Project: Creditworthiness

Loay Alhamwi (LOAY.HAMWI@GMAIL.COM)

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?
  Answer
  Identify the list of creditworthy of new loan applicants then approve them.

- What data is needed to inform those decisions?
  Answer
  1- Account-Balance
  2- Duration-of-Credit-Month
  3- Payment-Status-of-Previous-Credit
  4- Credit-Amount
  5- Purpose
  6- Value-Savings-Stocks
  7- Length-of-current-employment
  8- Instalment-per-cent
  9- Most-valuable-available-asset
  10- Age-years
  11- Type-of-apartment
  12- No-of-Credits-at-this-Bank

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Answer
  The model is Binary where the result of each customer is either **approved** (creditworthy) or **not approved** (non-creditworthy)

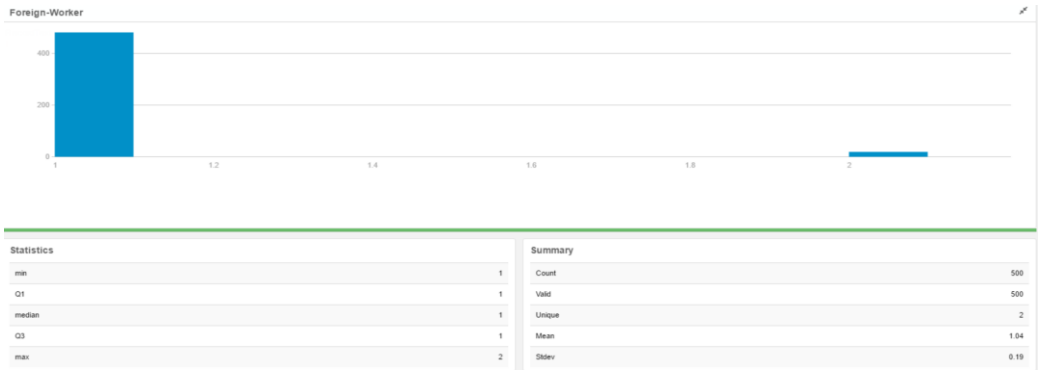
Step 2: Building the Training Set

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.***

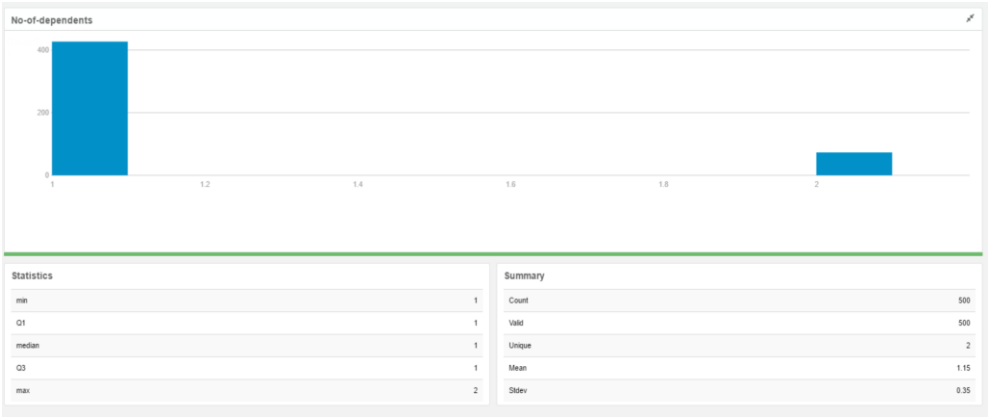
Answer this question:

In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
Answer
<u>Step1 - remove the following fields</u>

1. Foreign-Worker, it is low variability, this variable shows that the majority of the data is skewed towards "1" 481 records while the remaining 19 records "3".

