Project – Predicting Default Risk

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

- What decisions needs to be made?

Ans – There was a huge financial scandal that had hit a competitive bank a week ago and due to which there was a sudden influx of new people applying for loans at the bank. Due to the after effect there were nearly 500 loan applications to process this week as opposed to a typical 200 loan applications that is generally done per week. This new influx of loans has given the bank a great opportunity of using a process to predict the credit worthiness of the new loan applicants.

Thus the objective is to identify and predict whether customers who applied for loans are creditworthy enough to be extended a new loan or not.

- What data is needed to inform those decisions?

Ans – The data needed will the past data which is fed from the file “credit-data-training.xls”. The data needs to be checked for missing data and will be used to train the different models. The following are the data that would be used :

Credit-Application-Result
Account-Balance
Duration-of-credit-month
Payment-status-of-previous-credit
Purpose
Credit-amount
Value-savings-stocks
Length-of-current-employment
Most-valuable-available-asset
No.-of-credits-at-this-bank
Type-of-apartment
Instalment-per-cent
Age-years

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

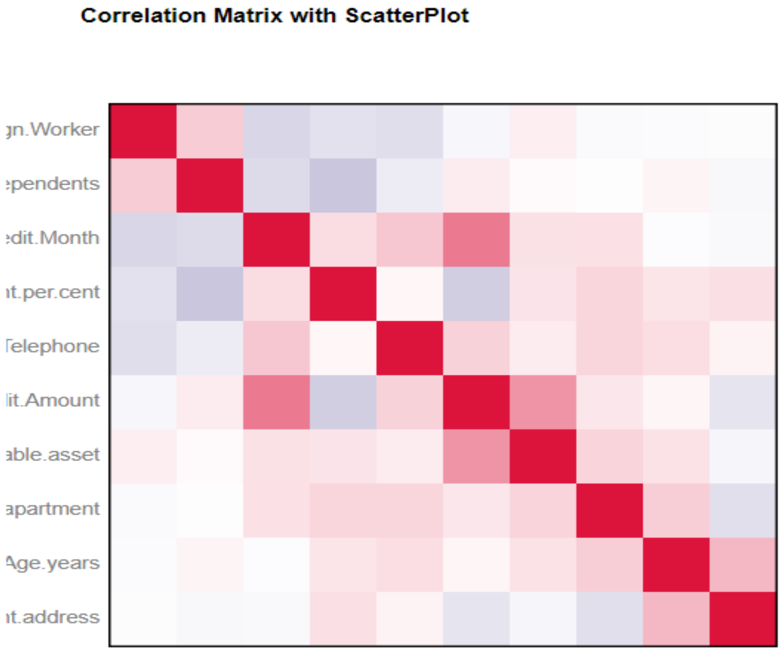
Ans – The Business problem asks to predict an outcome of credit worthiness by using Binary classification models such as logistic Regression, Decision tree, Random Forest Model and Boosted Model of the available data .

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.***

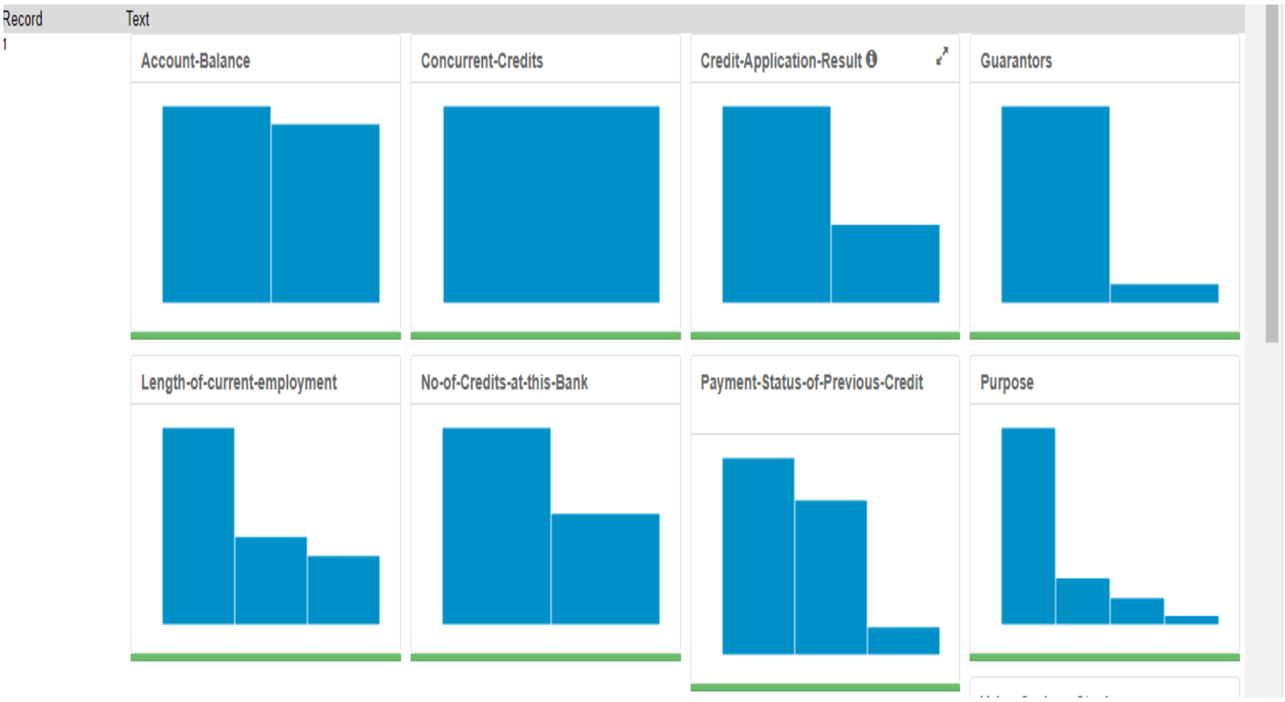
- In your clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

