Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- **What decisions needs to be made?**

We have had an influx of 500 loan applications, we must use the past data to prepare models that will help us in deciding whether an individual is credible enough or not to be given a loan. In order to prepare these models we must decide the most important parameters that would be right for our analysis. Then we must prepare 4 models and then decide the best model from amongst them.

- **What data is needed to inform those decisions?**

We need data on parameters that affect credibility for past loan seekers so that we can formulate the classification models. Then we also need the input parameters for the actual loan seekers, so that we can input their data into the classification model in order to predict whether or not they should be given the loan.

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

We need to make use of Binary Classification models in order to make the required decisions.

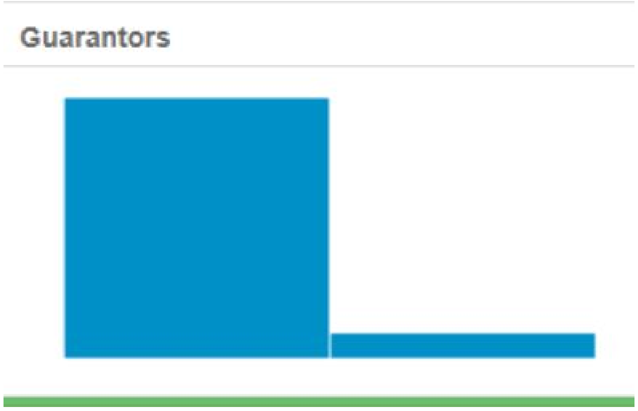
Step 2: Building the Training Set

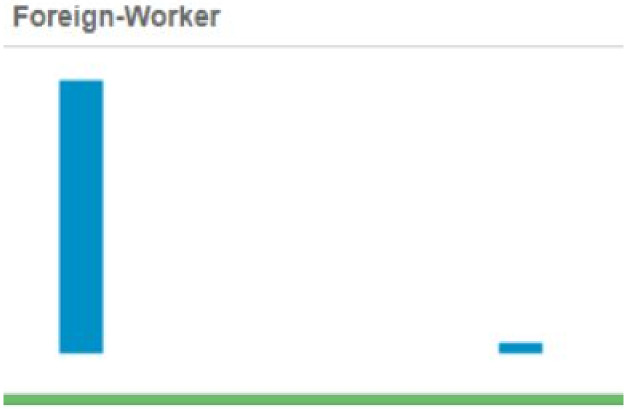
- In your clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

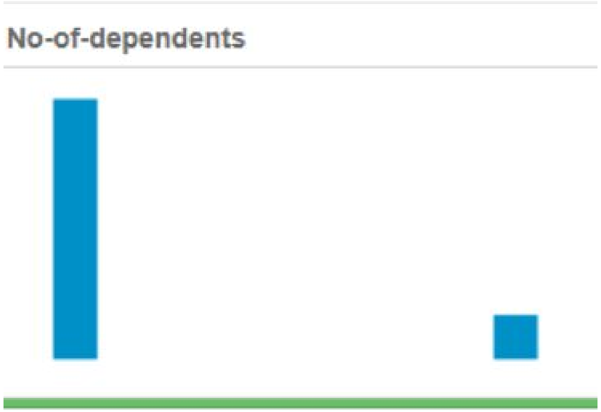
For my analysis I have removed the following parameters explanations for each are also specified below –

1. Guarantors, Foreign-Worker, No-of-dependents

These parameter shows the tendency of low variability, as data is skewed.

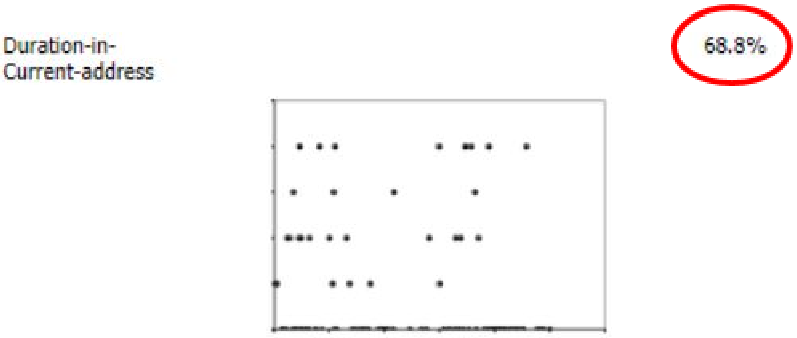






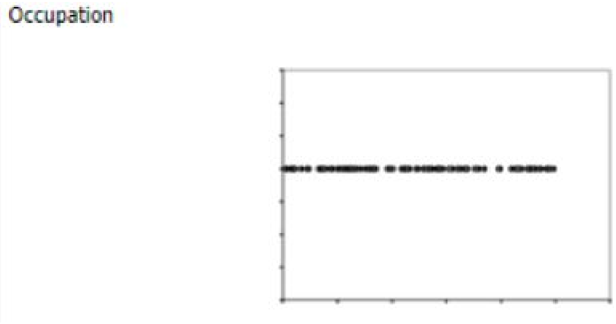
2. Duration in current address

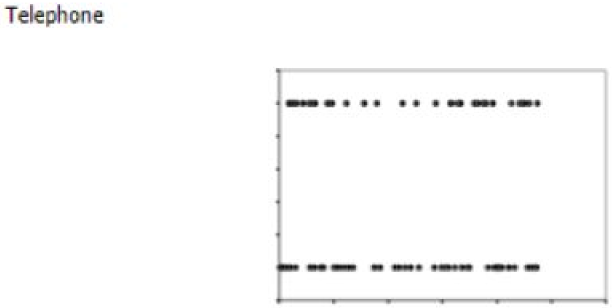
This parameter has been removed because it has a high percentage of missing values.



3. Occupation, Concurrent Credit and Telephone

These parameters have been removed because they have only a few unique value.

