Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

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 need to be made?

  The decision involved here is the determination of the credit worthiness of the bank's loan applicants for the purpose of determining which applicants to lend to and which ones to deny their loan request.

- What data is needed to inform those decisions?
  To determine the credit worthiness of a loan applicant we shall have to utilize past data of the bank's past loan applicants. The variables that we shall consider include credit amount, purpose, age of applicant, account balance, duration of credit, previous credit history, occupation status, possession of financial assets, employment status and history.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We are required to determine whether or not an applicant is credit worthy. That is to say we classify an applicant into one of two categories(credit worthy or credit unworthy). Therefore this necessitates the construction of a binary classification model.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

<u>Here are some guidelines to help guide your data cleanup:</u>

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

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

I used the Alteryx Association Analysis tool to check for inner correlation between predictor variables and upon observing the correlation matrix with scatter plot no variables came out as having high correlation between them.

The "***Duration_in_Current_Address***" field had 69% of values missing. Therefore it was best to remove the field.

I utilized the Alteryx field summary and frequency table tools to check for low variability. The graphs generated indicated the following fields had low variability:

Guarantors – 91.39%(None) and 8.61%(Yes)
Concurrent Credits – 100%(Other Banks/Debts)

Graph generated by the field summary tool showed the ***"Foreign Worker"*** field as having low variability.
Though the ***"Occupation"*** field didn't appear in any results generated by the two tools, using the summarize tool Group by function indicated that this field had only a single value(1) thus low variability.
I removed all the variables with low variability.

I removed the ***"Telephone"*** and ***"No_of_dependants"*** fields on the grounds that they had no logical relation with the target variable.

<u>Review Correction</u>:

The **Age-years** field had a few null values and using the Alteryx Imputation tool I imputed data using the **median** of the entire Age-year field.

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.

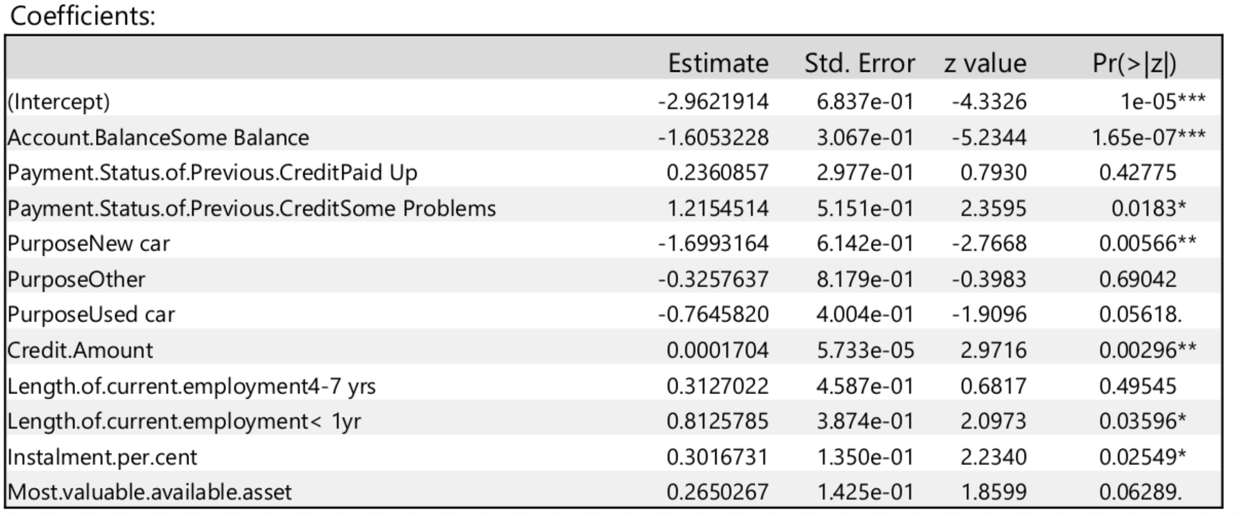
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

Logistic Regression using Stepwise tool:

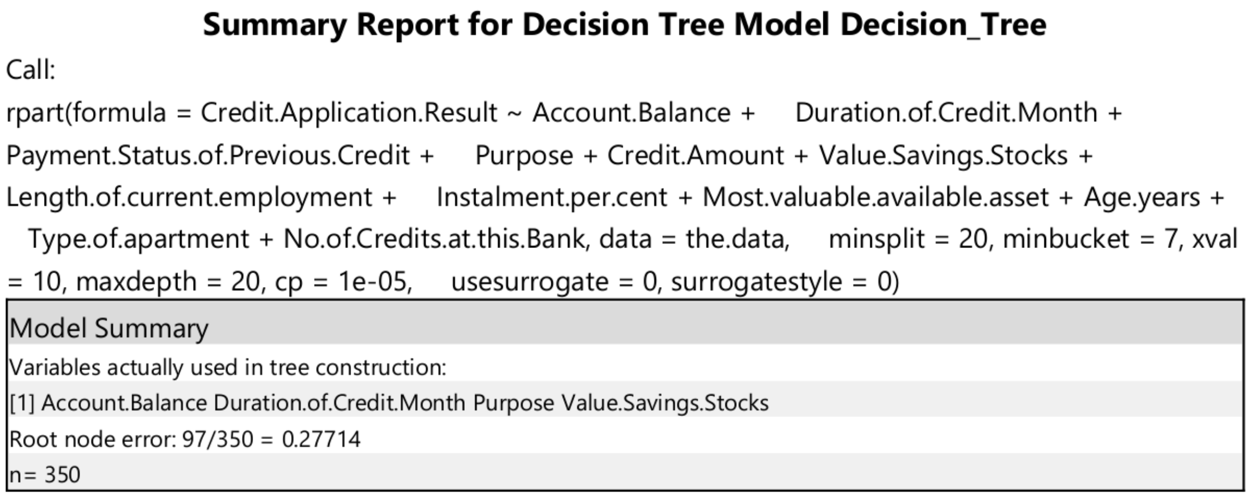
Account-Balance, Payment_Status_of_Previous_Credit, Purpose, Credit_Amount, Length_of_current_employment, and Installment_per_cent variables are significant to the Logistic Regression model as depicted by the report diagram below.



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.