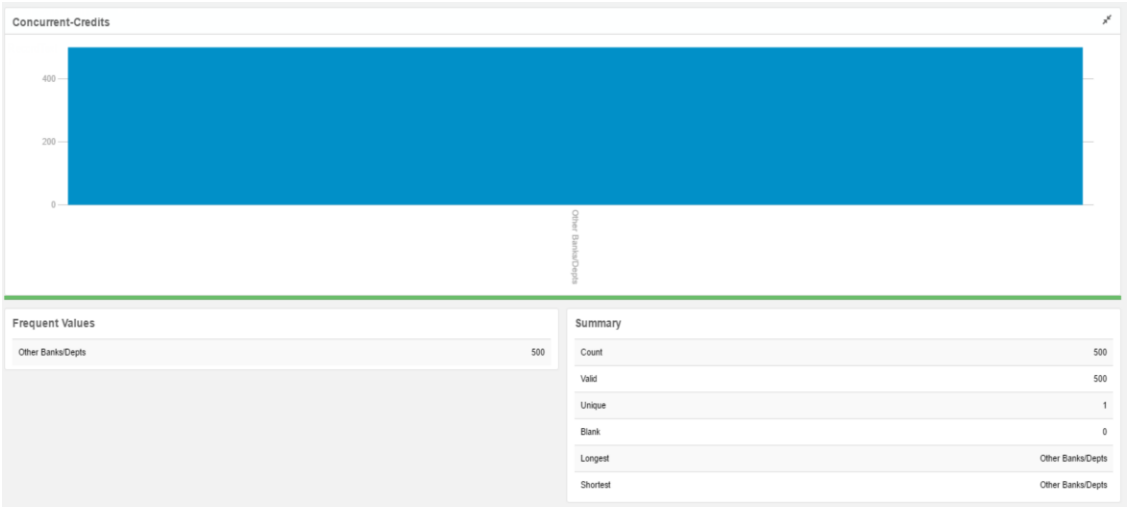


2. No-of-dependents, low variability, this variable show that the majority of the data is skewed towards "1" 427 records, while the remaining is 73 records "2".

