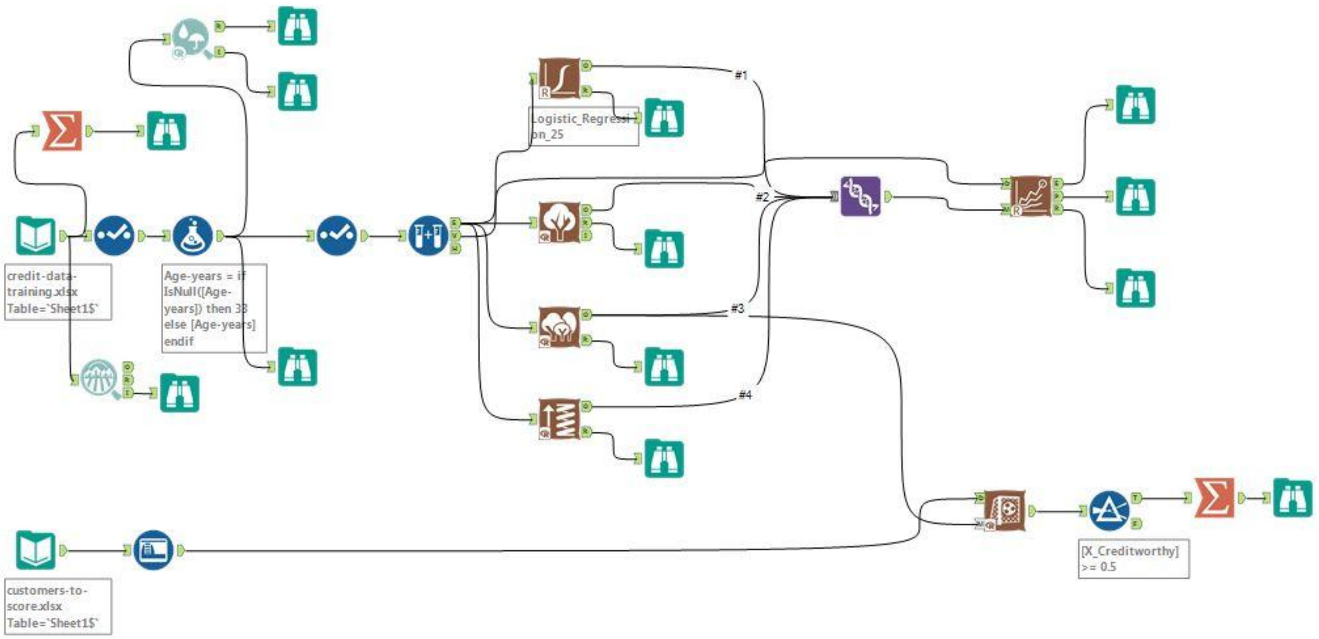
Alteryx Model used for the project:-


Logistic_Regression_25
#1
#2
#3
#4
credit-data-training.xlsx Table=`Sheet1$`
Age-years = if IsNull([Age-years]) then 33 else [Age-years] endif
customers-to-score.xlsx Table=`Sheet1$`
[X_Creditworthy] >= 0.5

Q1. What decisions need to be made?

The decision that needs to be made is to predict whether the applicant is creditworthy or not and whether can be given a loan based on his record.