Decision Tree



Summary Report for Decision Tree Model Decision_Tree

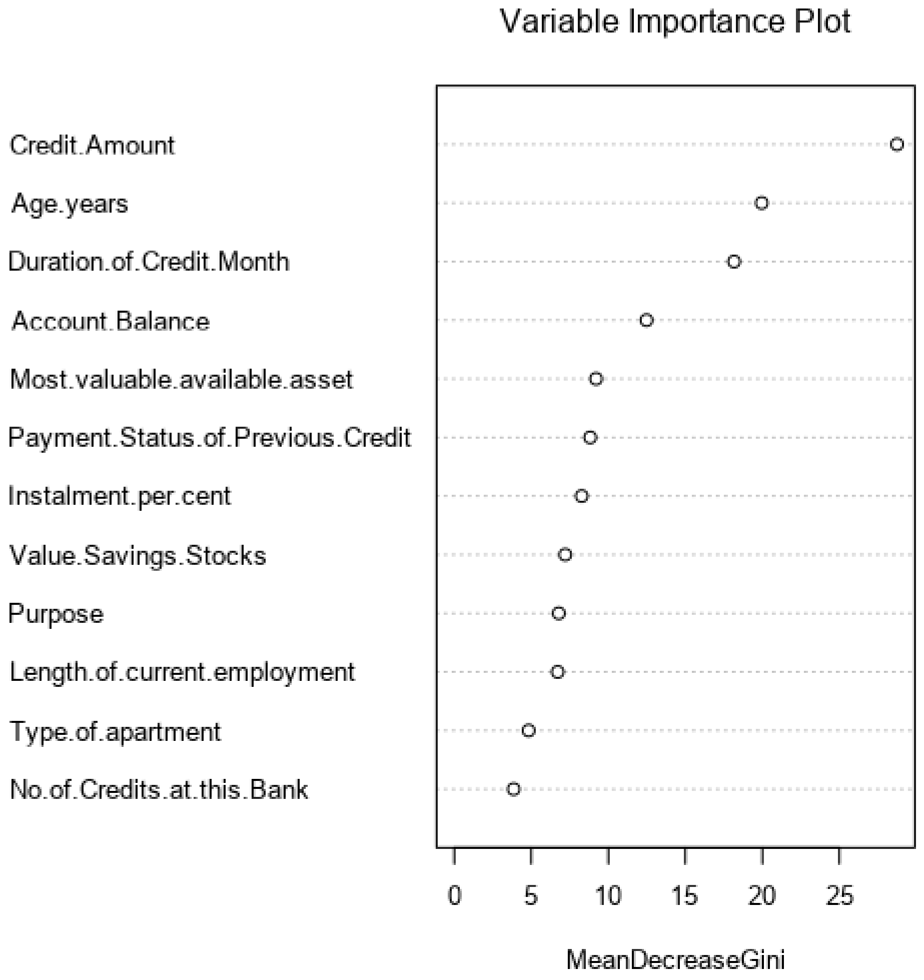
Call:

```
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 + Age.years +
  Type.of.apartment + No.of.Credits.at.this.Bank, 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 Purpose Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

For the Decision tree Account_Balance, Duration_of_Credit_Month, Purpose, and Value_Savings_Stocks were important for the model as shown by the report above.

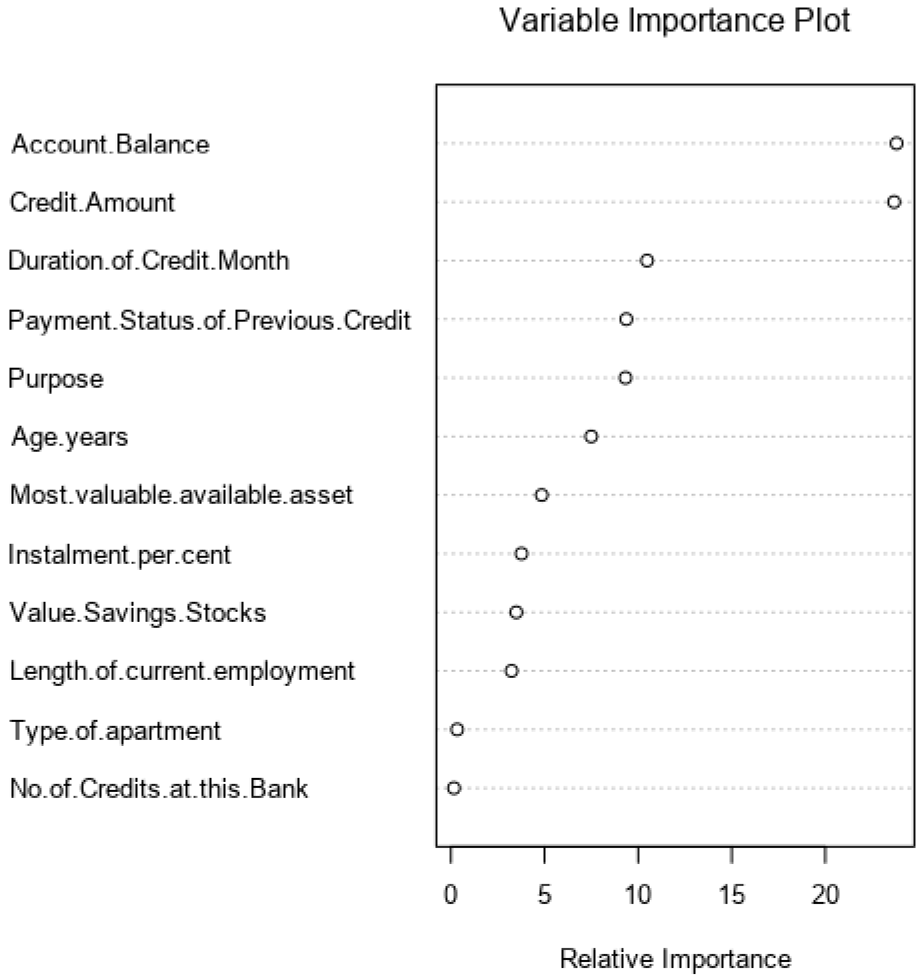
Forest Model Variable Importance Plot



The Variable Importance Plot depicts the order in which the variables were significant to the forest model with the top most variable being the most significant.

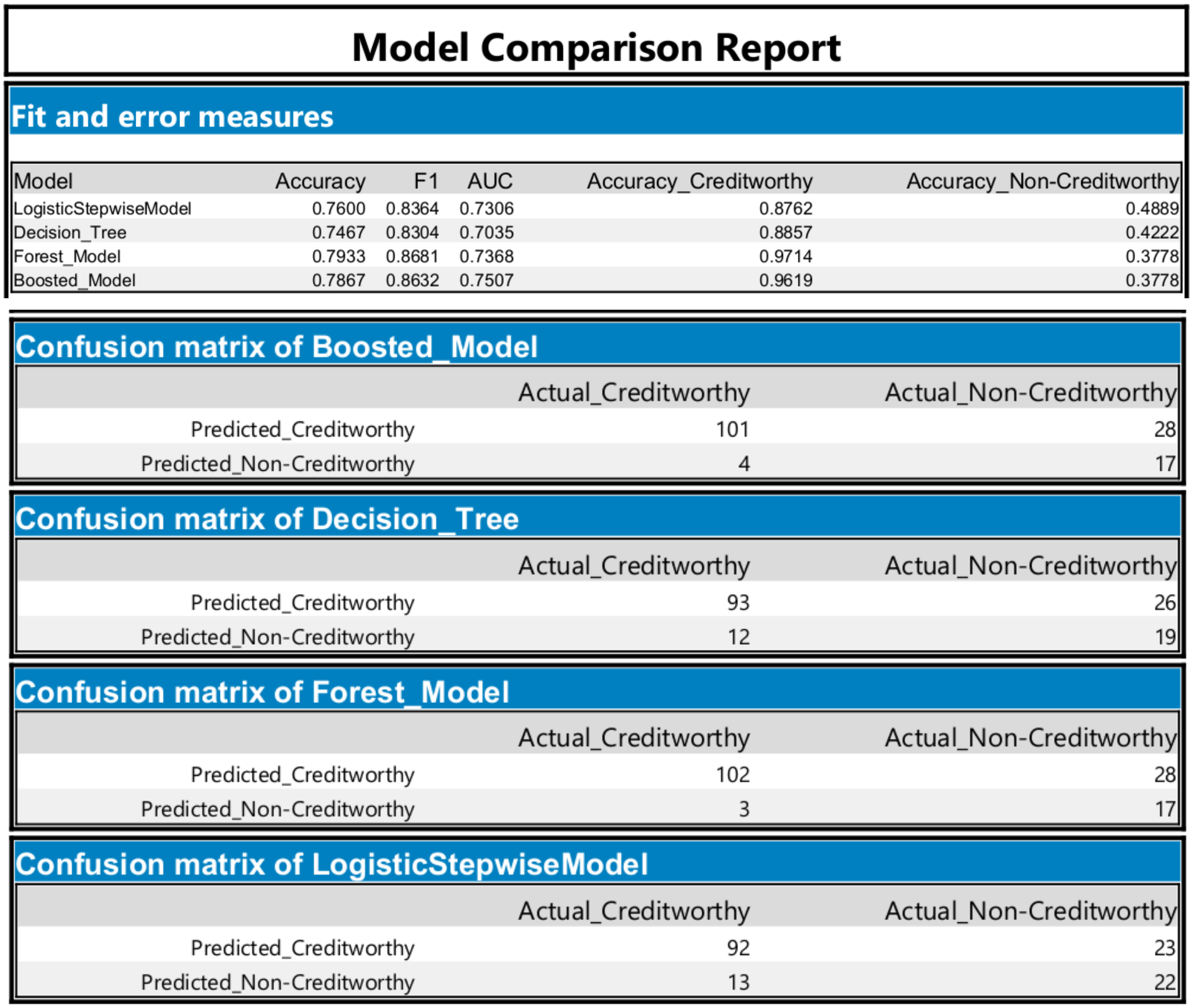
Boosted Model Variable Importance Plot

The Variable Importance Plot depicts the order in which the variables were significant to the Boosted model with the top most variable being the most significant.



- Validate your model against the Validation set. What was the overall percent accuracy? how the confusion matrix. Are there any biases seen in the model's predictions?

  The diagrams below show the report that was generated by the model comparison tool upon validating all the four models side by side.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticStepwiseModel	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7507	0.9619	0.3778

