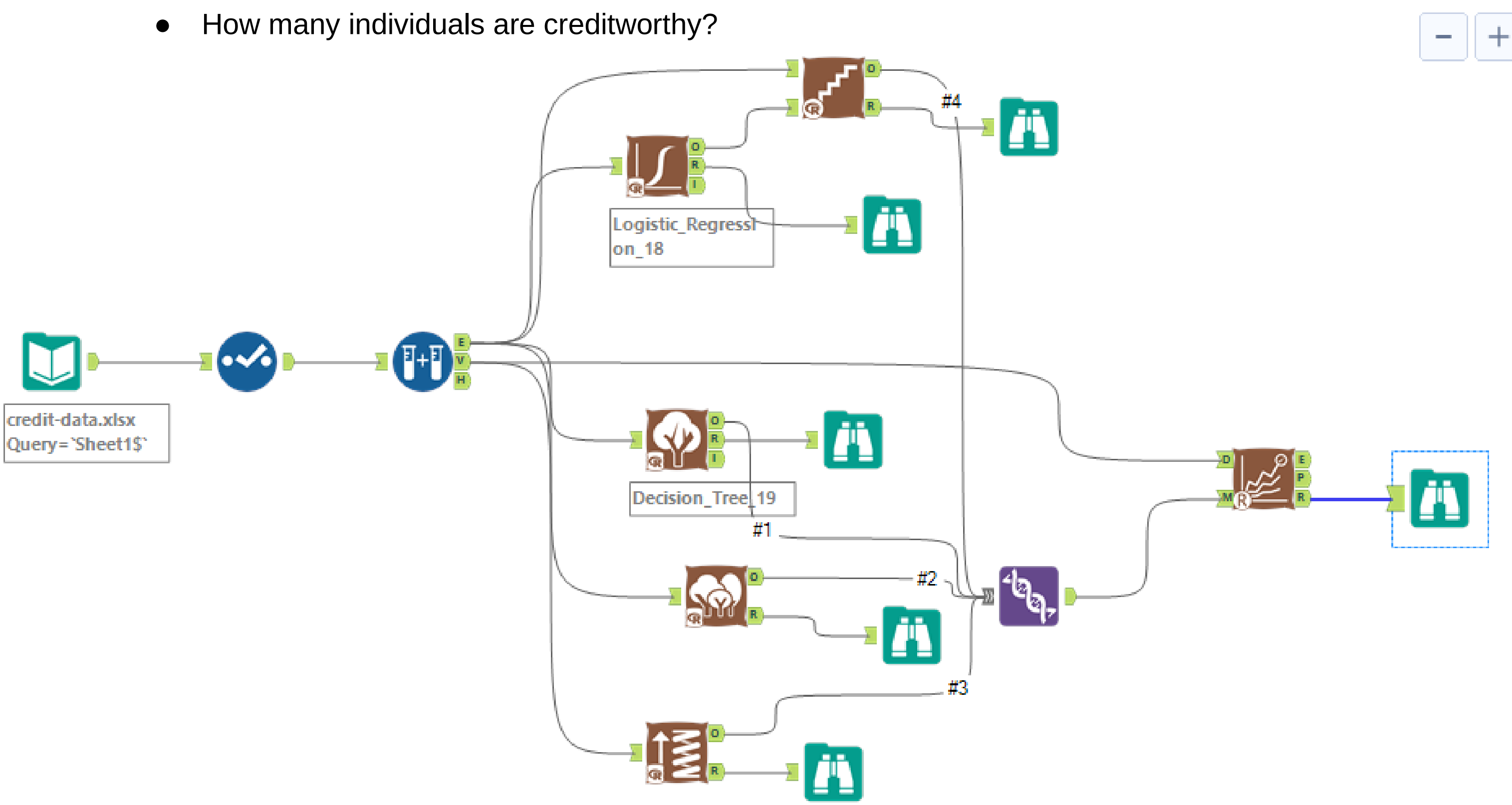
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- 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

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_19	0.7467	0.8304	0.7035	0.8857	0.4222
X	0.7933	0.8681	0.7368	0.9714	0.3778
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Test_Model	0.7667	0.8548	0.8080	0.9810	0.2667

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.

Confusion matrix of Decision_Tree_19

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

Confusion matrix of Stepwise

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

Confusion matrix of Test_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	33
Predicted_Non-Creditworthy	2	12

Confusion matrix of X

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

Figure 14: Side-by-side Model Comparison-confusion matrix.