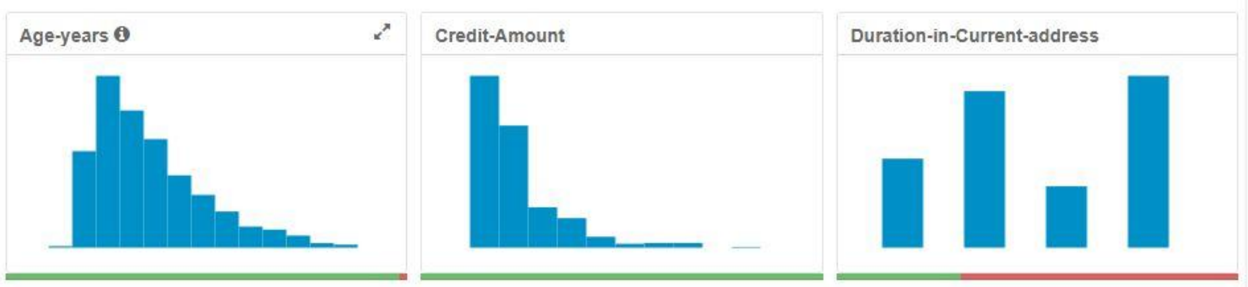
Q2. What data is needed to inform those decisions?

The data from the past applications that were either processed or were rejected is needed to make the decision. The important variables in the data seem like ‘Payment Status of previous credit’ as it can serve as an indicator if the applicant has promptly paid the previous loan amount of not. ‘Age-years’ seems like another important variable as generally with increasing age the financial stability seems to improve.

Q3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

As the solution is in terms of ‘Yes’ (the applicant should be given the loan) or ‘No’ (the applicant shouldn’t be given the loan), a binary classification model would be best suited for the problem.

Q4. In your cleanup process, which field(s) did you impute or remove?

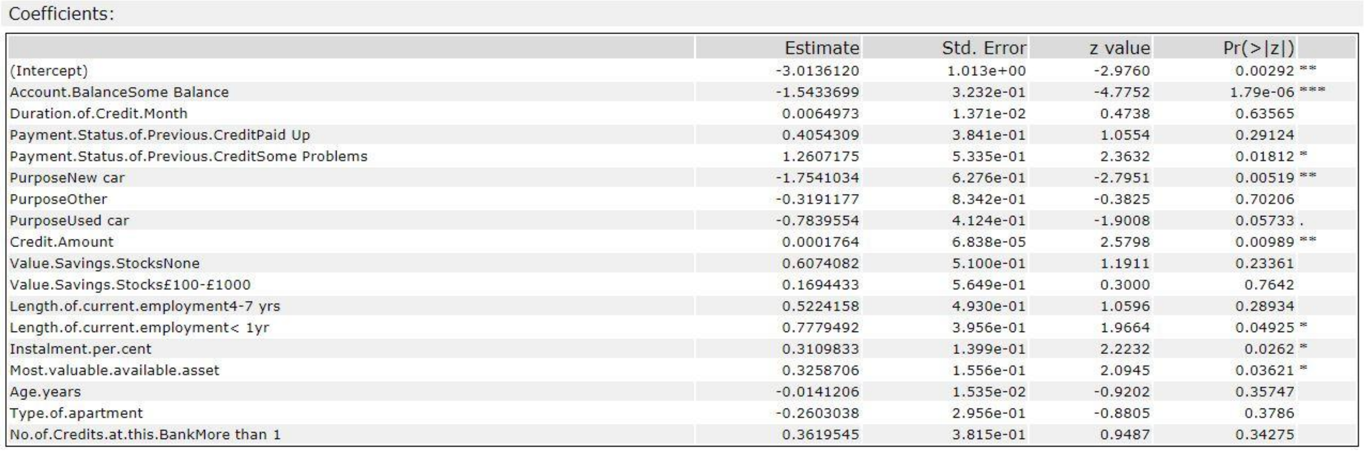


‘Duration-in-Current-address’ field was removed as it had more than half of the values missing. So, imputing or removing the missing values would cause bias data. The missing values in the field ‘Age-years’ were imputed with the median (33) to remove the chances of biased dataset.