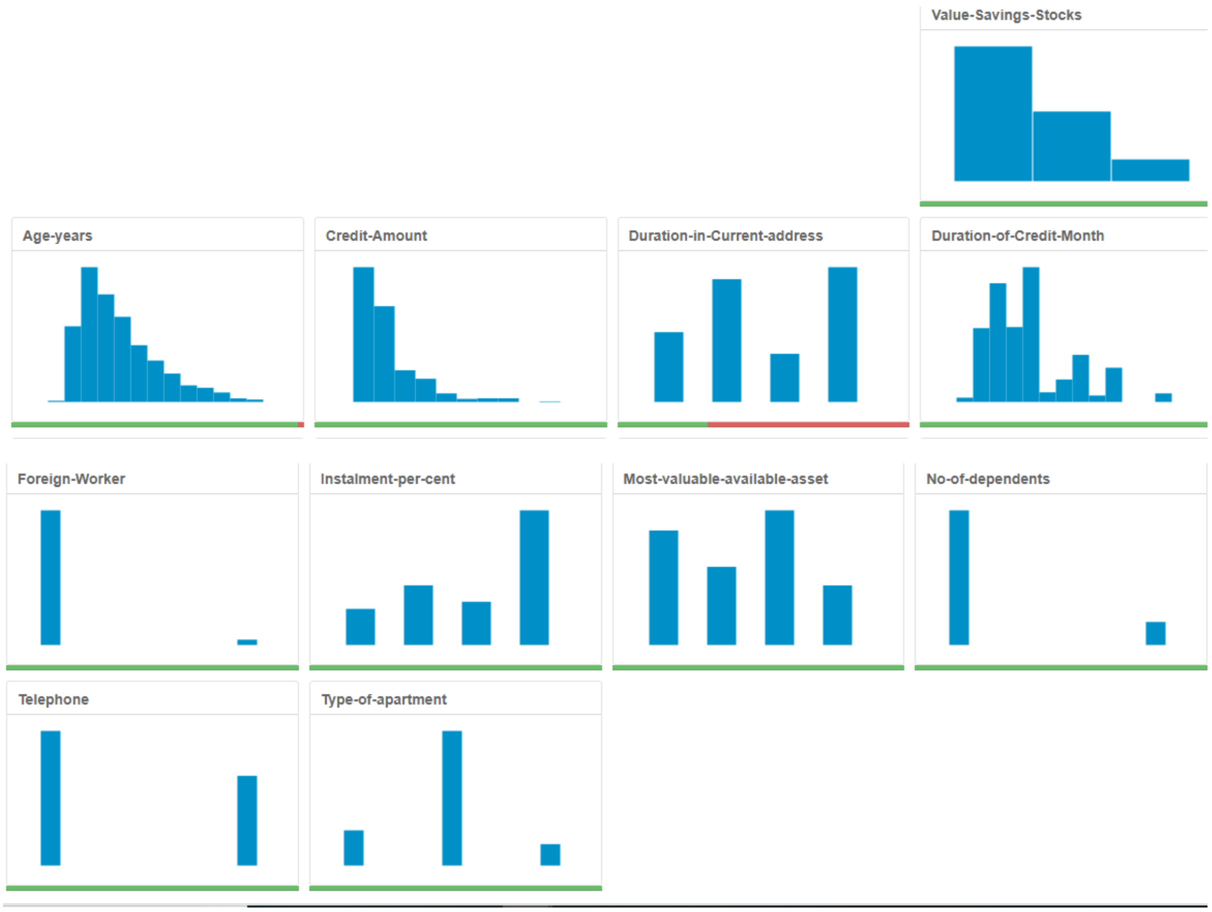
Ans – Association analysis was performed on he numerical variables and it was found that there are no variables which are highly correlated with each other with a correlation of more than 0.7



A field summary was established on the data fields and it was found that Duration-in-current-address has almost 60% missing data values and it was decided to remove the same. While it was also found that Age-years has 2% missing data but it is appropriate to impute the missing data with the median age as data is skewed to the left.