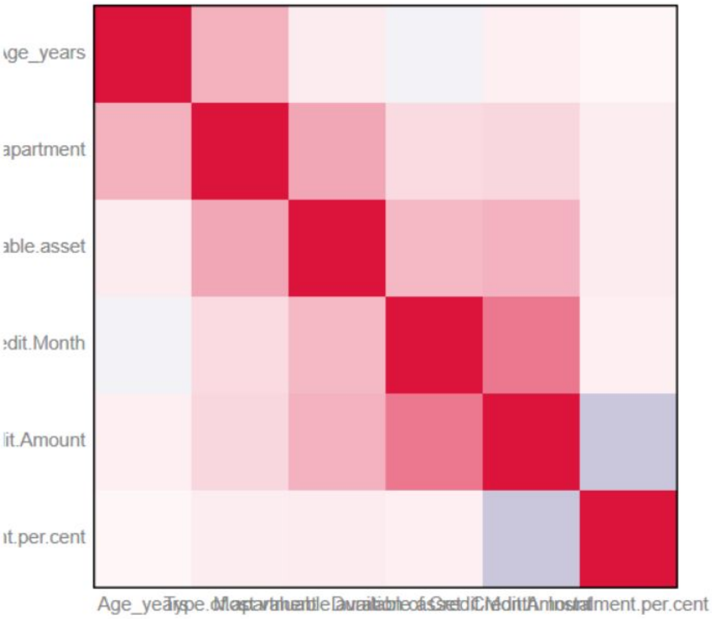

Occupation


Telephone

The parameters of Age has been imputed, wherever there were null variables the value has been replaced by the median age. The average now turns out to be 36.

Concurrent-Credits	0.0%	1

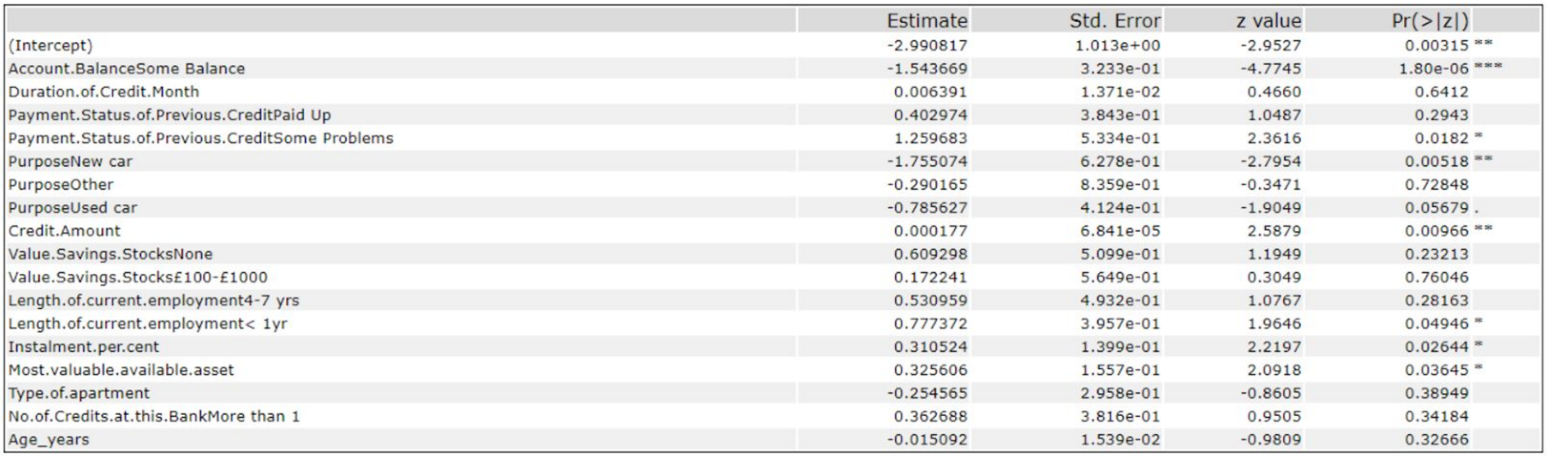
Lastly we check the correlation between the chosen parameter values, the matrix for the same is as follows –



No two parameters have high correlation amongst them.

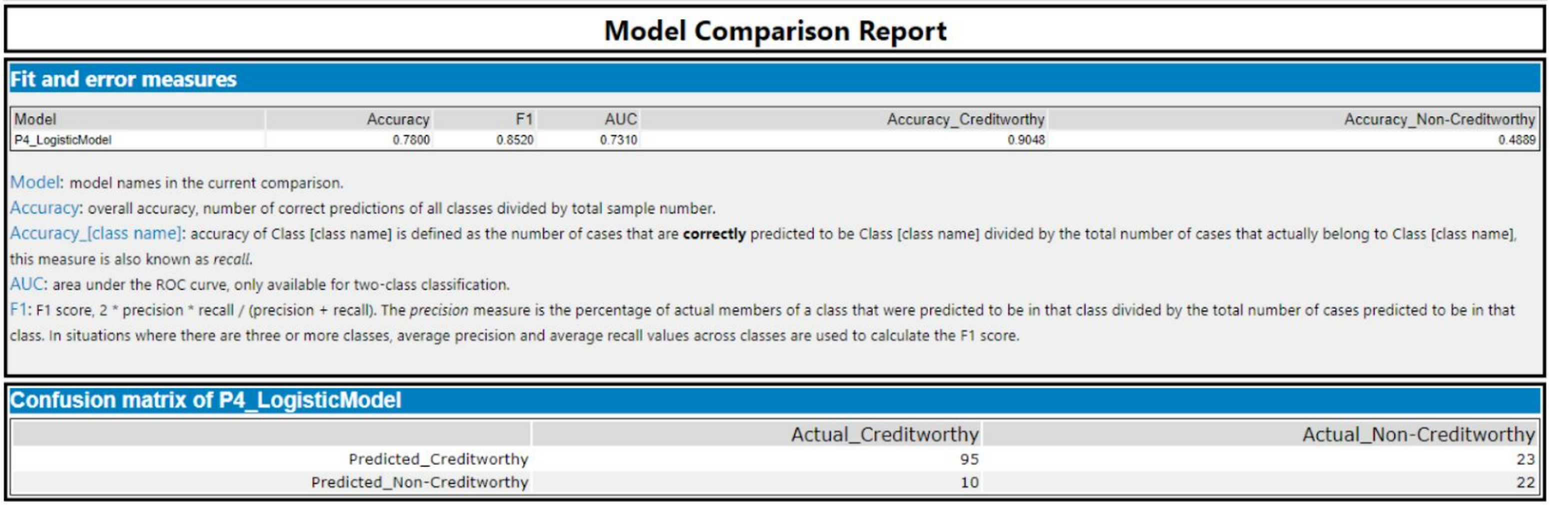
Step 3: Train your Classification Models