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Model

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

Confusion matrix of LogisticStepwiseModel

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

Review Correction:

Bias determination:

To determine if a model has bias we calculate the positive predictive value(PPV) and negative predictive value(NPV) and check if the two values are close. If the two values are close then we can conclude that the model has no bias otherwise it is biased.
Boosted Model bais check:

PPV = 101/129 = 78%   NPV = 17/21 = 80%
Conclusion: model has no bias as PPV and NPV are close.

Forest Model bais check:

PPV = 102/130 = 78%   NPV = 17/20 = 85%
Conclusion: model has no bias as PPV and NPV are close.

Decision Tree Model bais check:
PPV = 93/119 = 78%   NPV = 19/31 = 61%
Conclusion: model has bais as PPV and NPV are far apart

Logistic Regression Model bias check:
PPV = 92/115 = 80%   NPV = 22/35 = 62%
Conclusion: model has bais as PPV and NPV are far apart.

You should have four sets of questions answered. (500 word limit)

Step 4: Write up

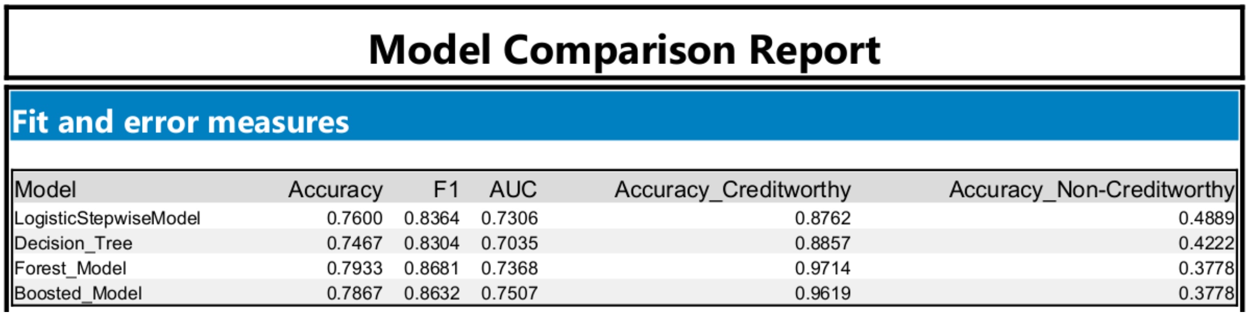
Answer these questions:

- 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

I chose the Forest Model to carry out the prediction.

The Forest Model provides the highest overall accuracy(79.33%) and the averaged results from the forest model helps deal with the decision tree model's bias to over-fit the data.