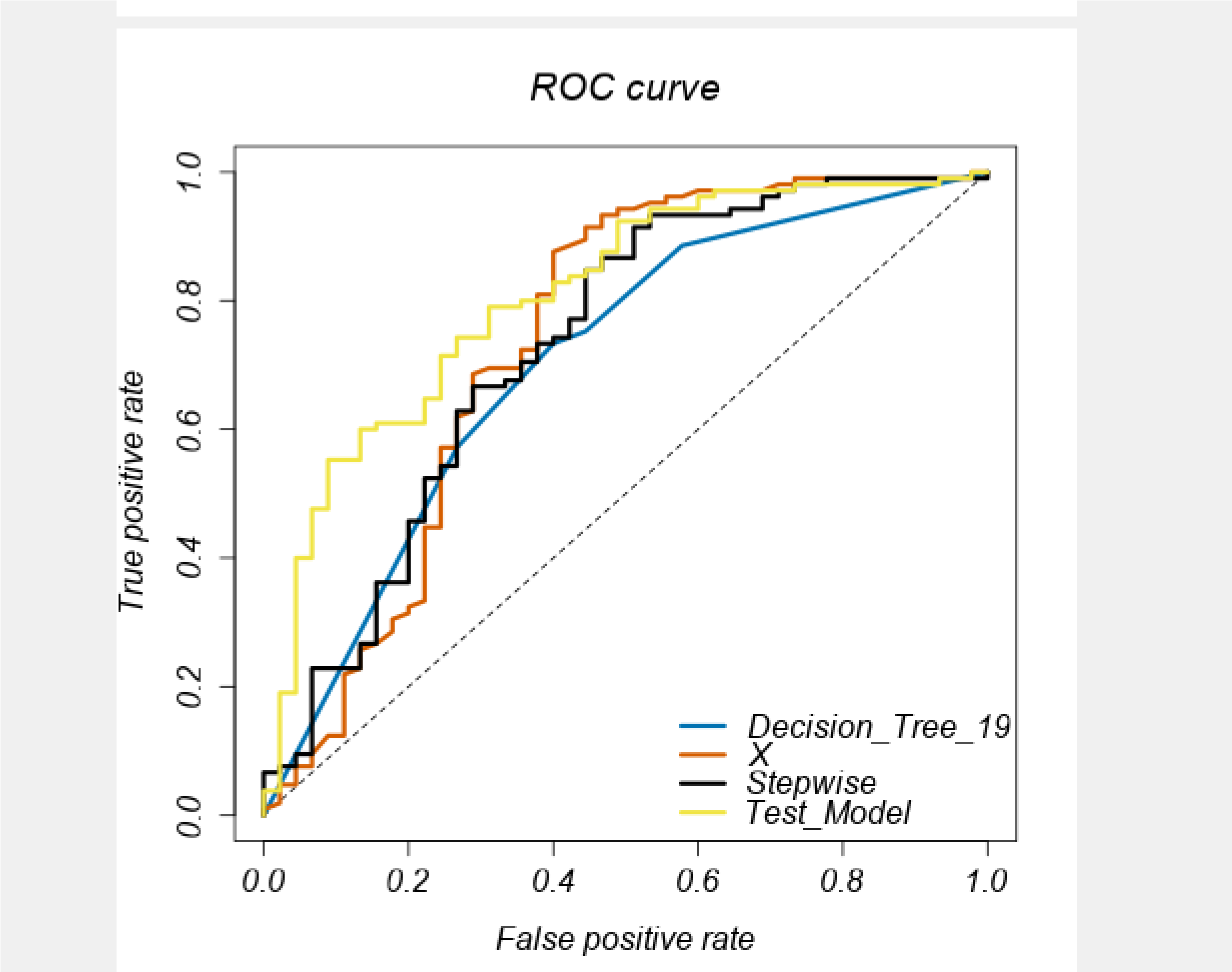


Figure 15: ROC curve for all 4 Classification models.

According to its high accuracy of 80% presented, Forest Model is the best choice. At 96% and 42% for creditworthy and non-creditworthy respectively, the accuracies for these two groups are the highest compared to other models.

The model is not biased towards a group as the accuracy difference between creditworthy and non-creditworthy are very minimal. Based on the ROC curve, the forest model reaches the **positives rate** at the fastest rate or hugs the most positive side of the graph. These values regarding bias and accuracies are important for a lender and the loan customer since they equalize the opportunities for loan acceptance or denial based on each customer's individual ability to responsibly utilize the loan.

416 individuals are creditworthy.