1.LOGISTIC MODEL



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315 **
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06 ***
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182 *
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518 **
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679 .
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966 **
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946 *
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644 *
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645 *
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184
Age_years	-0.015092	1.539e-02	-0.9809	0.32666

The predictor variables that are most significant are Balance, Previous credit, Purpose (new car), Credit Amount, Payment status, Length of current employment, Instalment per cent, and the most valuable asset available.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
P4_LogisticModel	0.7800	0.8520	0.7310	0.9048	0.4889

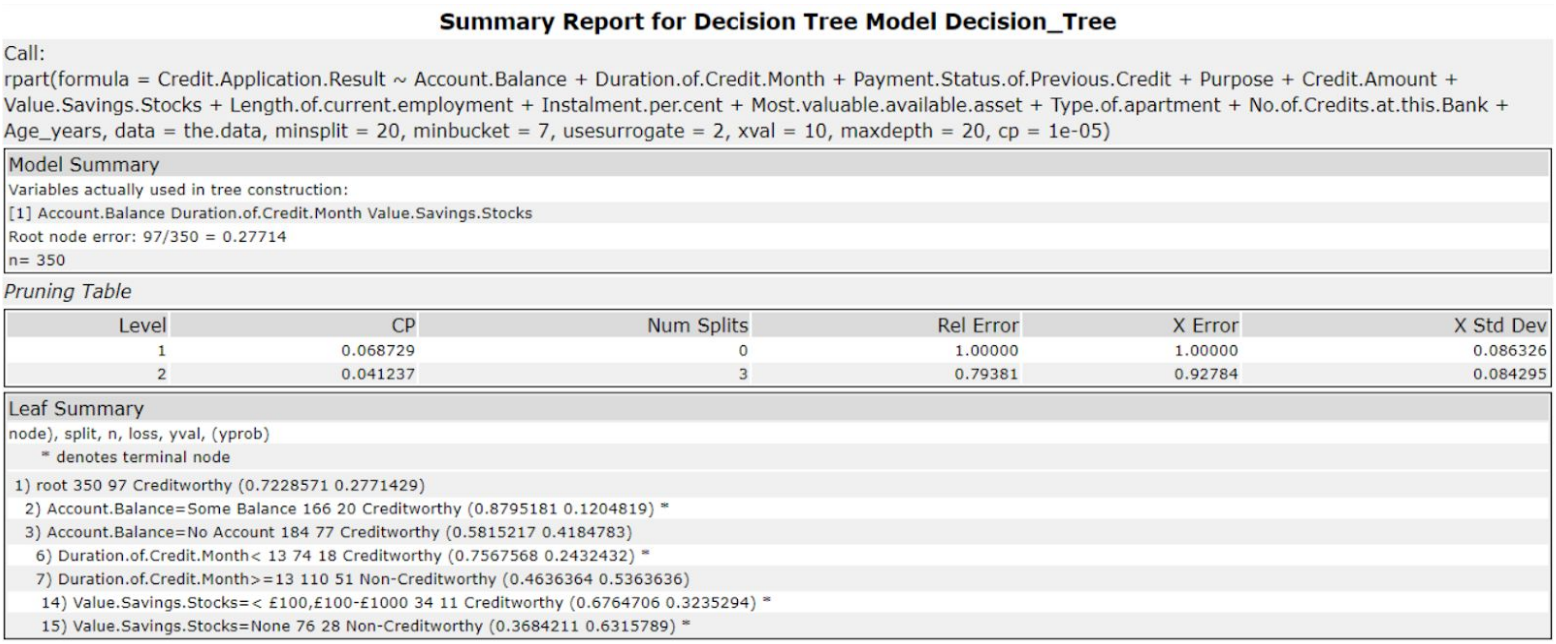
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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 P4_LogisticModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

The model has an accuracy rate of 78%, the model is biased towards predicting customers as non-credit-worthy as the accuracy rate pertaining to that parameter is 48%.

2. DECISION TREE MODEL



Summary Report for Decision Tree Model Decision_Tree

```
Call:
rpart(formula = 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 + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank +
Age_years, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)
```

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
```

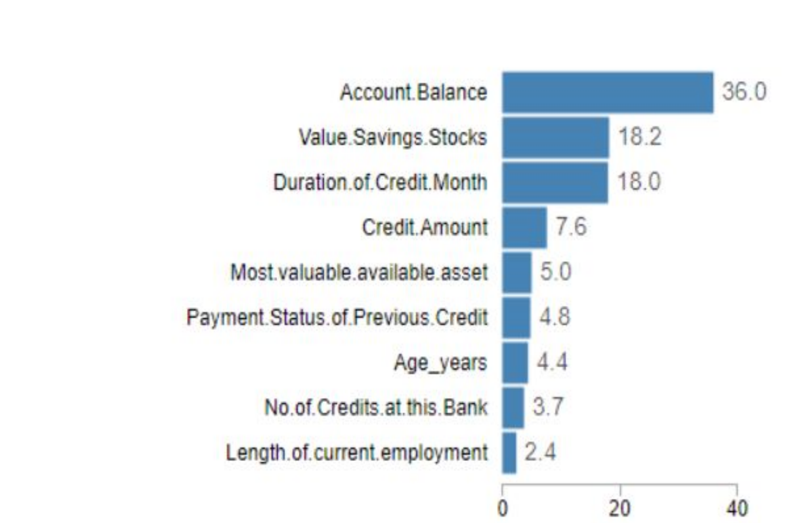
Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

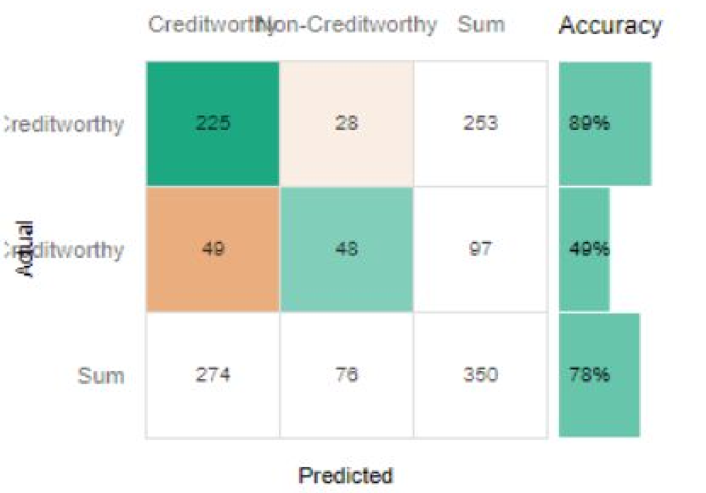
Leaf Summary