Q5. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

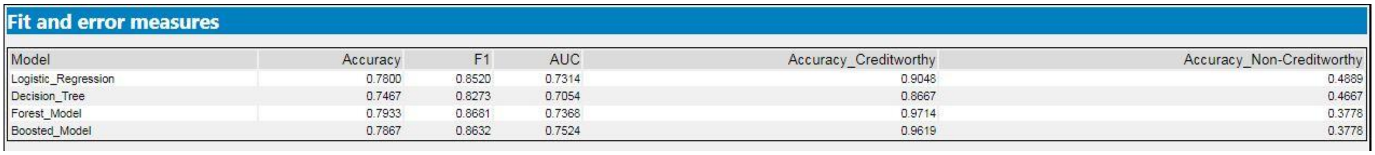
Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

Logistic Regression: - The variable ‘Account Balance’ has the high p value.


Coefficients:

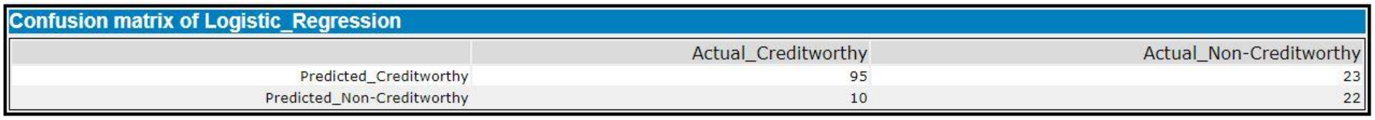
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275

The overall accuracy for the logistic regression model is 0.78.


Fit and error measures

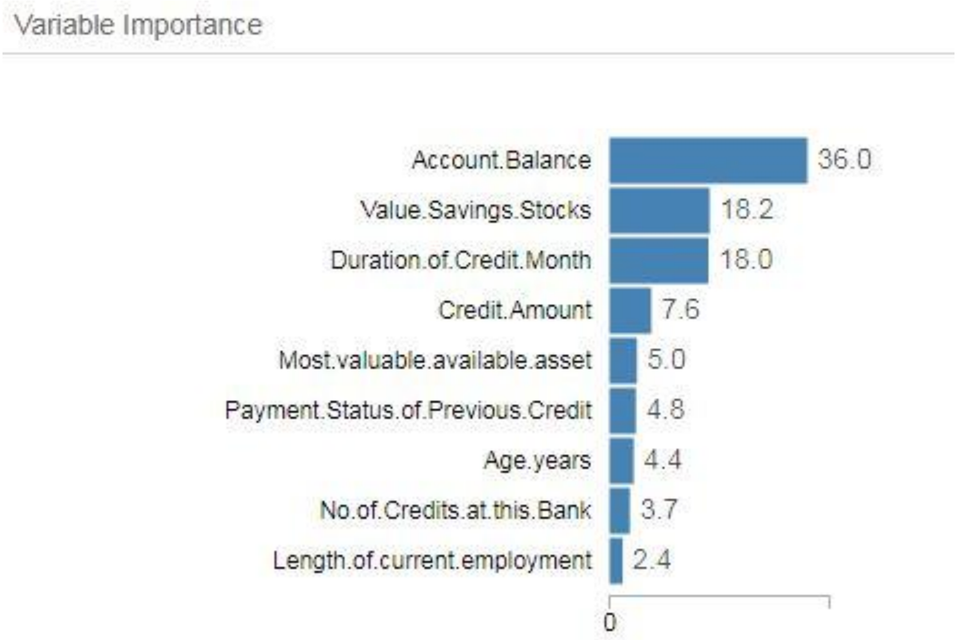
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion Matrix for Logistic regression model is

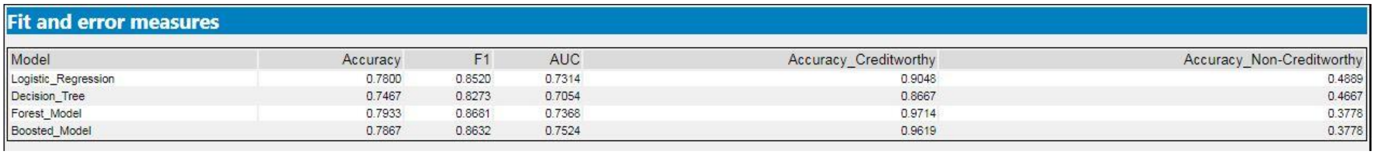

Confusion matrix of Logistic_Regression

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

Decision Tree: - ‘Account balance’, ‘Value Saving Stocks and ‘Duration of Credit Month’ seem important variables