Similarly, concurrent-credits and Occupation has one value while guarantors, Foreign workers, and No. of dependents were removed due to the low variability in order not to skew our analysis. Telephone field also need to be removed due to irrelevancy to the customer credit worthiness.


Record
Text
1
Account-Balance
Concurrent-Credits
Credit-Application-Result
Guarantors
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit
Purpose


Value-Savings-Stocks
Age-years
Credit-Amount
Duration-in-Current-address
Duration-of-Credit-Month
Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset
No-of-dependents
Telephone
Type-of-apartment

Step 3: Train your Classification Models

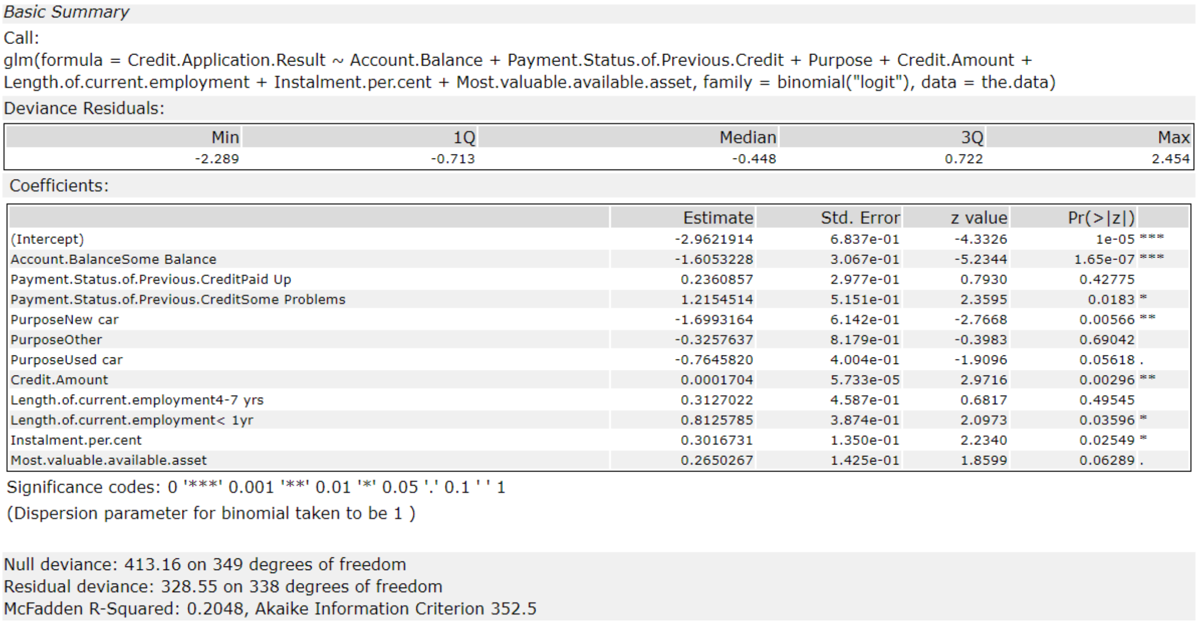
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

*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.
- 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?

A) <u>Logistic Regression (Stepwise)</u>

Credit-Application-Result was used as the target variable and it was found that Account-Balance, Purpose New car and Credit-Amount were most significant with P values <0.05.



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

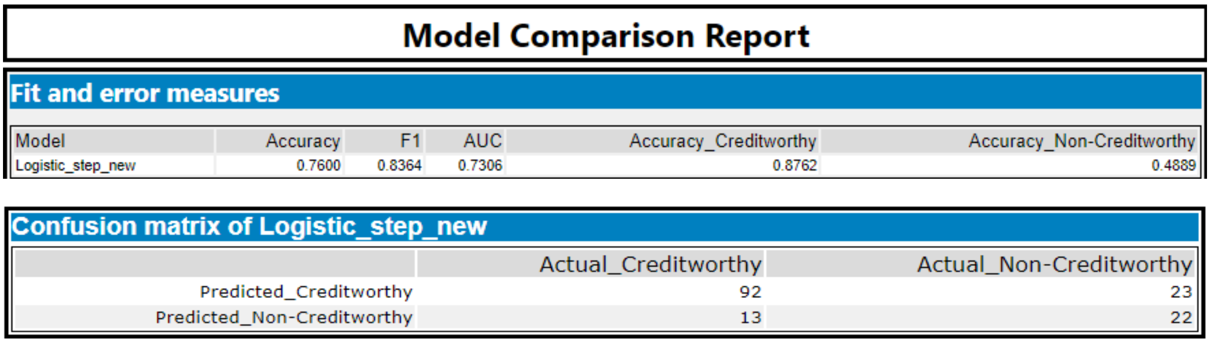
Coefficients:

	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, Akaike Information Criterion 352.5



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_step_new	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Logistic_step_new