```
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) *
```

Variable Importance



Confusion Matrix



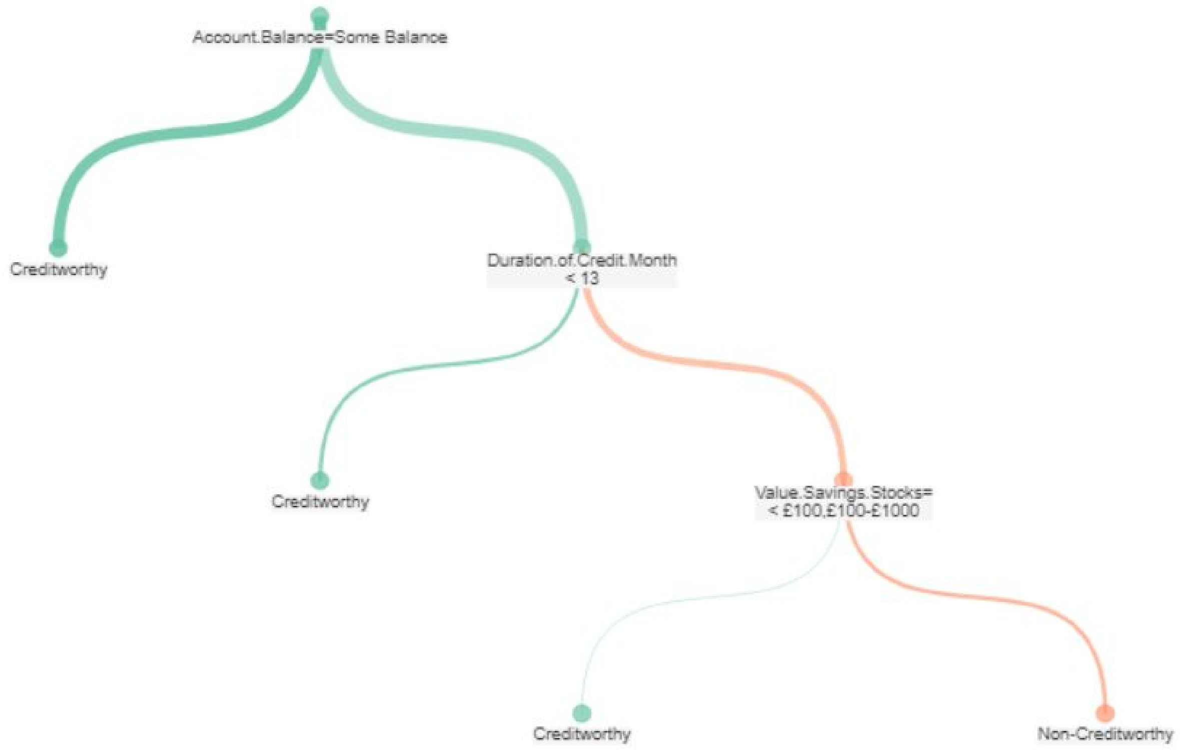
Actual	Creditworth	Non-Creditworthy	Sum	Accuracy
reditworthy	225	28	253	89%
rditworthy	49	48	97	49%
Sum	274	76	350	78%

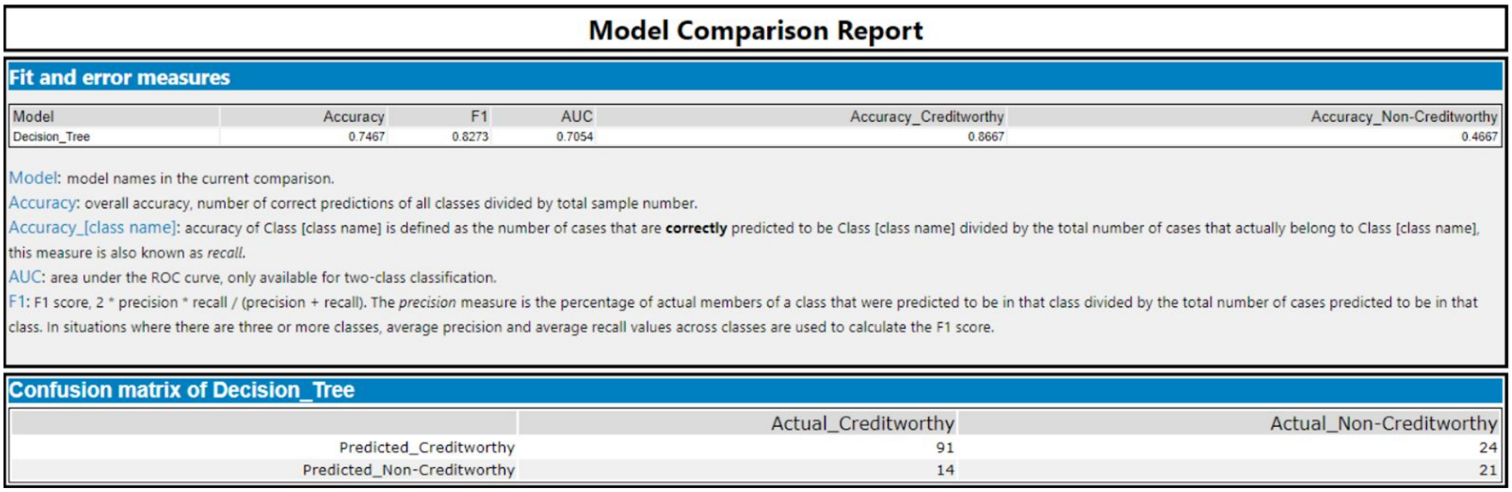
Predicted

The model has an accuracy rate of 74%, the model is biased towards predicting customers as non-credit-worthy as the the accuracy rate pertaining to this parameter stands at 46%.

The most important variables are account balance, value saving stock and duration of credit month.