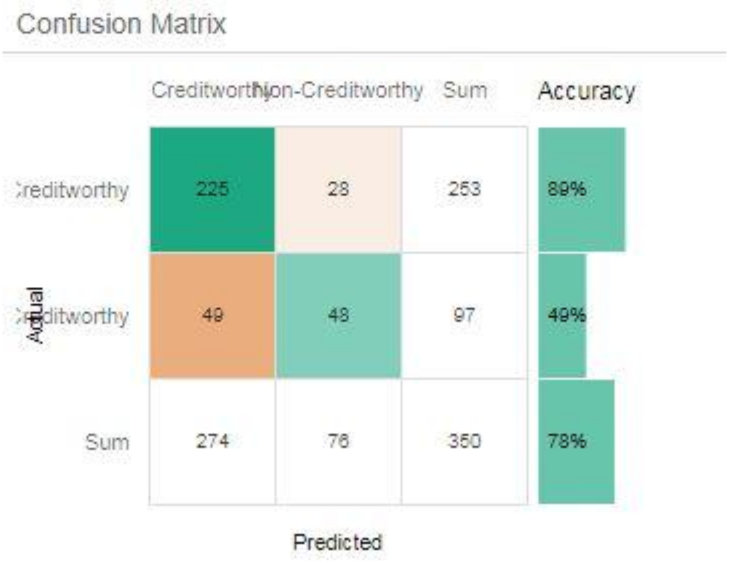
The overall accuracy for the decision tree model is 0.74.



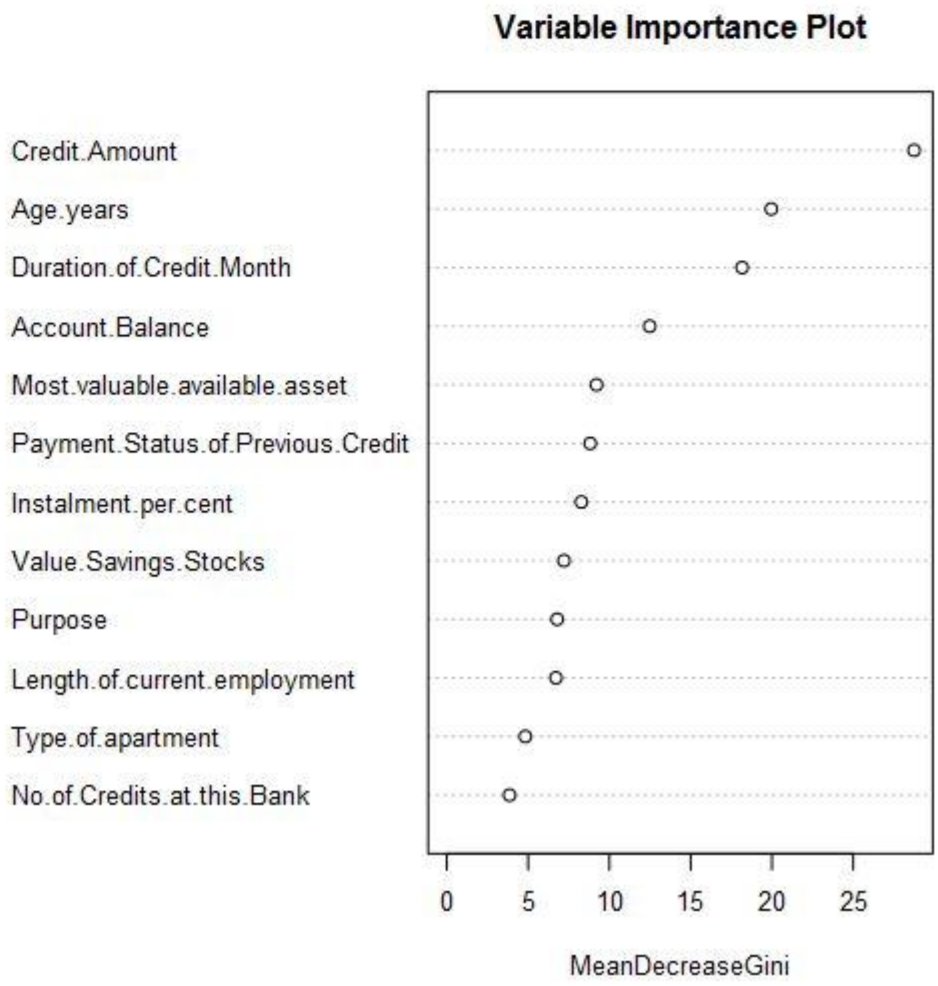
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778

