Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Answer:

As an analyst, I have the responsibility to predict whether a customer is approved for a loan or not.

2. What data is needed to inform those decisions?

Answer:

We need information about the credit approvals from the past loan applicants and the list of customers that need to be processed in the next few days for loan approval.

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

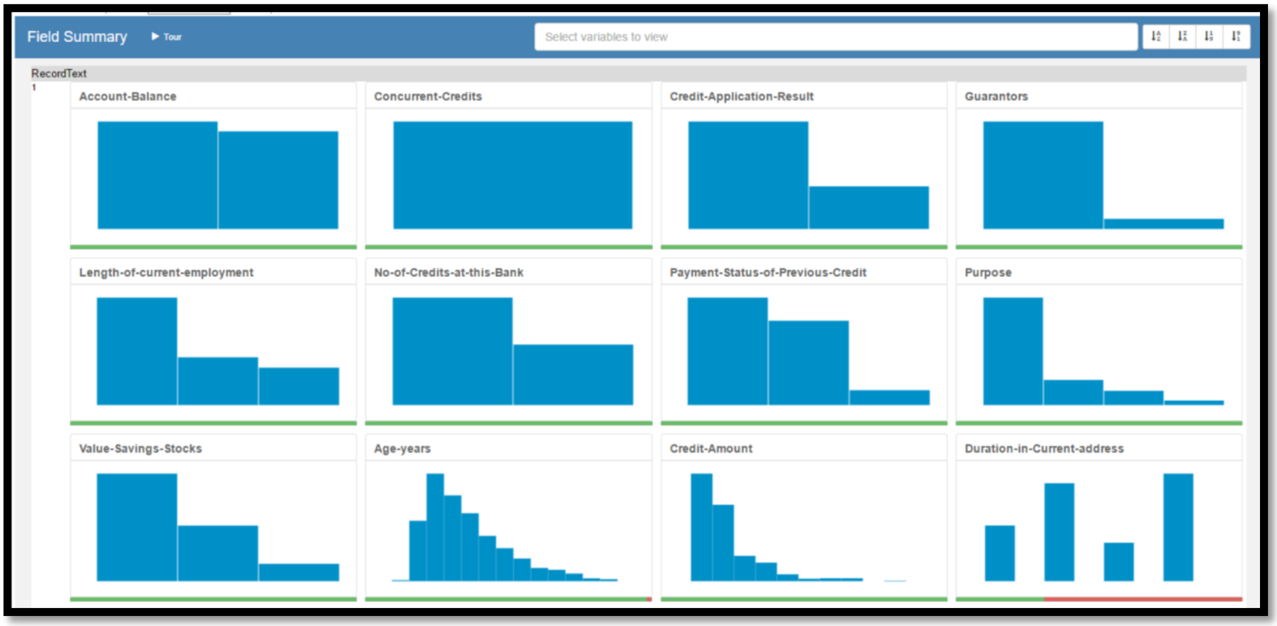
Answer:

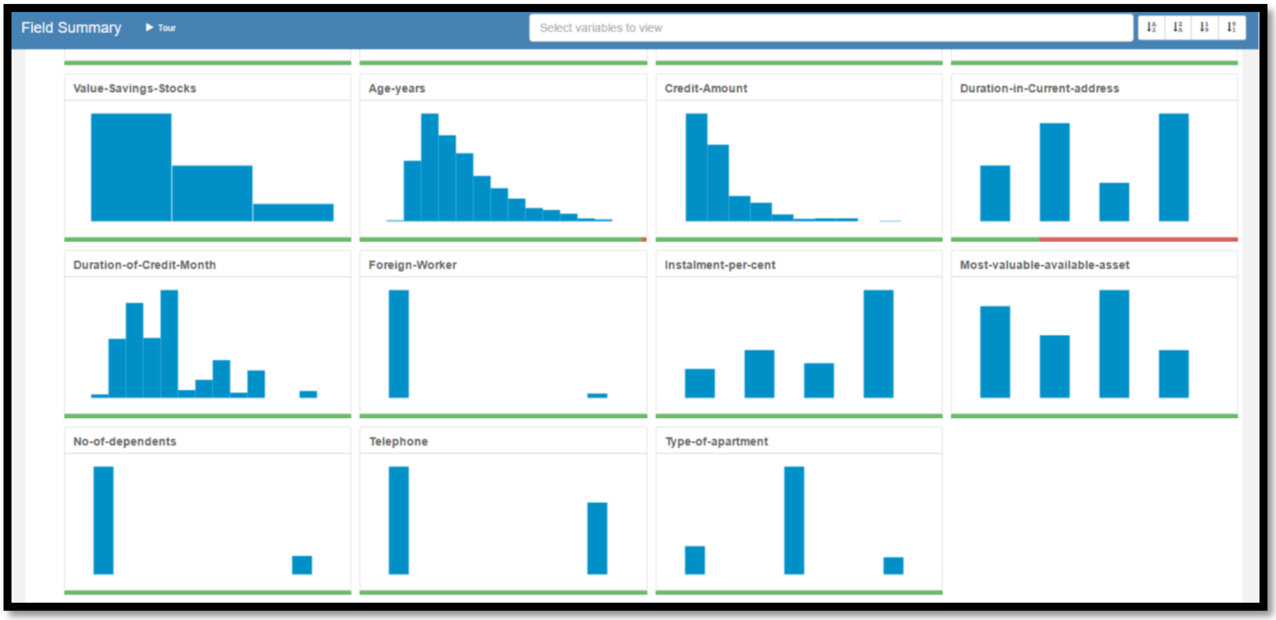
We need a **Binary model** to help make our decision as the predicted variable only has two categories: Credit worthy and Non-credit worthy.

