Project -Predict Default Risk

Key Decisions –

Q.1) What decisions need to be made?

Ans – 1) The business decision that needs to be made is that whether a customer is eligible for a loan.

Q.2) What data is needed to inform those decisions?

Ans – 2) We need data on all the past applications or the customers who have previously applied for loan & the outcome of those customers based on specific criteria.

Q.3) What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Ans – 3) Since we need to decide whether a client will default or not, Binary models like Logistic regression, Forest Model, Boosted Model & Decision tree will be required for this analysis.

Step 2: Building the Training Set

The dataset “credit-data-training” consists of 20 columns, which includes both numeric and non-numeric parameters. On running the field summary tool, the columns which had very high missing records and parameters that were not statistically significant to the dependent variable were removed, and some of the missing records of age were replaced with a median on age.

Step 3: Train your classification model

Estimation and Validation samples were created where 70% of the dataset is Estimation, and 30% is for Validation.

A) Logistic Regression –

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

Q.1) Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.

Ans – 1)

Record | Report

1

Report for Logistic Regression Model Stepwise_Credit

2 *Basic Summary*

3 Call:

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

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.352	-0.731	-0.456	0.769	2.458

6 Coefficients:

7

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.5783608	6.414e-01	-4.0202	6e-05	***
Account.BalanceSome Balance	-1.5715598	3.037e-01	-5.1742	2.28e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2117362	2.952e-01	0.7174	0.47316	
Payment.Status.of.Previous.CreditSome Problems	1.3053044	5.089e-01	2.5648	0.01032	*
PurposeNew car	-1.6344313	6.137e-01	-2.6633	0.00774	**
PurposeOther	-0.4435055	8.242e-01	-0.5381	0.59049	
PurposeUsed car	-0.7315961	3.976e-01	-1.8400	0.06577	.
Credit.Amount	0.0002076	5.453e-05	3.8070	0.00014	***
Length.of.current.employment4-7 yrs	0.3678284	4.537e-01	0.8107	0.41752	
Length.of.current.employment< 1yr	0.7564408	3.833e-01	1.9733	0.04846	*
Instalment.per.cent	0.3426933	1.325e-01	2.5873	0.00967	**

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

(Dispersion parameter for binomial taken to be 1)

Fig 1.1 – Logistic Regression Report

As we can see, the variables whose p-value is less than 0.05 are marked as essential variables that means that those variables are statistically significant to the dependent variables, which is

"Credit.Application.Result" in our case. The crucial variables for Logistic Regression model include –

- ❖ AccountBalanceSome Balance
- ❖ PaymentStatusOfPreviousCredit
- ❖ PurposeNewCar
- ❖ CreditAmount

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Credit	0.7800	0.8507	0.7352	0.8952	0.5111

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.

3

Confusion matrix of Stepwise_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

4

Performance Diagnostic Plots

5

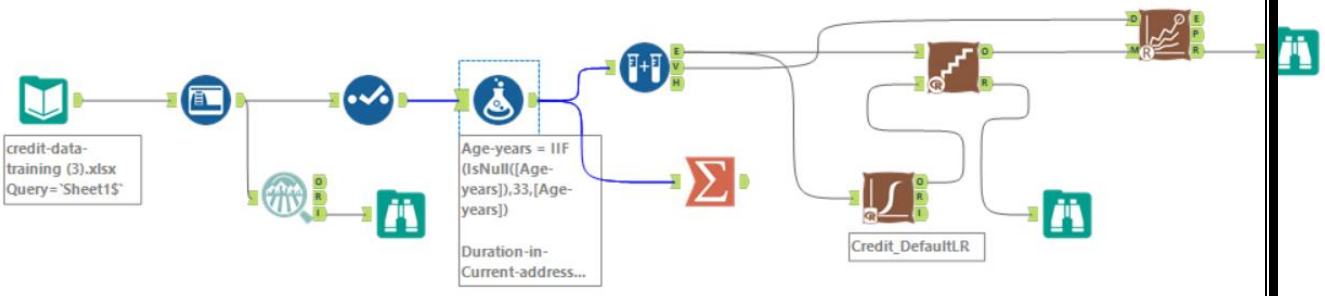


Fig 1.2 – Workflow of Logistic Regression

B. Decision Tree

Q.1) Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.