Decision Tree


Account.Balance=Some Balance
Creditworthy
Duration.of.Credit.Month < 13
Creditworthy
Value.Savings.Stocks= < £100,£100-£1000
Creditworthy
Non-Creditworthy



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
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*.

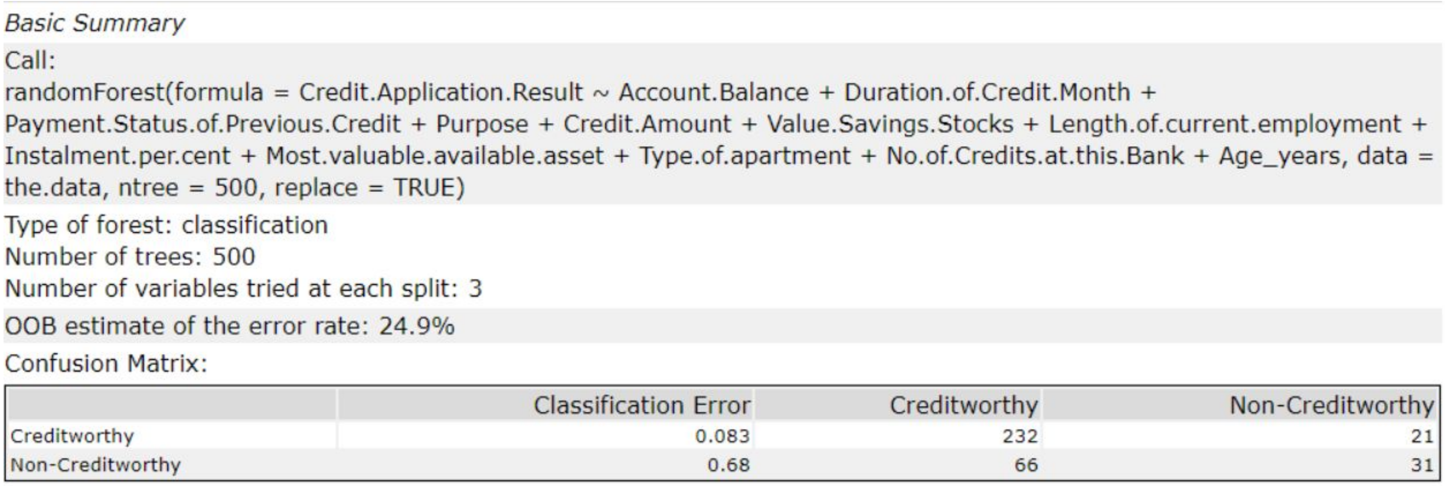
AUC: area under the ROC curve, only available for two-class classification.

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 Decision_Tree

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

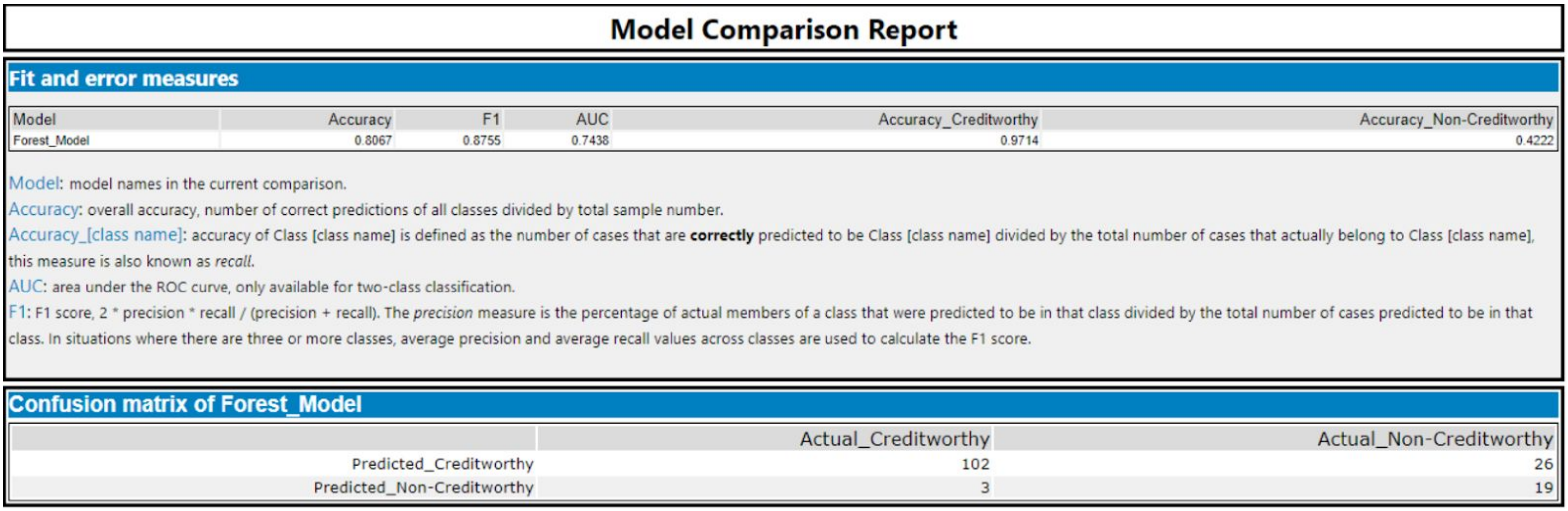
3. RANDOM FOREST MODEL



Basic Summary