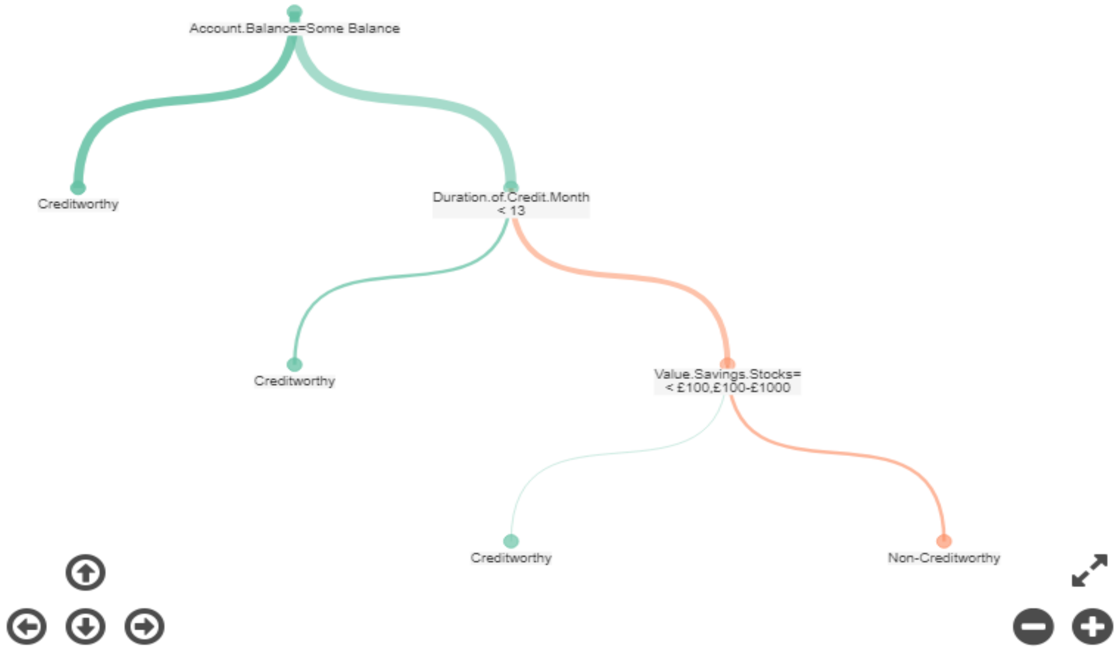
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Overall accuracy was found to be 76% and the accuracy for credit worthy is more than non credit worthy at 87% and 49% respectively.

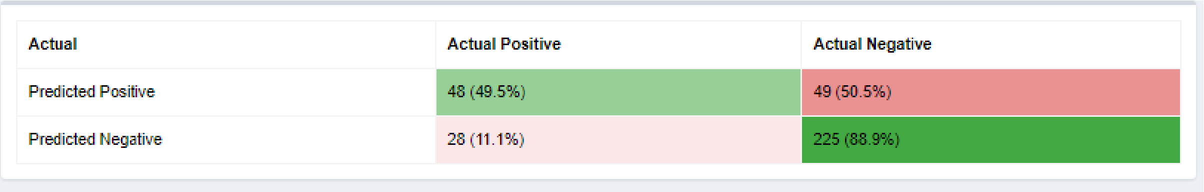
B. Decision Tree

Using Credit-Application result as the target variables, Account Balance, value savings stocks and Duration of credit month are the top 3 most important variables. The overall accuracy is 74.67 %.

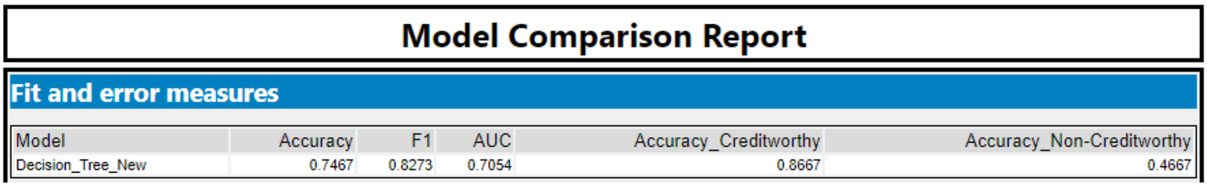
Accuracy for credit worthy is 86% and non-credit worthy is 46 %   .