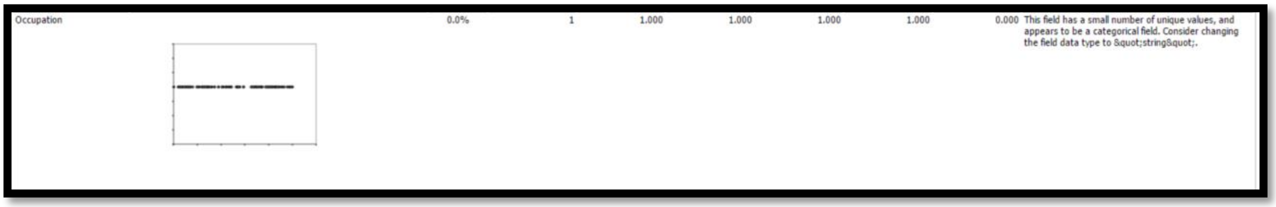
Step 2: Building the Training Set

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

Answer:







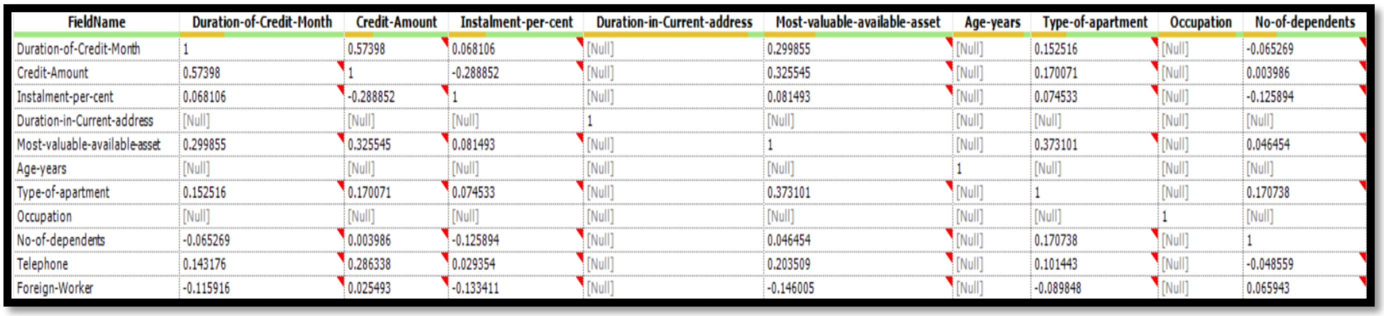
```
Occupation                              0.0%     1     1.000     1.000     1.000     1.000     0.000  This field has a small number of unique values, and
                                                                                                        appears to be a categorical field. Consider changing
                                                                                                        the field data type to &quot;string&quot;.
```

All the fields in which didn't show variability or had a lot of null values were de-selected. The fields that were de-selected using the 'Select' tool are:

a) Guarantors (low variability)
b) Duration-in-Current-Address (null values)
c) Concurrent-Credits (no variability)
d) Occupation (low variability)
e) No-of-dependents (low variability and biased)
f) Telephone (low variability)
g) Foreign-Worker (low variability and biased)

All the null values in Age-years were imputed by its median value (i.e. 33) so that a normal distribution is achieved. Hence, the data isn't biased.

Also, using 'Pearson Coefficient' tool to analyze correlation between variables, it was evident that there was no correlation between the variables since pearson co-efficient didn't have any value greater than 0.7.



FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents
Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269
Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986
Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894
Duration-in-Current-address	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454
Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738
Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]
No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1
Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559
Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943

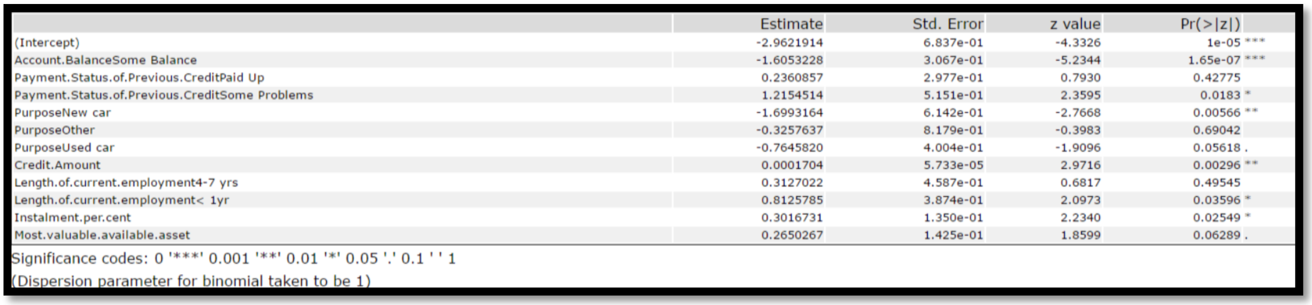
Step 3: Train your Classification Models

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

<u>Answer</u>:

Logistic Regression –

- Account Balance
- Credit Amount
- Purpose



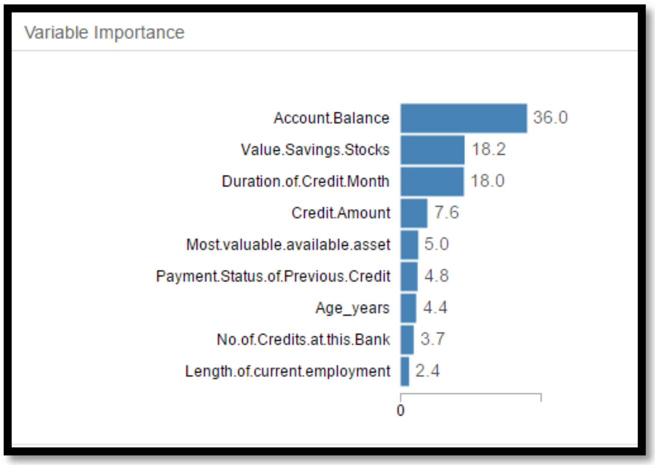
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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)

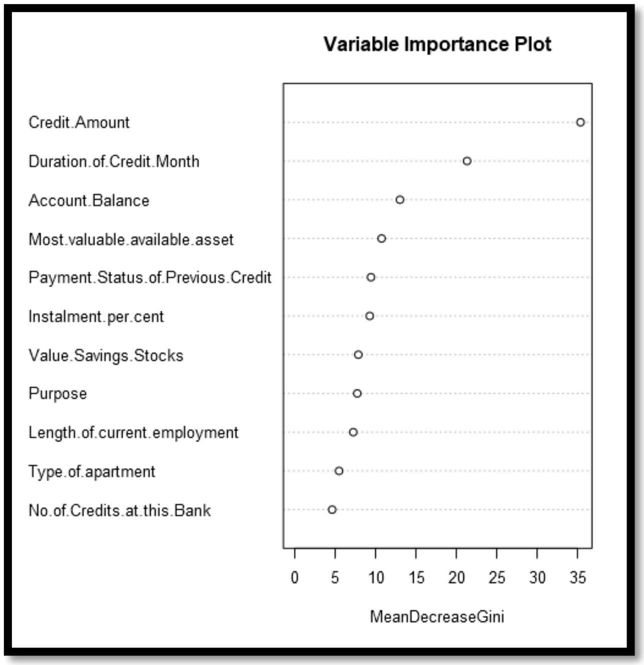
Decision Tree –

- Account.Balance
- Value.Savings.Stocks
- Duration.of.Credit.Month
- Credit.Amount