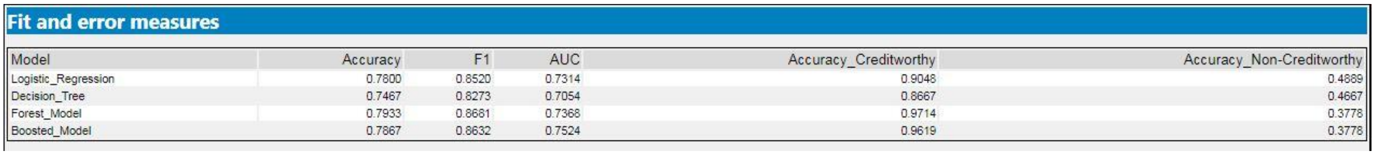
Confusion Matrix for decision tree model.



Forest Model: - ‘Credit Amount’, ‘Age-Years’ and ‘Duration of credit month’ seem important variables



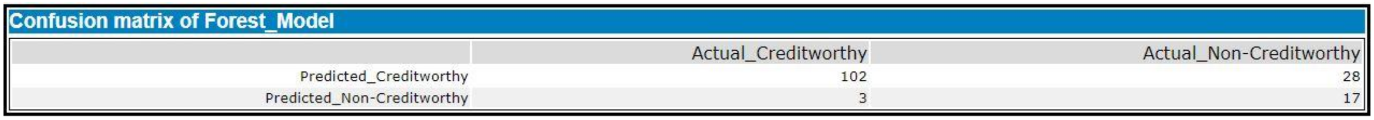
The overall accuracy of the forest model is 0.79