Ans – 1)

Record Report

1

Summary Report for Decision Tree Model Decision_Client

2

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)
```

3

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

4

Pruning Table

5

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

6

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) *
```

7

Plots

Fig 2.1 – Decision Tree Model Report

As we can see, the variables whose p-value is less than 0.05 are marked as essential variables that means that those variables are statistically significant to the dependent variables, which is "Credit.Application.Result" in our case. The critical variables for the Decision Tree model are shown in Leaf Summary in the above image; all the variables marked as * are statistically significant.

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Client	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_Client

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

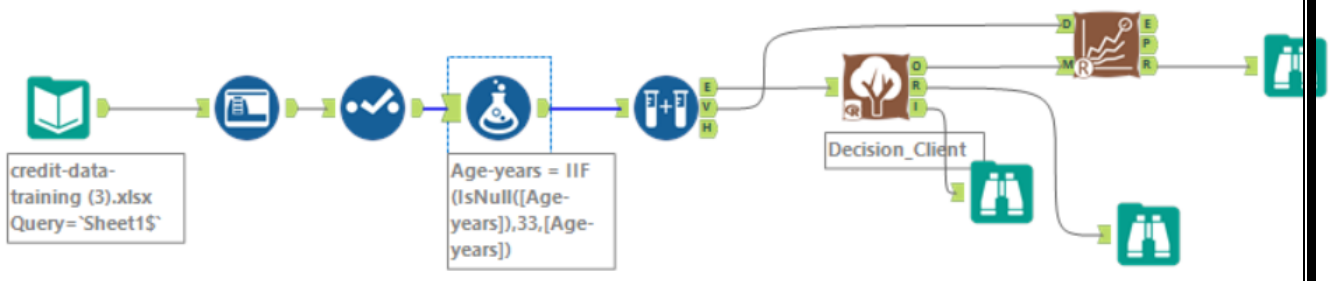


Fig 2.2 – Workflow of Decision Tree Model

C. FOREST MODEL

Q.1) Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.

Ans – 1) As we can see, the variables whose p-value is less than 0.05 are marked as essential variables that means that those variables are statistically significant to the dependent variables, which is “Credit.Application.Result” in our case. The critical variables for the Forest model are shown in the below image.

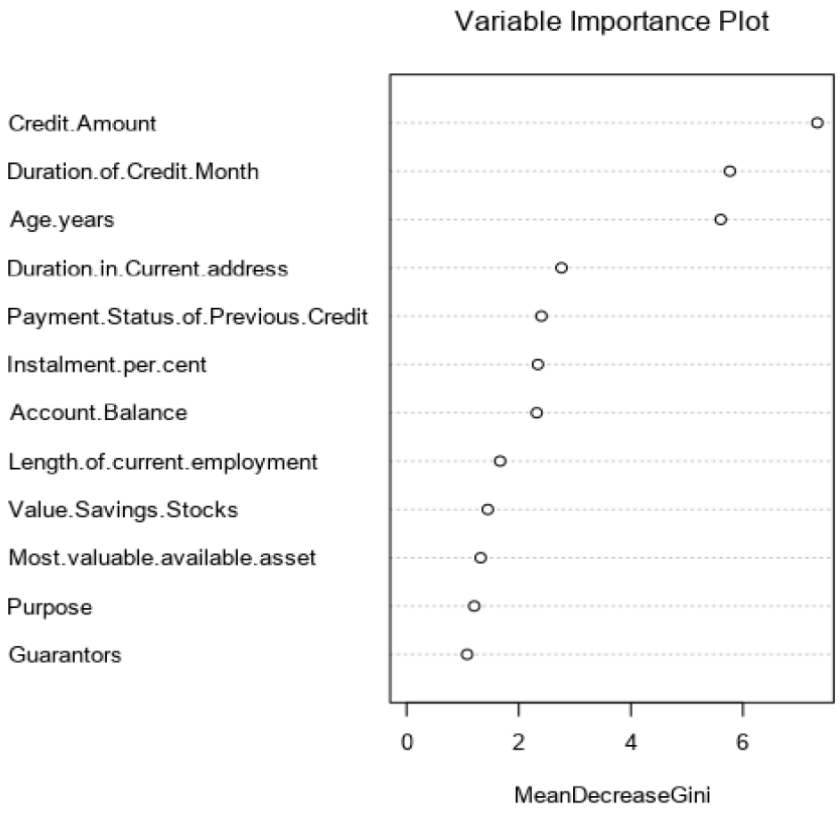


Fig 3.1 – Important variables in Forest Model

The first three variables are essential variables for Forest model i.e.