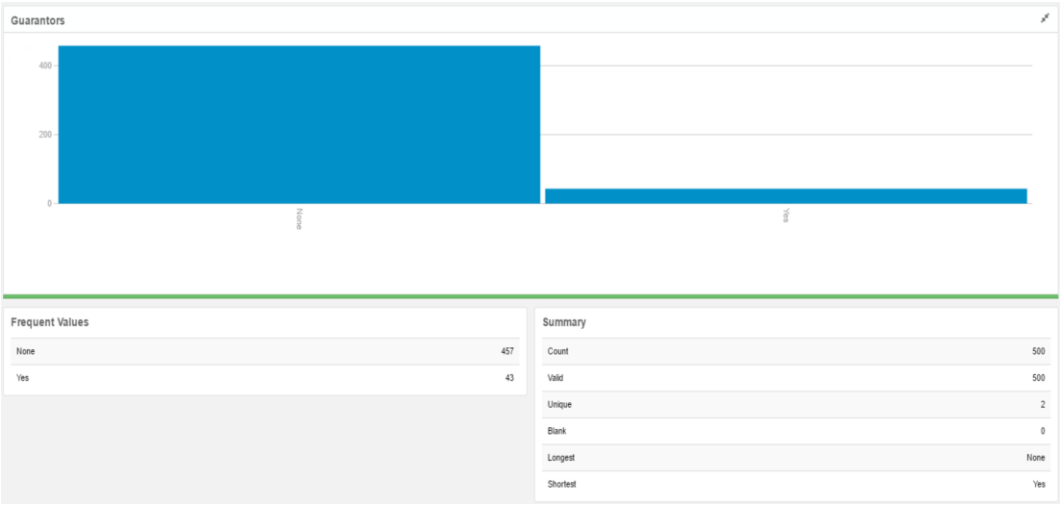


3. Occupation, it is low variability, variable only contains one value is 1.

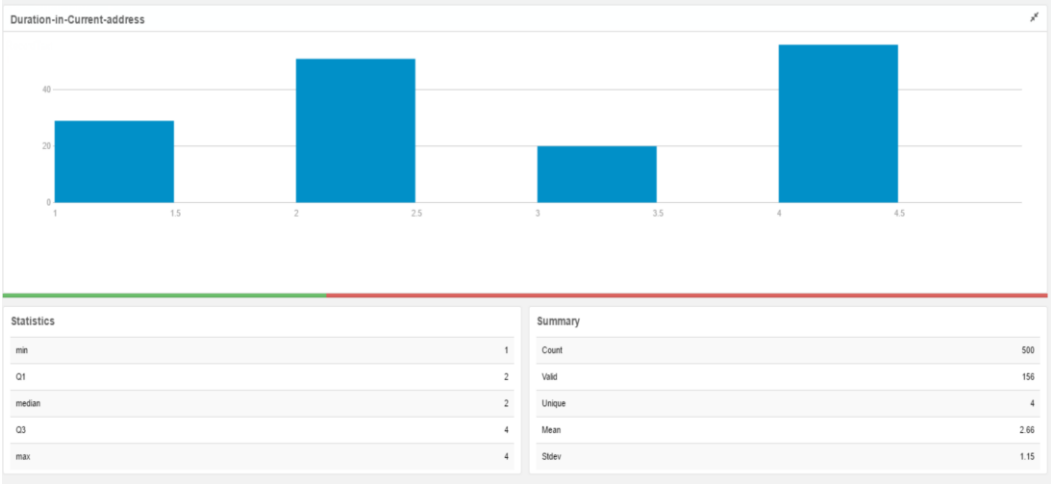
4. Concurrent-Credits, It is low variability, variable only contains one value “Other Banks/Depts”



5. Guarantors, it is low variability, this variable shows that the majority of the data is skewed towards “None” 457 records, while the remaining 43 records “Yes”



6. Duration-in-Current-address, 344 empty records of 500.



7. Telephone because we have two unique values only.

Step 2- Impute the null values of Age-years

The median of Age-years is 33

After cleaning the data, the final data set is 13 columns.

Step 3: Train your Classification Models

- 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

  As per the table below, the predictor variables are significant where P value < 0.05
    - Account.BalanceSome Balance
    - Payment.Status.of.Previous.CreditSome Problems
    - PurposeNew car
    - Credit.Amount
    - Length.of.current.employment< 1yr
    - Instalment.per.cent

Record Report

Report for Logistic Regression Model Stepwise_CreditWorthy

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

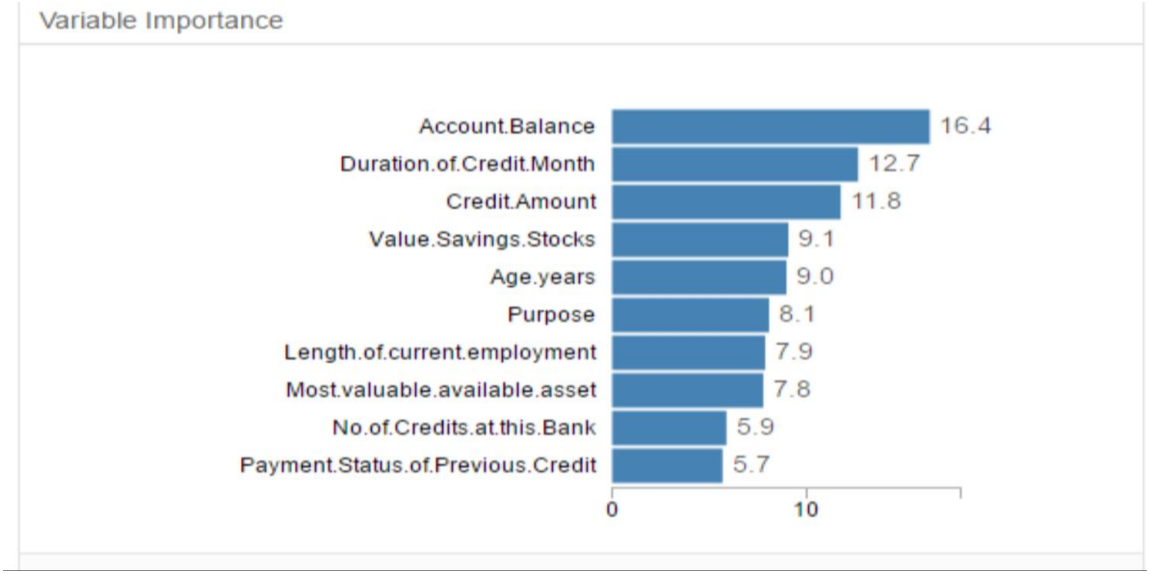
(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