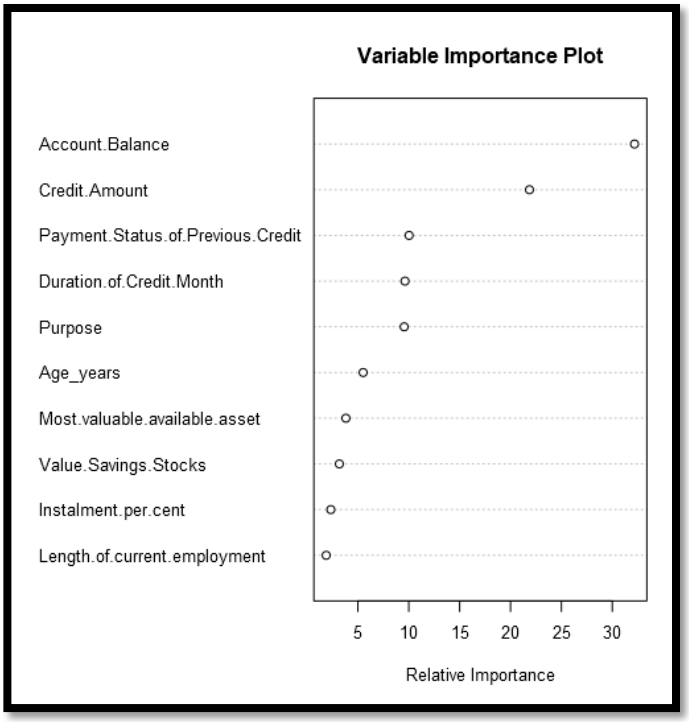
Forest Model –

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



Boosted Model -

- Account.Balance
- Credit.Amount
- Payment Status of Previous Credit

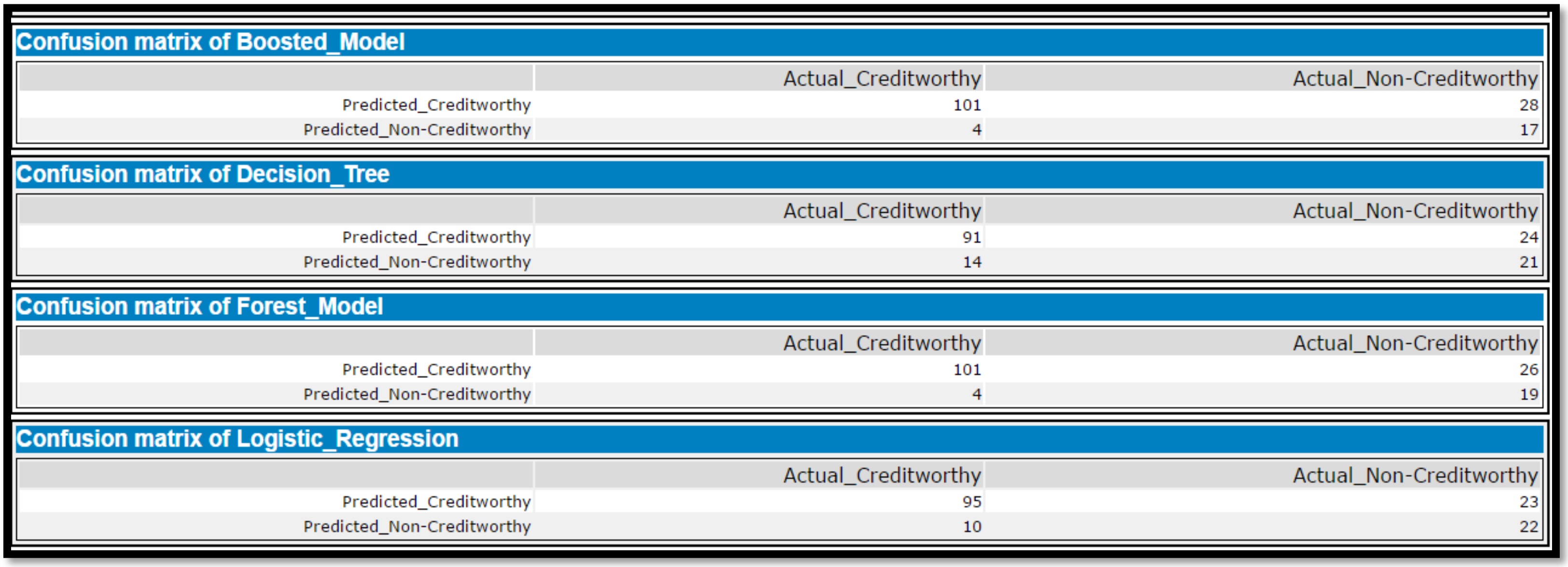


2.