- ❖ Credit Amount
- ❖ Duration of Credit Month
- ❖ Age.years

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Client	0.7333	0.8361	0.5998	0.9714	0.1778

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.

3

Confusion matrix of FM_Client

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	37
Predicted_Non-Creditworthy	3	8

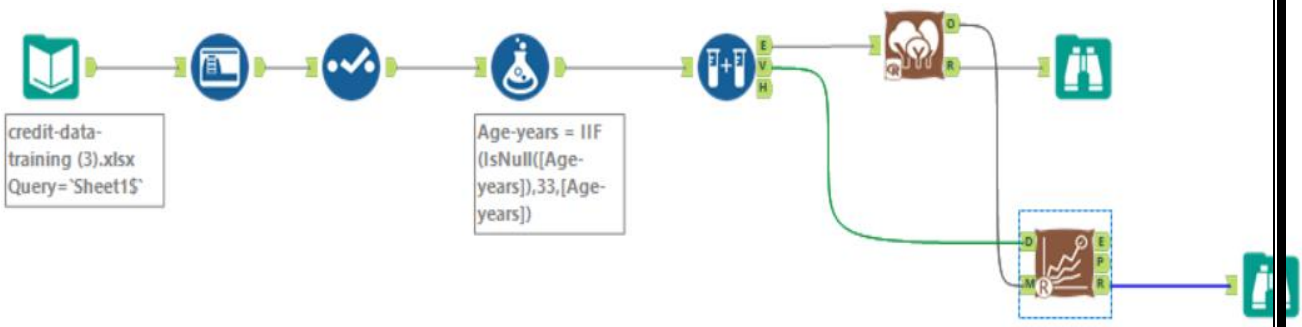


Fig 3.2 - WORKFLOW of Forest Model

D. BOOSTED MODEL

Q.1) Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.

Ans – 1) As we can see, the variables whose p-value is less than 0.05 are marked as essential variables that means that those variables are statistically significant to the dependent variables, which is “Credit.Application.Result” in our case. The crucial variables for the Boosted model are shown in the below image.

1) Credit Balance
2) Account Balance

Record 1 Report

Report for Boosted Model OppalModel

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2018

2 **Plots:**

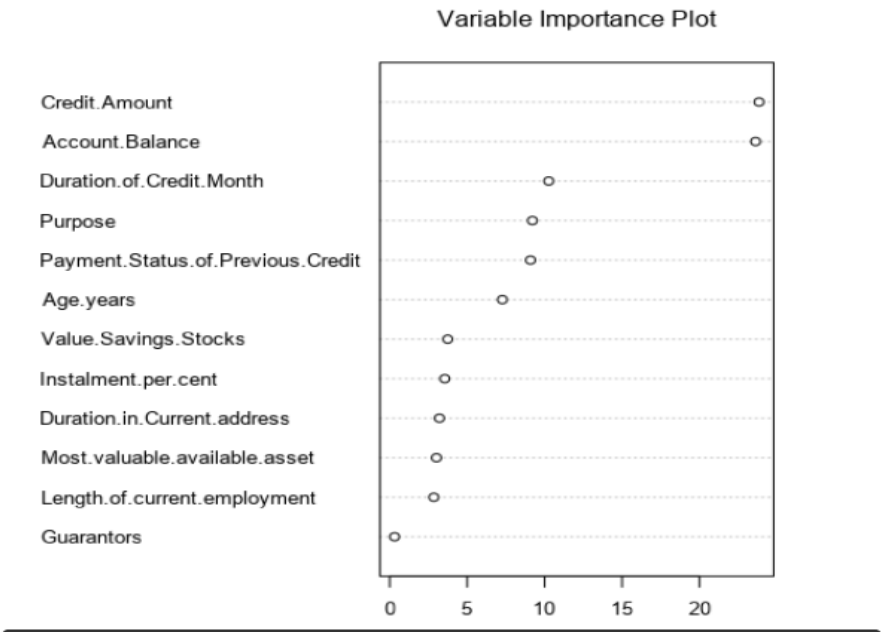


Fig – 4.1 Important Variables in Boosted Model

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Oppai_Model	0.7267	0.8285	0.7047	0.9429	0.2222

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.