```
Call:
randomForest(formula = 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 + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years, data =
the.data, ntree = 500, replace = TRUE)
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 24.9%
Confusion Matrix:
```

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.083	232	21
Non-Creditworthy	0.68	66	31



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8067	0.8755	0.7438	0.9714	0.4222

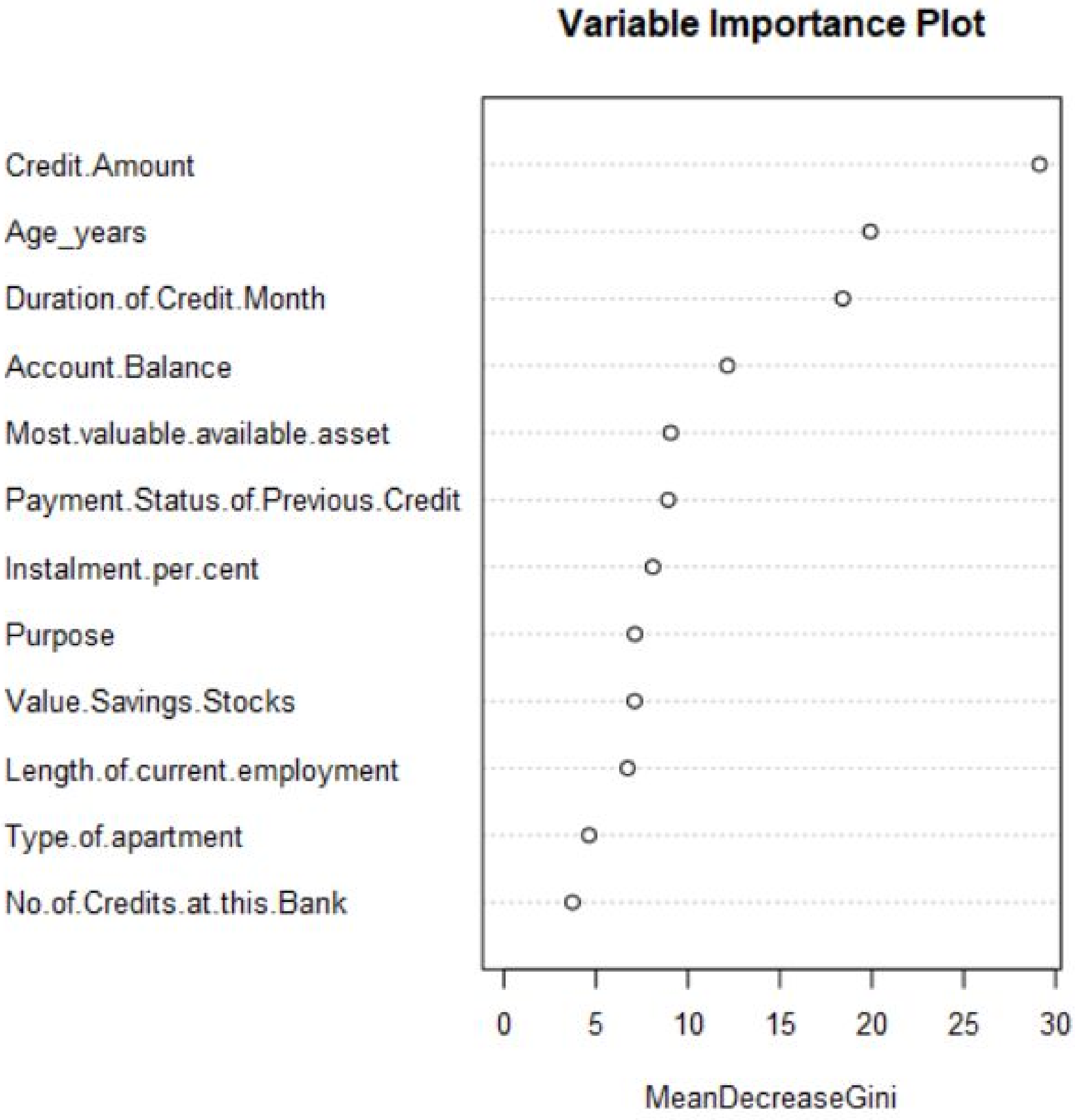
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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

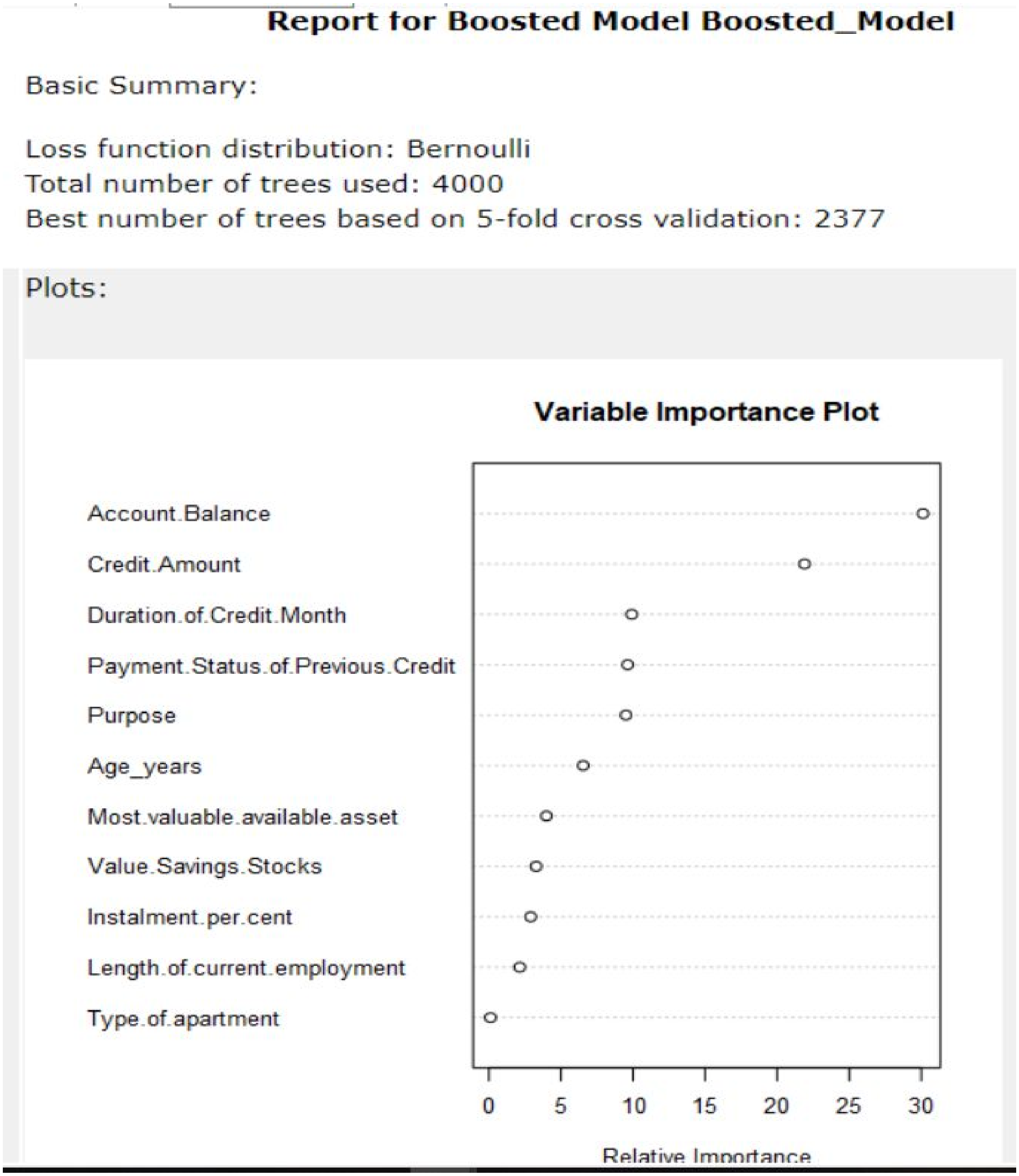
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

The forest model predicts the status of a customer with 80% accuracy. The model is biased towards predicting customers as non-credit-worthy as the accuracy rate for the same is 42%, whereas the accuracy rate with which the model predicts credit-worthy is high at 97%.

The most important variables are credit amount, age years and duration of credit month.


Variable Importance Plot