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

RPart Decision Tree Classification



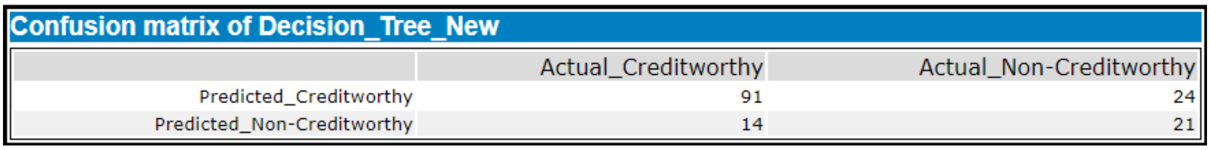
Actual	Actual Positive	Actual Negative
Predicted Positive	48 (49.5%)	49 (50.5%)
Predicted Negative	28 (11.1%)	225 (88.9%)



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_New	0.7467	0.8273	0.7054	0.8667	0.4667



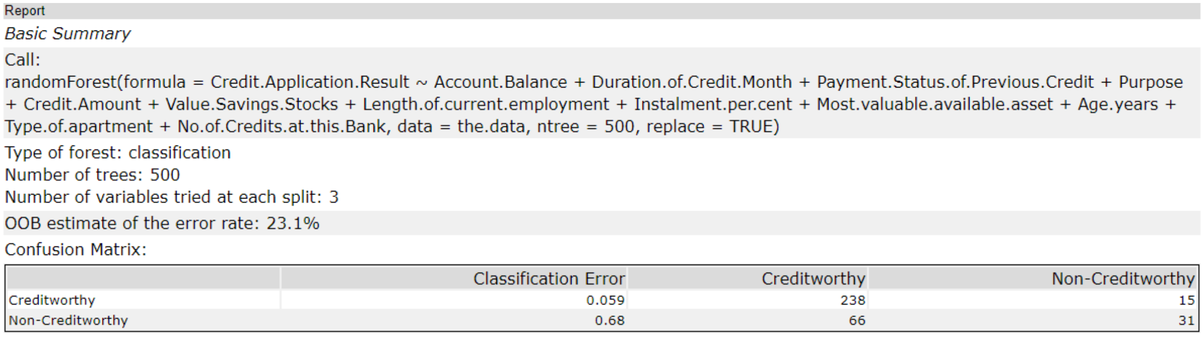
Confusion matrix of Decision_Tree_New

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

3. Forest Model

Using Credit application result as the target variable it was found that Credit amount, Age years and Duration of credit month were the top 3 significant variables.

The overall accuracy was found to be 80% and the accuracy for credit worthy and non-credit worthy was 96 % and 42% respectively.



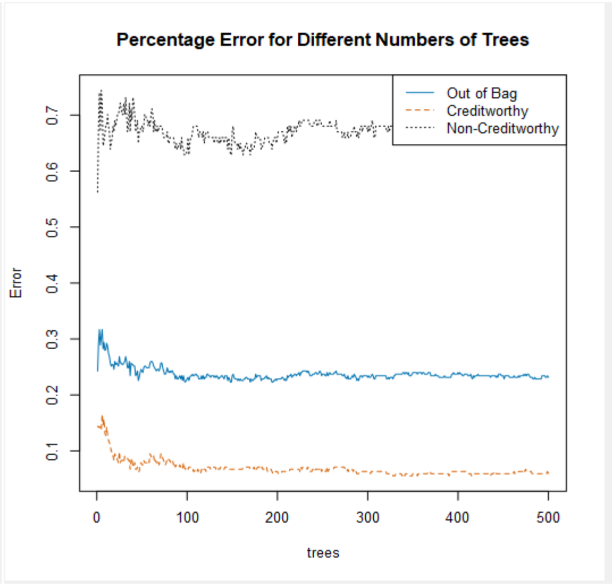
Report

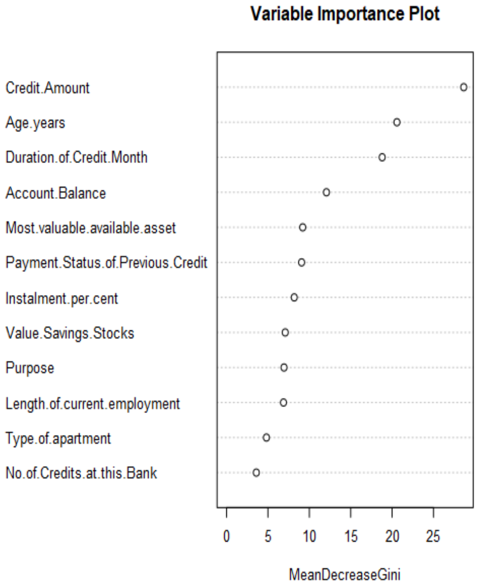
Basic Summary

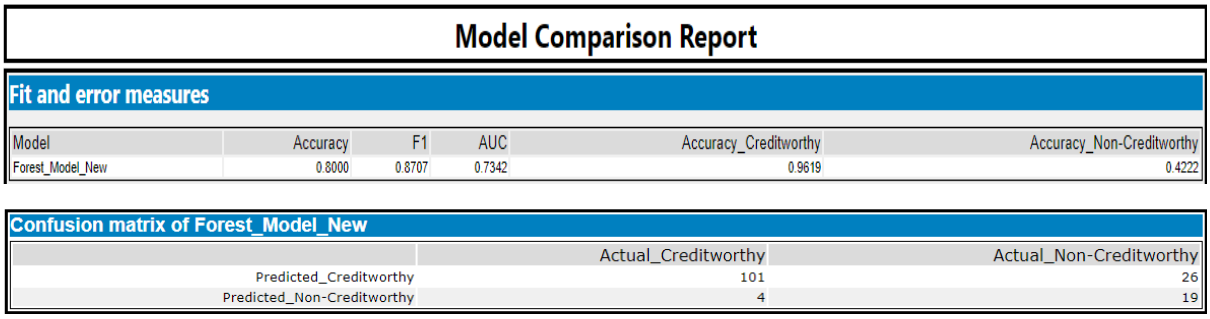
```
Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose
+ Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years +
Type.of.apartment + No.of.Credits.at.this.Bank, 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: 23.1%
Confusion Matrix:
```