3

Confusion matrix of Oppai_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	35
Predicted_Non-Creditworthy	6	10

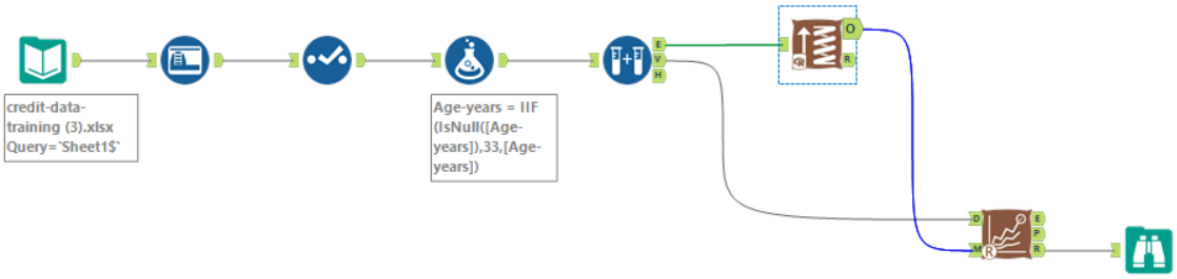


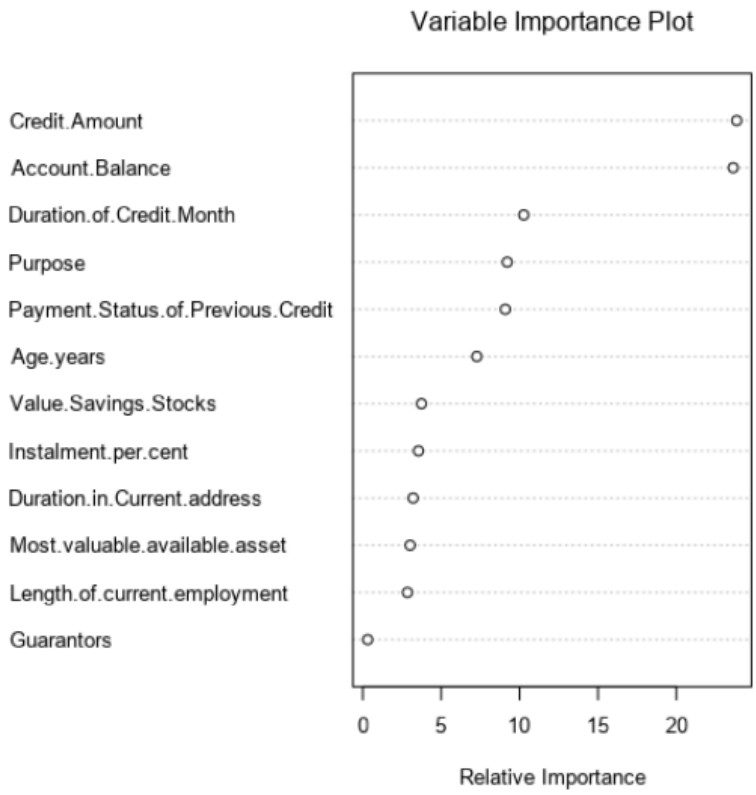
Fig – 4.2 Workflow of Boosted Model

Report for Boosted Model OppalModel

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2018

Plots:



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

CONCLUSION –

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Step_wise	0.7800	0.8507	0.7352	0.8952	0.5111
Forest_Model	0.7333	0.8361	0.5998	0.9714	0.1778
Decision_Tree_8	0.7467	0.8273	0.7054	0.8667	0.4667
Boosted_Model	0.7267	0.8285	0.7047	0.9429	0.2222

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*.