Accuracies within the Forest model "Creditworthy"(97.14%) and "Non-Creditworthy"(37.78%). Though the forest model has the least accuracy together with the boosted model in predicting "Non-Creditworthy" this is balanced out by it having the highest accuracy in predicting "Creditworthy" outcomes by a greater margin. The Forest Model had the most correctly predicted values(119) and least wrongly predicted values(31).



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticStepwiseModel	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7507	0.9619	0.3778

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

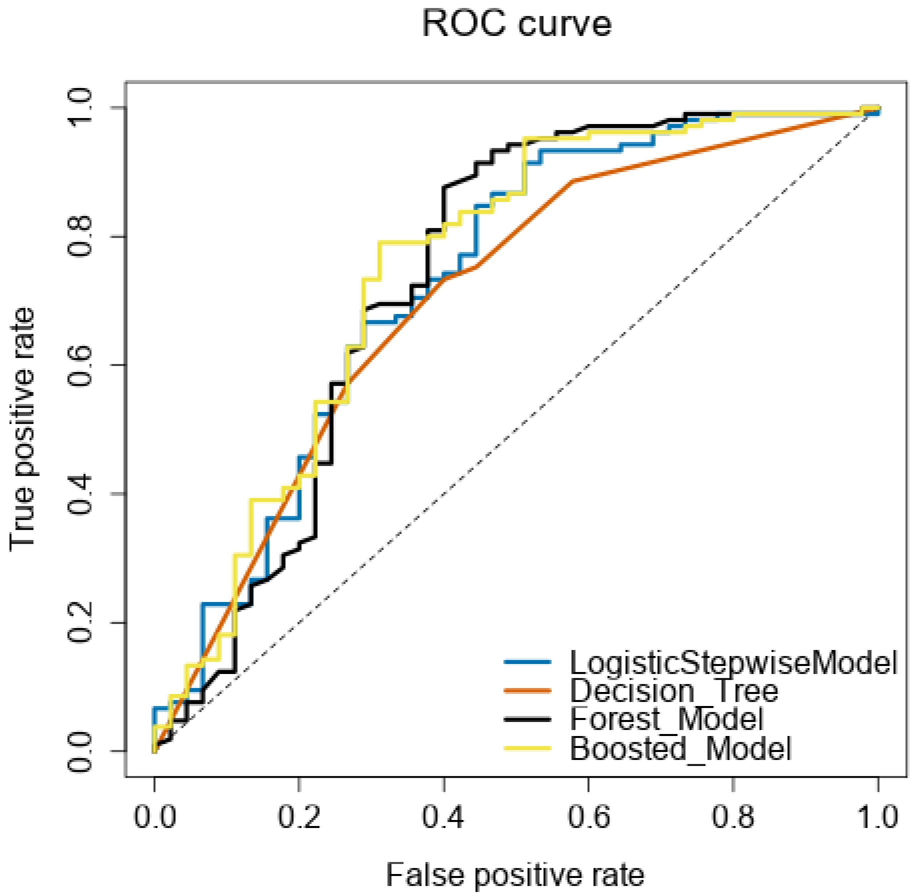
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Forest_Model

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

Confusion matrix of LogisticStepwiseModel

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



The ROC curve classifier that gives the curve closer to the top-left corner indicates a better performance. In our case that happens to be the Forest model curve.

Upon calculating the Forest model PPV(78%) and NPV(85%) and observing that the values were close to each other we concluded that the model had no bias.

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

- How many individuals are creditworthy?
  Creditworthy – 408 individuals
  Non-Creditworthy – 92 individuals

<u>Before you Submit</u>

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.