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778

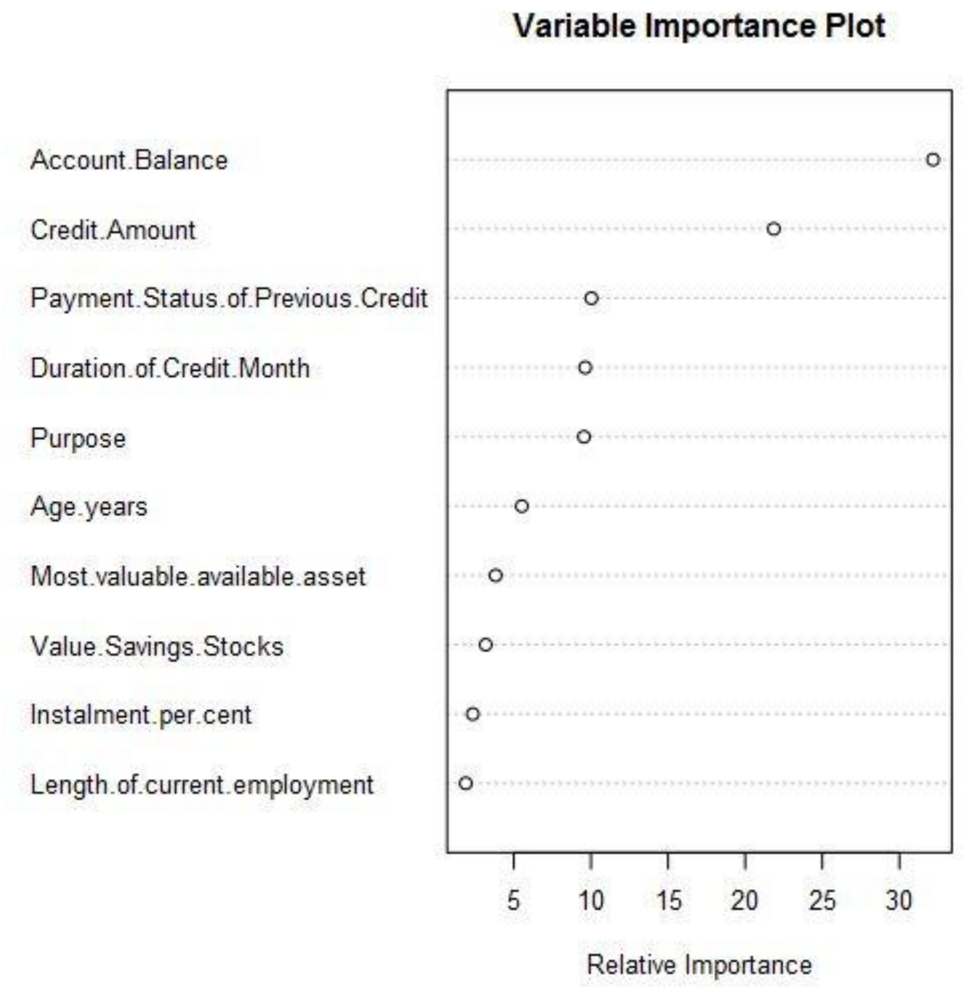
Confusion matrix for the forest model:-



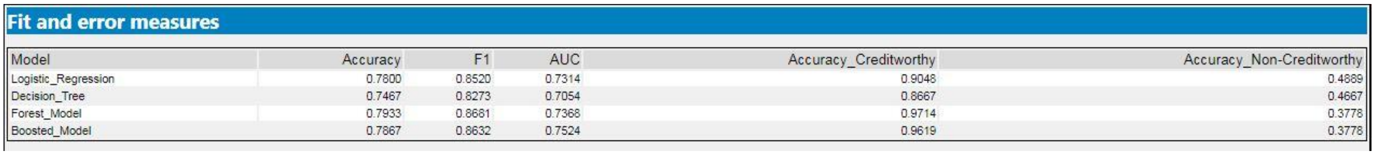
Confusion matrix of Forest_Model

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

Boosted Model: - ‘Account Balance’, ‘Credit Amount’ and ‘Payment Status’ seem like important variables.



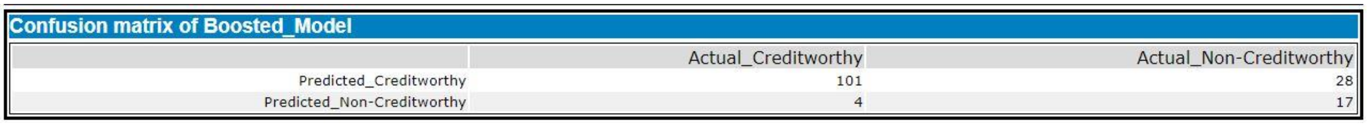
Overall accuracy of the model is 0.76.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix for the model