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.059	238	15
Non-Creditworthy	0.68	66	31

Plots







Model Comparison Report

Fit and error measures

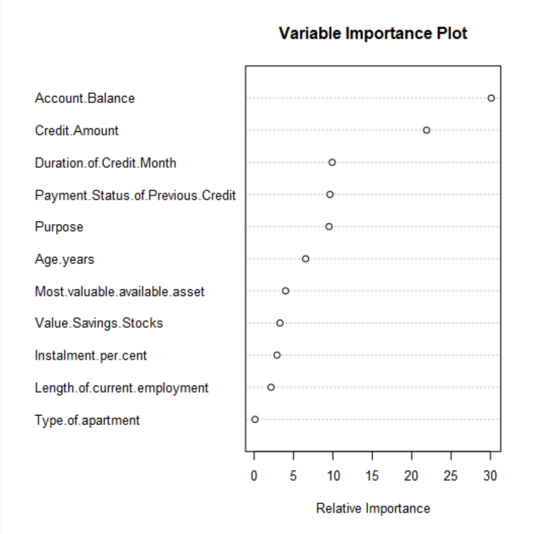
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model_New	0.8000	0.8707	0.7342	0.9619	0.4222

Confusion matrix of Forest_Model_New

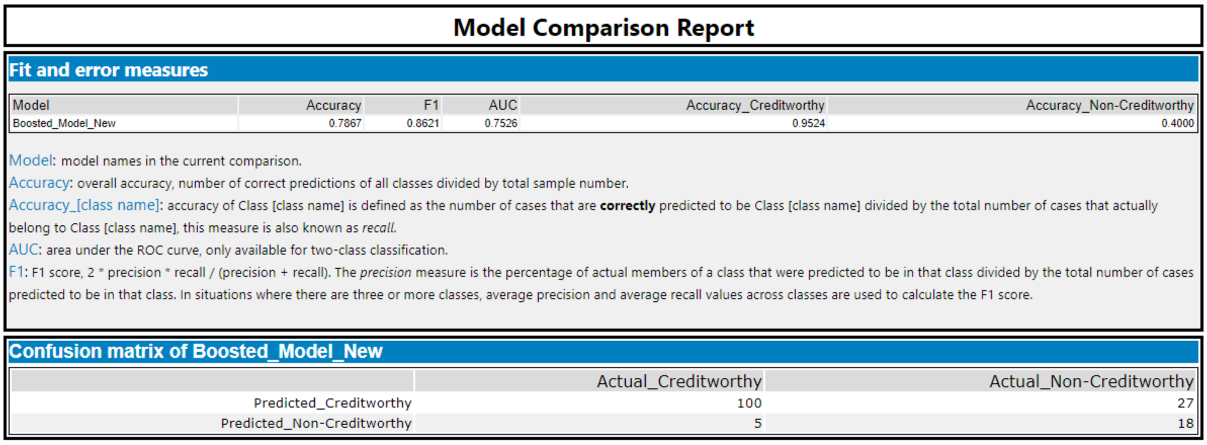
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

4. Boosted Model

Account balance, credit amount and Duration of credit month are top 3 significant variables. Overall accuracy is 78.67%. Accuracies for credit worthy and non-credit worthy are 95% and 40% respectively.



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.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model_New	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_New