DECISION TREE MODEL

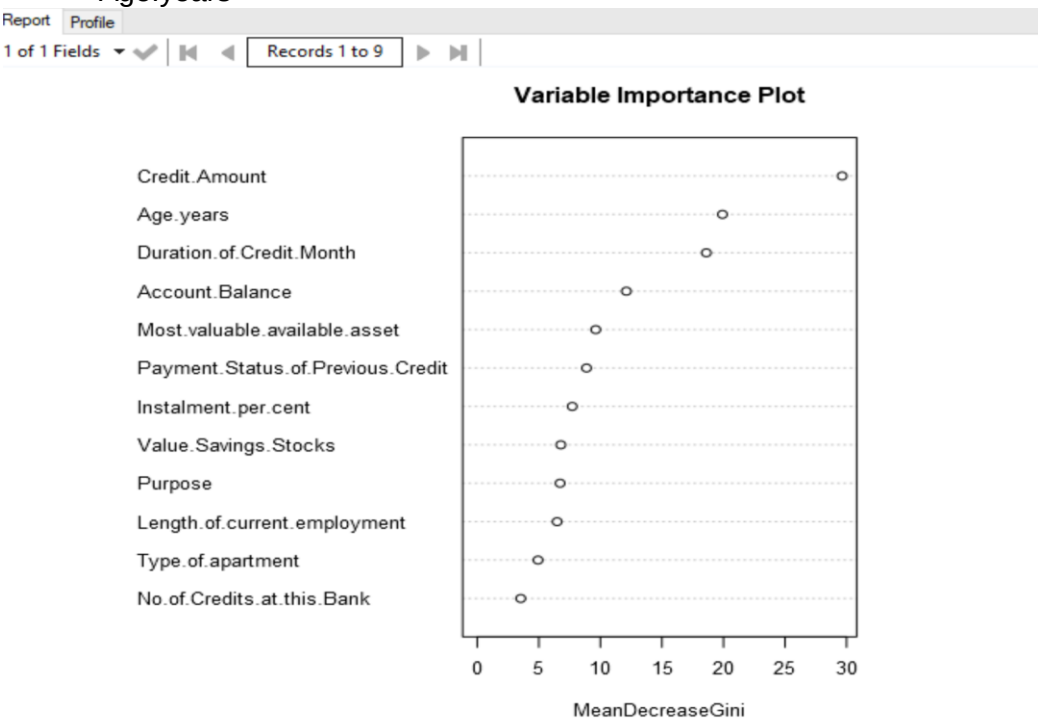
The most important variables are:

- Account.Balance
- Duration.of.Credit.Month
- Credit.Amount



FOREST MODEL

- The most important variables are:
- Credit.Amount
- Age.years



Boosted Model

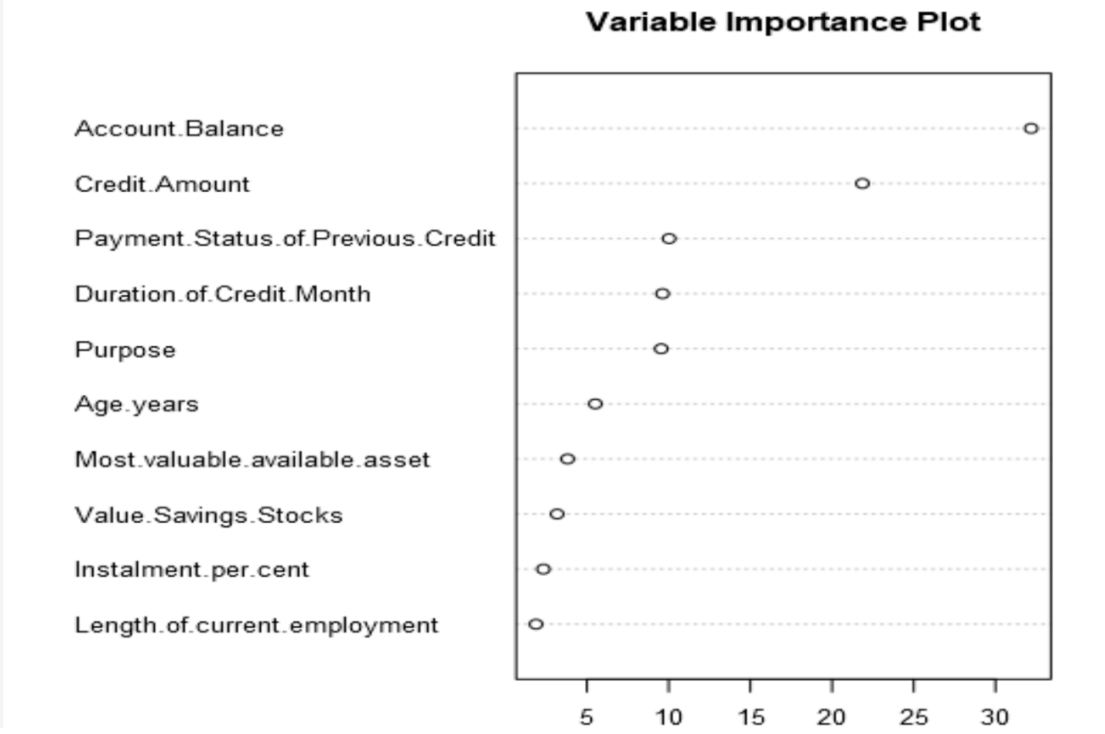
The most important variables are:

- Account.Balance
- Credit.Amount

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



- 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?
  As per the table below taken from model Comparison Report, the accuracy is:
  Stepwise Model: 0.7600
  Decision_Tree_Model: 0.6667
  Boosted_Model: 0.7867
  Forest_Model: 0.8067

  There are bias in all models, in the Forest_Model we have 102 records that are predicted Creditworthy that were actually Creditworthy, yet we have 26 records that were predicted Creditworthy that were actually Non-Creditworthy.

We have 3 records that are predicted Non-Creditworthy that were actually Creditworthy, yet we have 19 records that were predicted Non-Creditworthy that were actually Non-Creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Model	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_Tree_Model	0.6667	0.7685	0.6272	0.7477	0.4359
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095
Forest_Model	0.8067	0.8755	0.7382	0.7969	0.8636

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of Forest_Model

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

Confusion matrix of Stepwise_Model

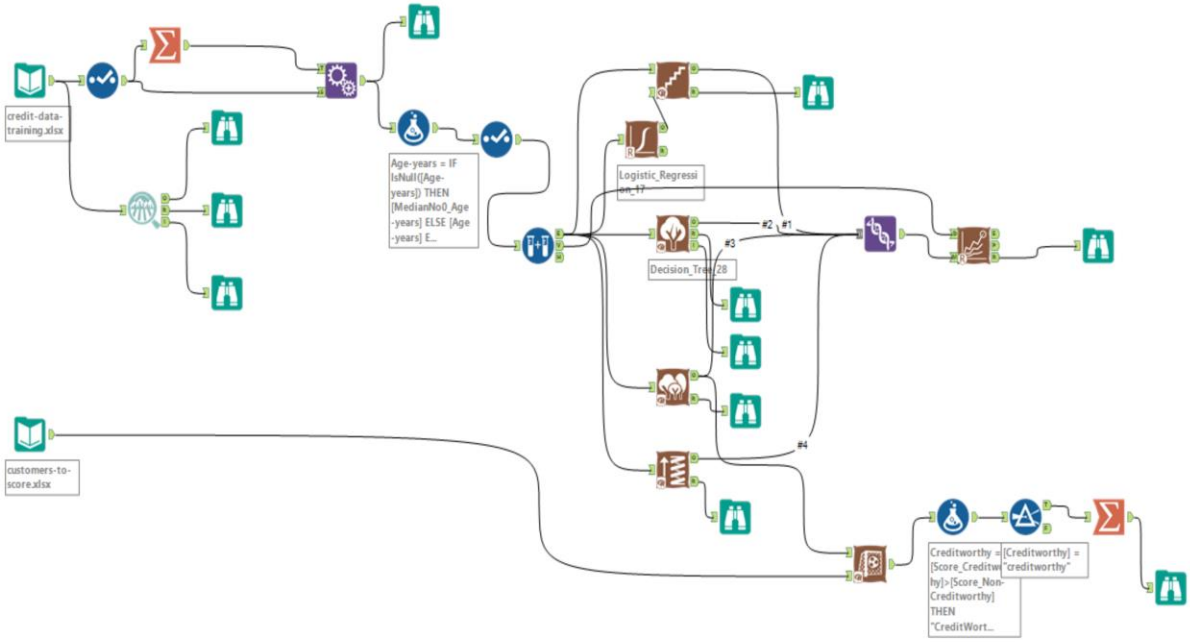
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Step 4: Writeup