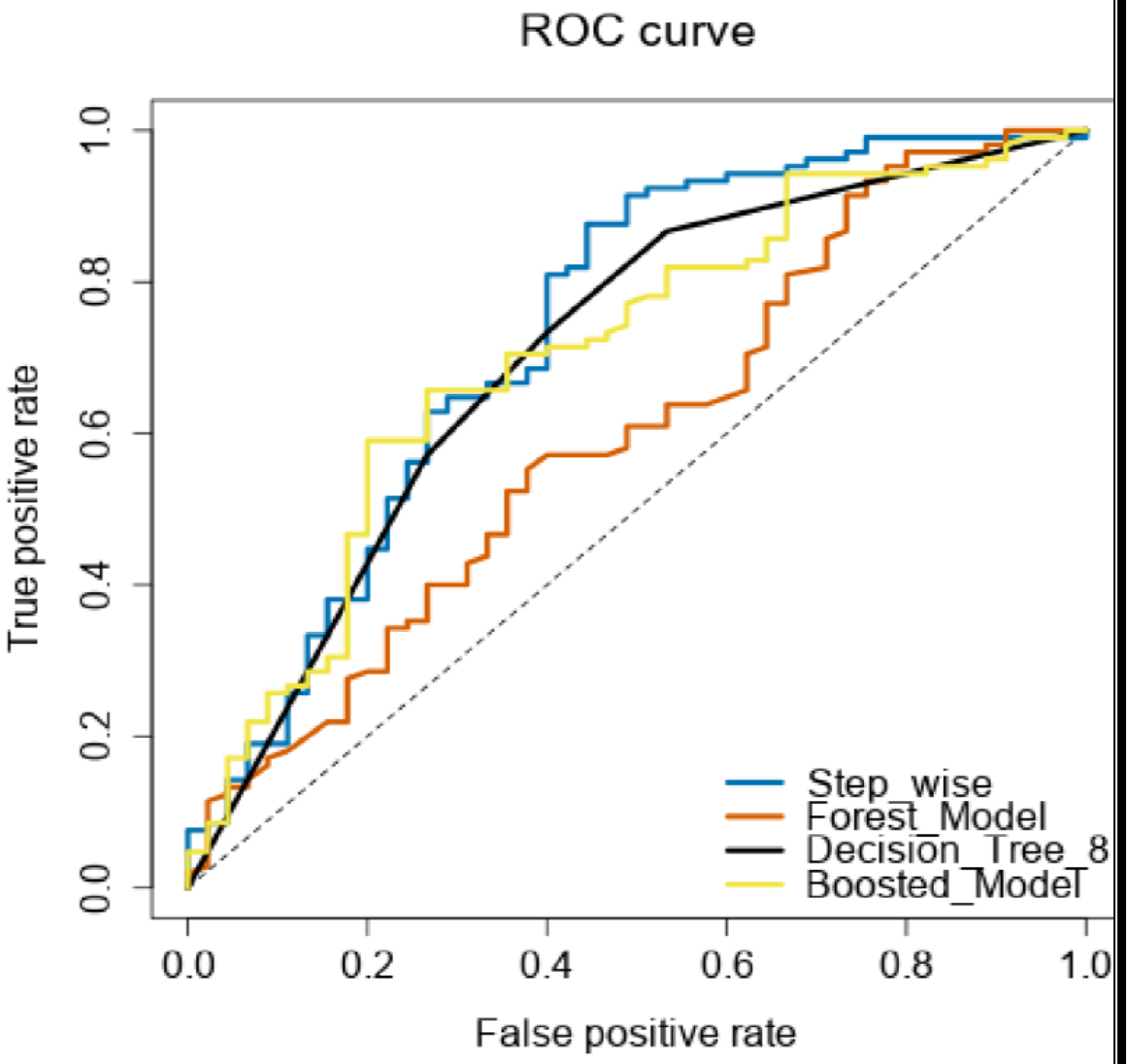
AUC: area under the ROC curve, only available for two-class classification.

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 model chosen for the final analysis is “Stepwise Logistic Regression” because –

I. It has the best overall accuracy when compared with other models.
II. For predicting non-creditworthy customers, the best model was “Stepwise Logistic Regression,” and for predicting creditworthy customers, the best model was the “Forest Model.”

III. The ROC curve shows that the Decision Tree has the best overall true positive rates.



Hence, the Logistic Regression model is used since it has the best fit overall; also, it's overall accuracy is more compared to other models.