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

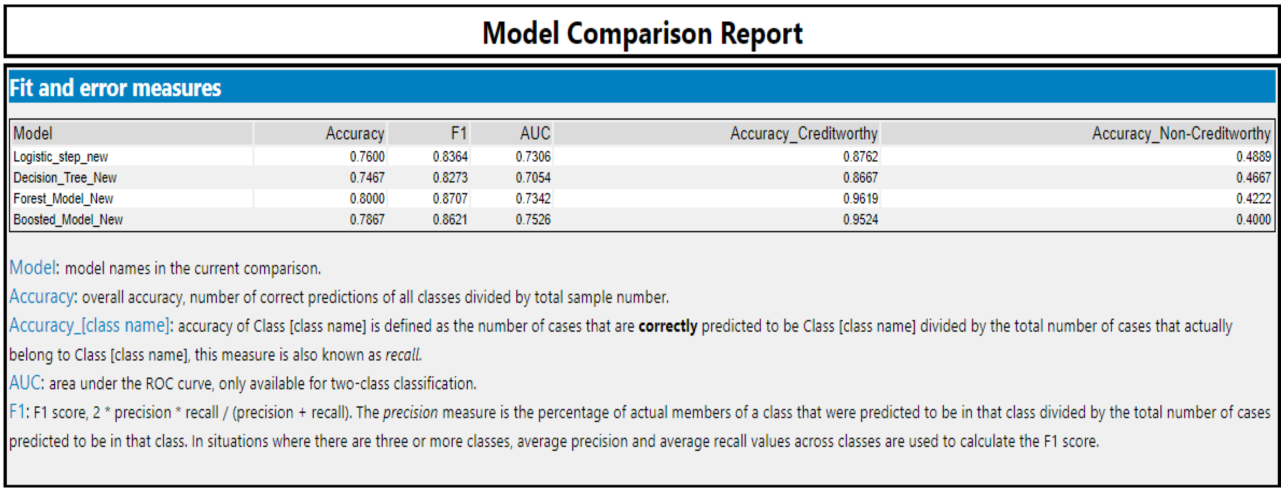
Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score-Creditworthy is greater than Score-Noncreditworthy, the person should be labelled as "Creditworthy"

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:
  A) Overall Accuracy against your Validation set
  B) Accuracies within "Creditworthy" and "Non-Creditworthy" segments
  C) ROC graph
  D) Bias in the Confusion Matrices

Ans - Forest Model has been chosen since it has the highest overall accuracy of 80%. It also has a F1 accuracy of 87% which is highest of all. The accuracies of credit worthy and non-credit worthy segments has 96% and 42 % respectively with less bias towards any decisions. This is really crucial in making the most predicted outcome by offering loan to the best suited credit worthy customers.

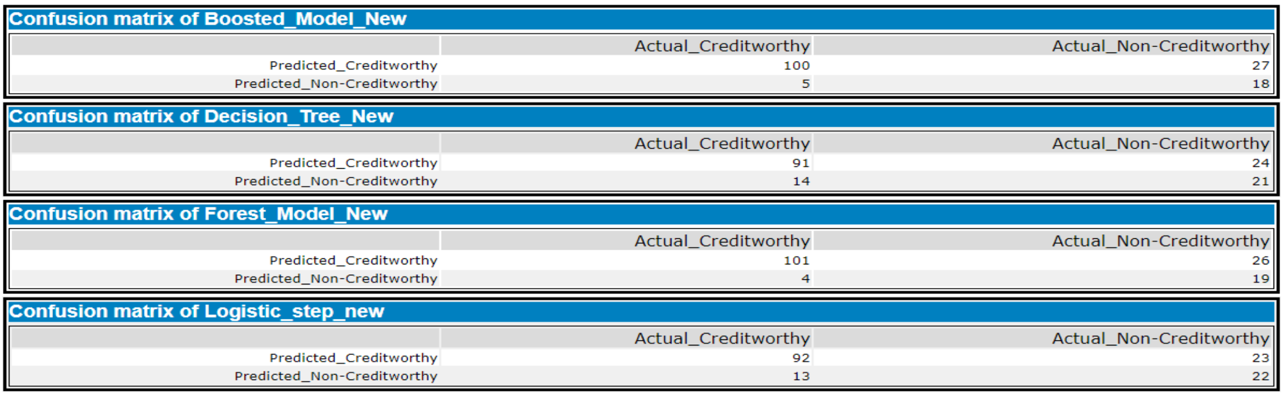


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_step_new	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_New	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model_New	0.8000	0.8707	0.7342	0.9619	0.4222
Boosted_Model_New	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_New