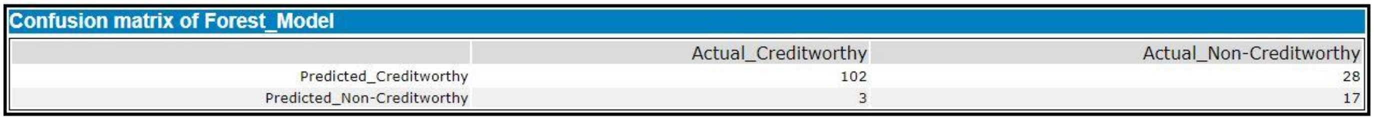
Confusion matrix of Boosted_Model

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

Q6. 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

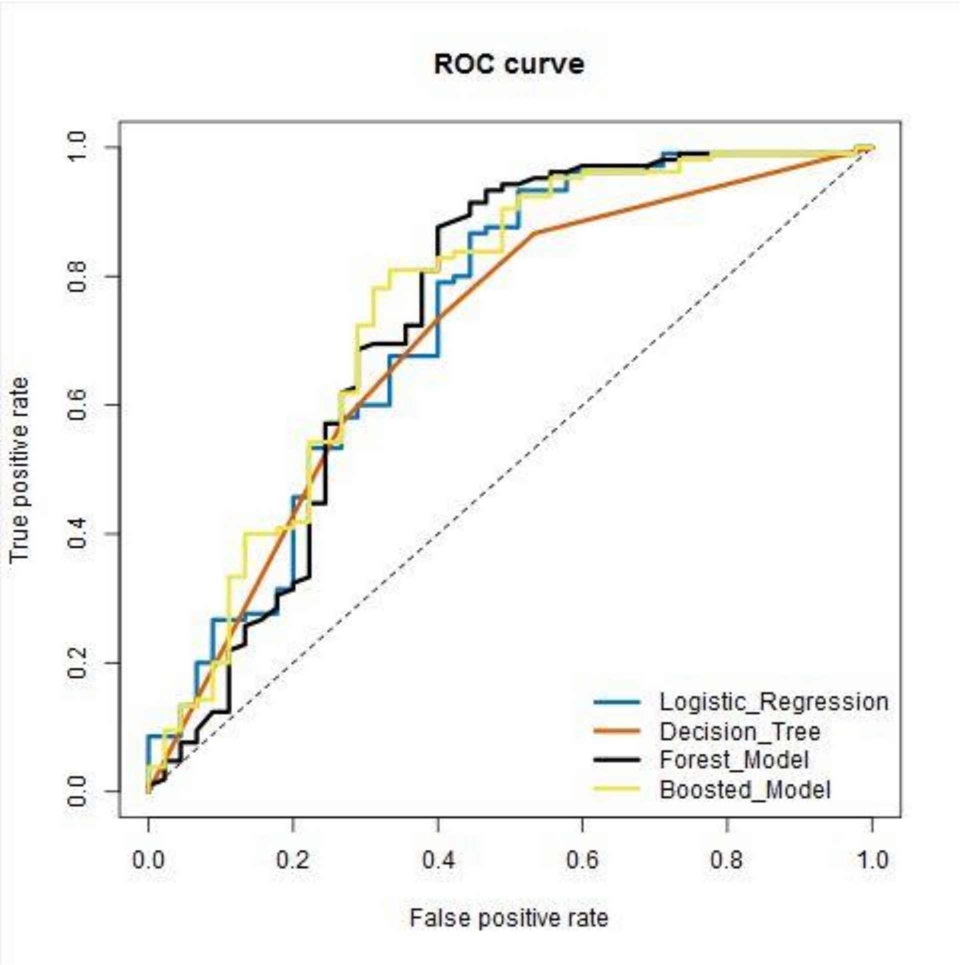
I chose the 'Forest Model' because it has the highest accuracy among the four models (0.79). As shown in the confusion matrix the accuracy for creditworthy is excellent.



Confusion matrix of Forest_Model

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

ROC curve comparison for all the models is given below.



Q7. How many individuals are creditworthy?

I have taken into account all the individuals with creditworthy probability > 0.5. This yielded the number 410. Thus 410 applicants are creditworthy.