Credit.Amount
Age_years
Duration.of.Credit.Month
Account.Balance
Most.valuable.available.asset
Payment.Status.of.Previous.Credit
Instalment.per.cent
Purpose
Value.Savings.Stocks
Length.of.current.employment
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
30
MeanDecreaseGini

4. BOOSTED MODEL

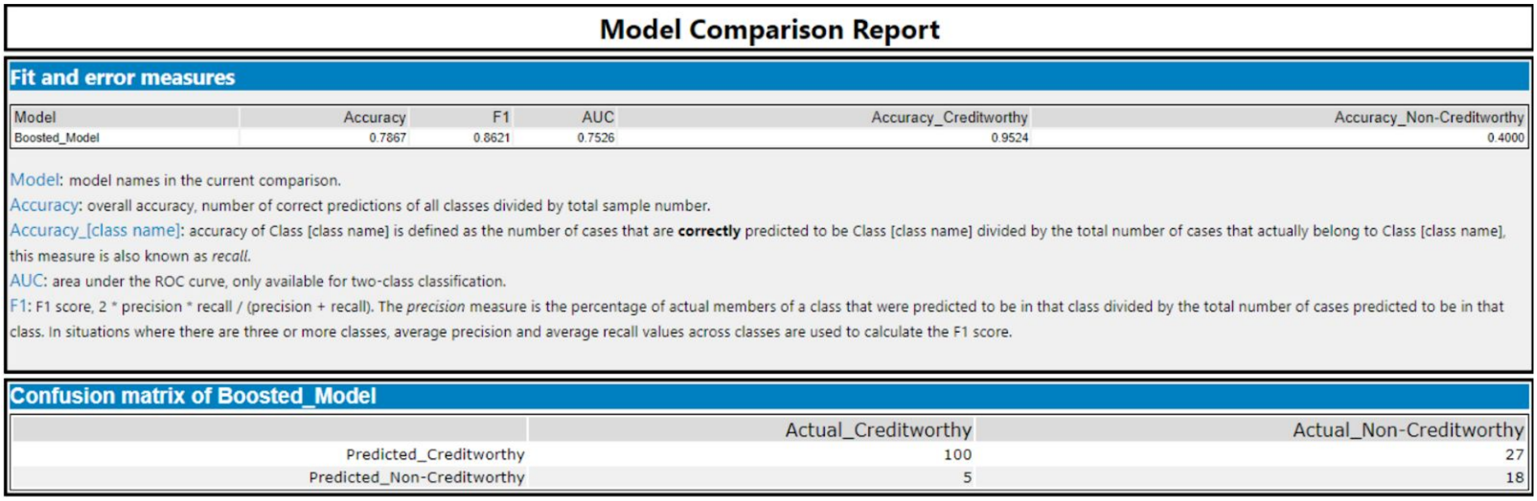


Report for Boosted Model Boosted_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2377

Plots:



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8621	0.7526	0.9524	0.4000

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 Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

The boosted model has an accuracy rate of 78%, the model is biased towards predicting customers as non-credit worthy with the accuracy rate at 40%, whereas the accuracy rate for credit-worthy is 95%.

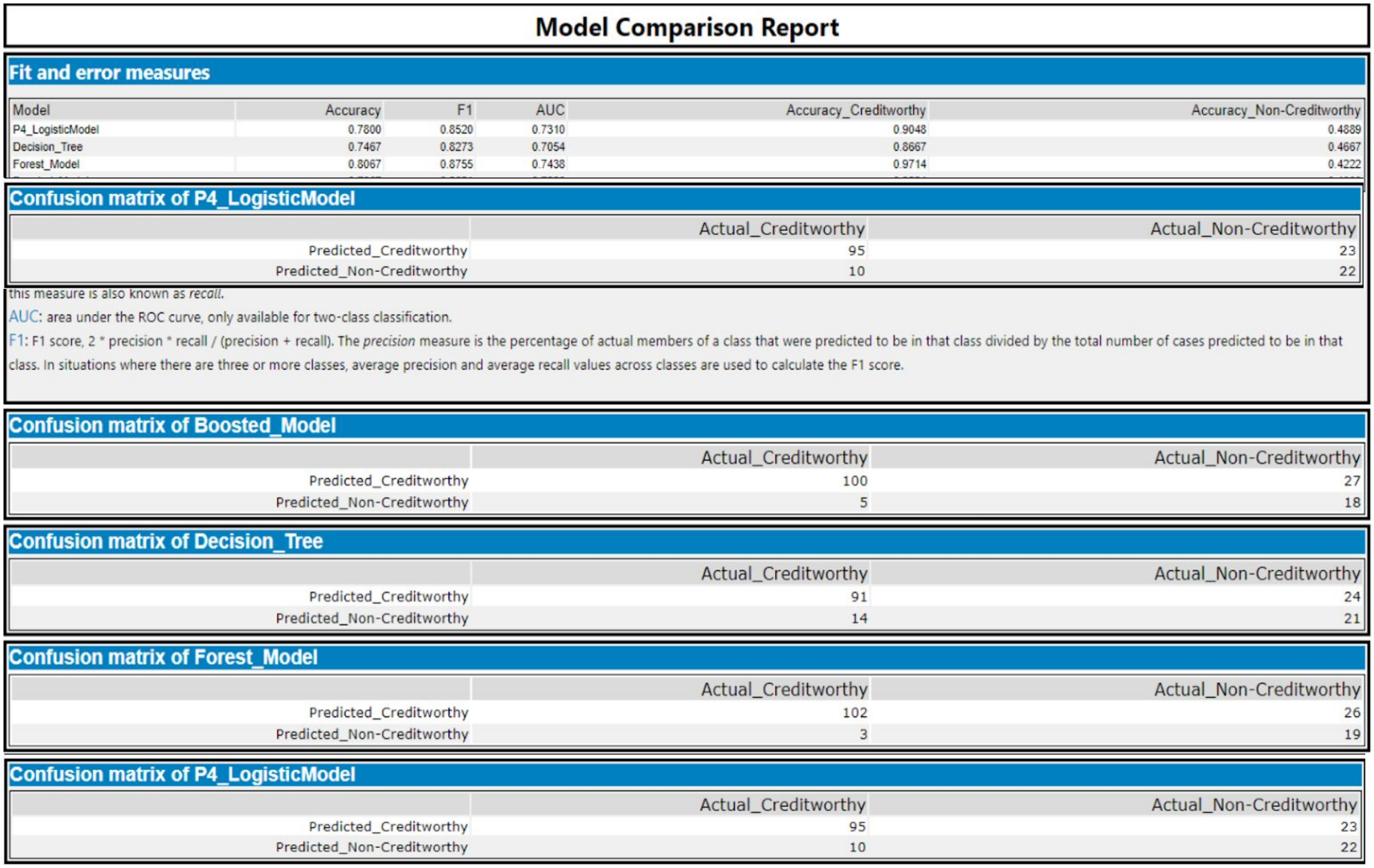
The most important predictor variables are account balance, credit amount, and duration of credit mont.

Step 4: Writeup