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

Confusion matrix of Decision_Tree_New

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

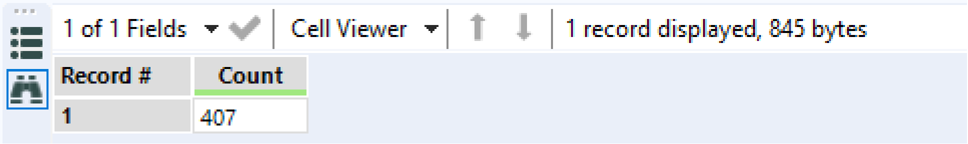
Confusion matrix of Forest_Model_New

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Logistic_step_new

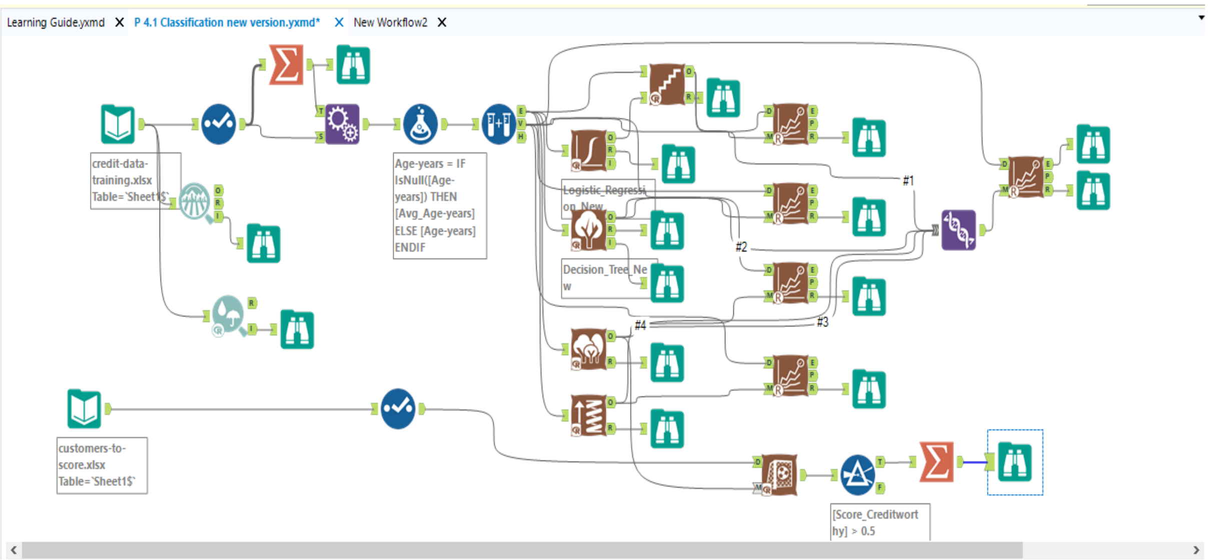
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

From the modelling it was found that the final count of individuals who are creditworthy are 407.



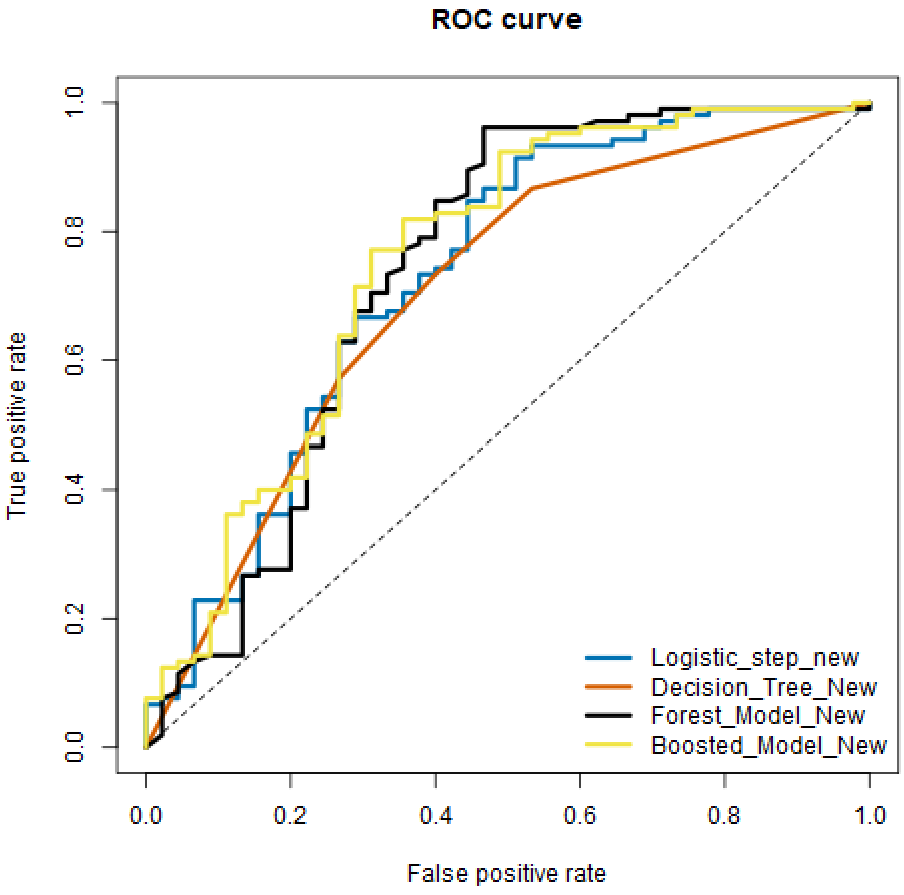
Record #	Count
1	407

<u>**ALTERYX WORKFLOW**</u>


Learning Guide.yxmd
P 4.1 Classification new version.yxmd*
New Workflow2
credit-data-training.xlsx Table=`Sheet1$`
Logistic_Regression_New
Decision_Tree_New
#1
#2
#3
#4
customers-to-score.xlsx Table=`Sheet1$`

```
Age-years = IF
IsNull([Age-
years]) THEN
[Avg_Age-years]
ELSE [Age-years]
ENDIF
```

```
[Score_Creditwort
hy] > 0.5
```


ROC curve
True positive rate
False positive rate
Logistic_step_new
Decision_Tree_New
Forest_Model_New
Boosted_Model_New