Answer:

Below the steps which I fellow

1- Creating analytical data set where I removed 7 columns and cleaned the data and imputed the Age-year
2- Created Sample where estimation sample is 70% while validation sample is 30%
3- Added and configured Logistic regression model and step wise tool.
4- Added and configured Decision Tree Model
5- Added and configured Forest Model
6- Added and configured Boosted Model
7- Union All models then compare the models using the tool Model Comparison



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

    Answer:

    The most accurate model is Forest Model where the accuracy is 0.8067 is greater than the accuracy of other models, because of this I decided to use the Forest Model.

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments

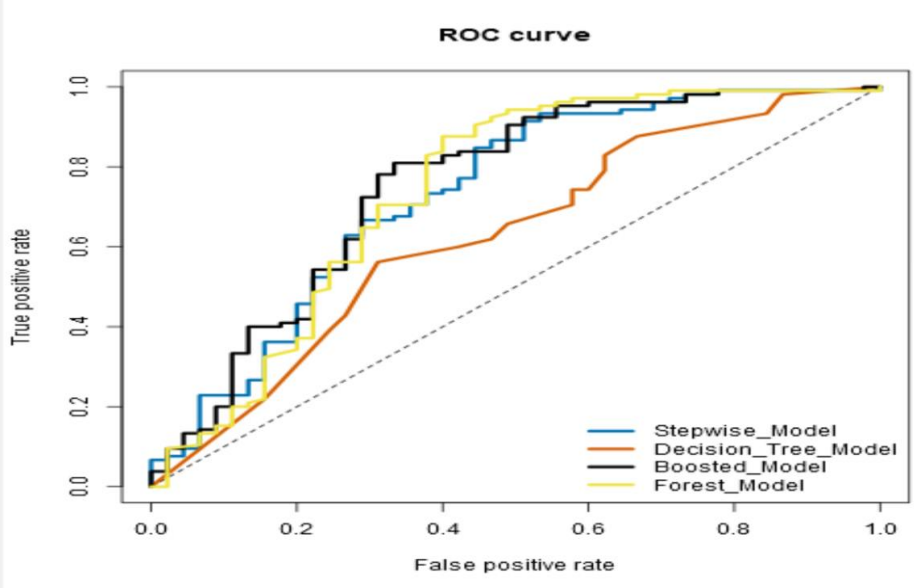
Answer:

Forest Model, The accuracy of Creditworthy is 0.7969 is greater than all models except stepwise model but the overall accuracy is greater than all models while the accuracy of Non-Creditworthy is 0.8636 is greater than the other models, I see the Forest model is quite strong compared to other models.

- ROC graph

Answer:

It is a diagnostic plot, the yellow line represents the Forest model, and it performs slightly better than other models



- Bias in the Confusion Matrices

Answer:

We have 102 records that are predicted Creditworthy that were actually Creditworthy, yet we have 26 records that were predicted Creditworthy that were actually Non-Creditworthy.

We have 3 records that are predicted Non-Creditworthy that were actually Creditworthy, yet we have 19 records that were predicted Non-Creditworthy that were actually Non-Creditworthy.

Confusion matrix of Forest_Model

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

- How many individuals are creditworthy?

  Answer: 408

  Here is the steps to score new data set
  - Added Score tool where the first input is R model object produced by Forest Model and data input of records.
  - Creating new column to indicates that the person is creditworthy or not creditworthy depending on the formula “IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN "CreditWorthy" ELSE "Not_CreditWorthy" ENDIF”
  - Count the individuals who are creditworthy using the Filter and Summarize tool