The logistic model is the best model from amongst the 4 models as it has the highest, accuracy rate of 80%. It has the highest accuracy rate when it comes to predicting the credit-worthy status of a customer, the value stands at 97%. The confusion matrix for the forest model is also the best as it has the lowest number of predictions that are wrong.

The ROC curve for the forest model is also the best as it can be observed through the graph.

Therefore the forest model is the best, using this model we find that 408 customers are credit worthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
P4_LogisticModel	0.7800	0.8520	0.7310	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.8067	0.8755	0.7438	0.9714	0.4222

Confusion matrix of P4_LogisticModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

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 Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of Decision_Tree

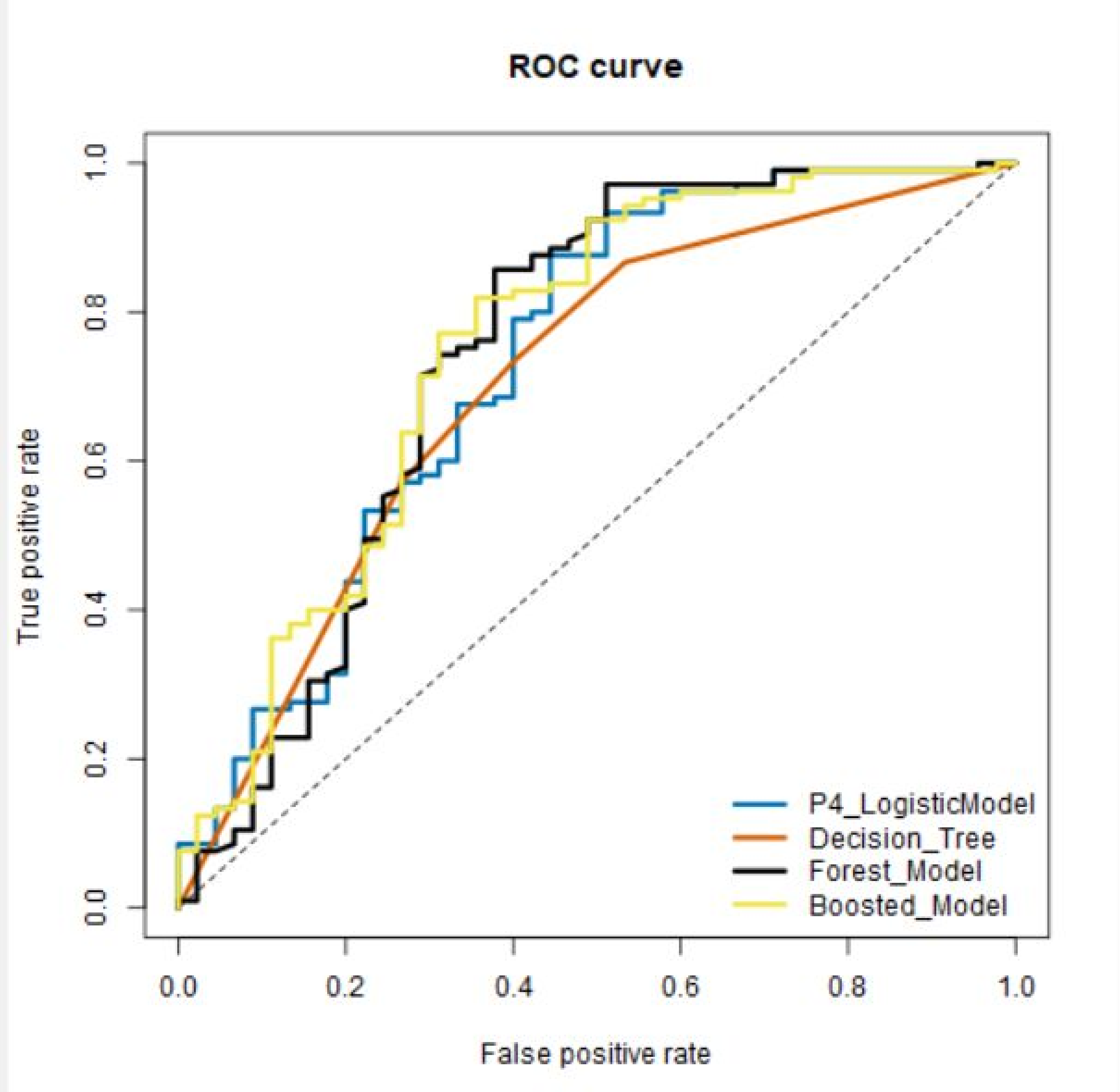
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of P4_LogisticModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
P4_LogisticModel
Decision_Tree
Forest_Model
Boosted_Model