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

<u>**Answer**</u>**:**

Here are the confusion matrices of all the models –



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	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Forest_Model

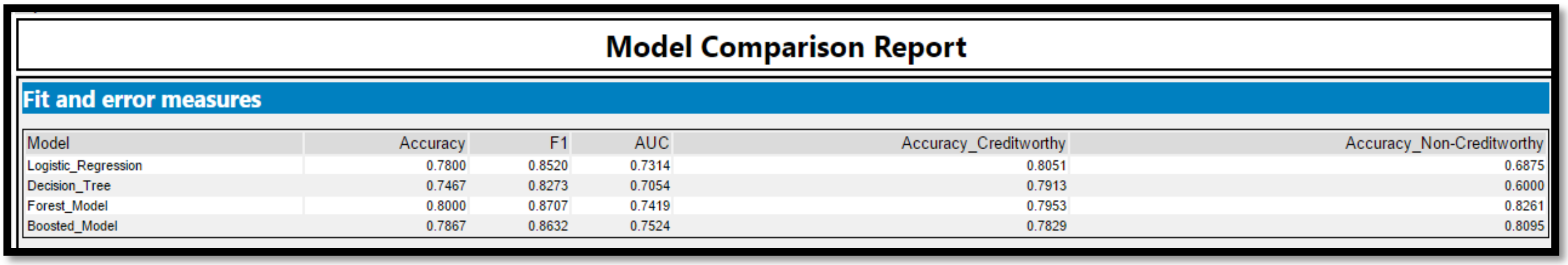
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Logistic_Regression

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

There is a bias towards 'Creditworthy' as the accuracy of Creditworthy for all the models is much higher than the predicted accuracies of non-creditworthy.
Therefore, we can conclude a bias towards Creditworthy.

And given below are the overall accuracies:



Model Comparison Report

Fit and error measures

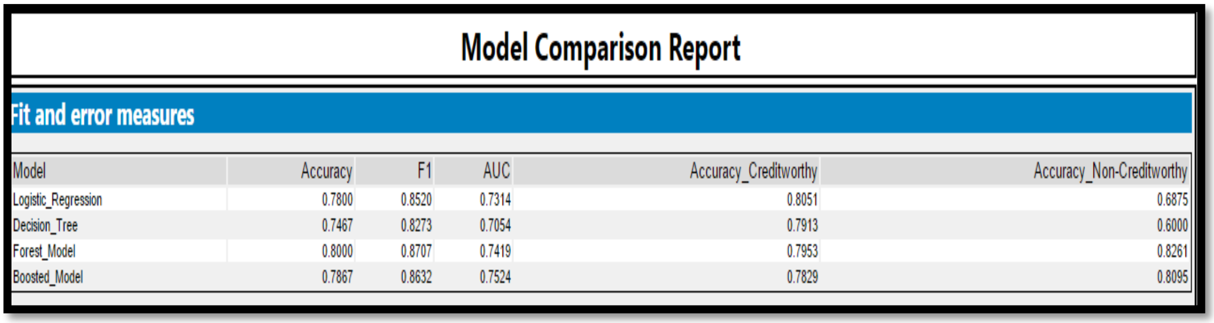
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

All the models have a very close set of overall accuracy ranging from about 74-80% with Decision Tree having the least accuracy and Forest Model having the highest accuracy.

Step 4: Writeup

1. 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

Answer:
From the model comparison report, it is evident that Forest Model provides the best and the most accurate results.



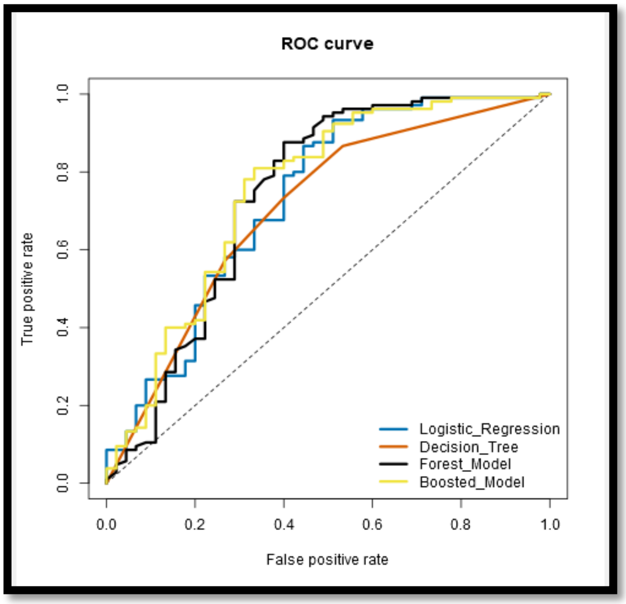
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression	0.7800	0.8520	0.7314	0.8051	0.6875
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Looking at the ROC curve and the Gain chart, it can be seen that Forest model reaches the top the quickest and the highest.

Bias in the confusion matrix was highest in the case of Forest Model with Creditworthy accuracy at about 80% and 82% Non-creditworthy accuracy.

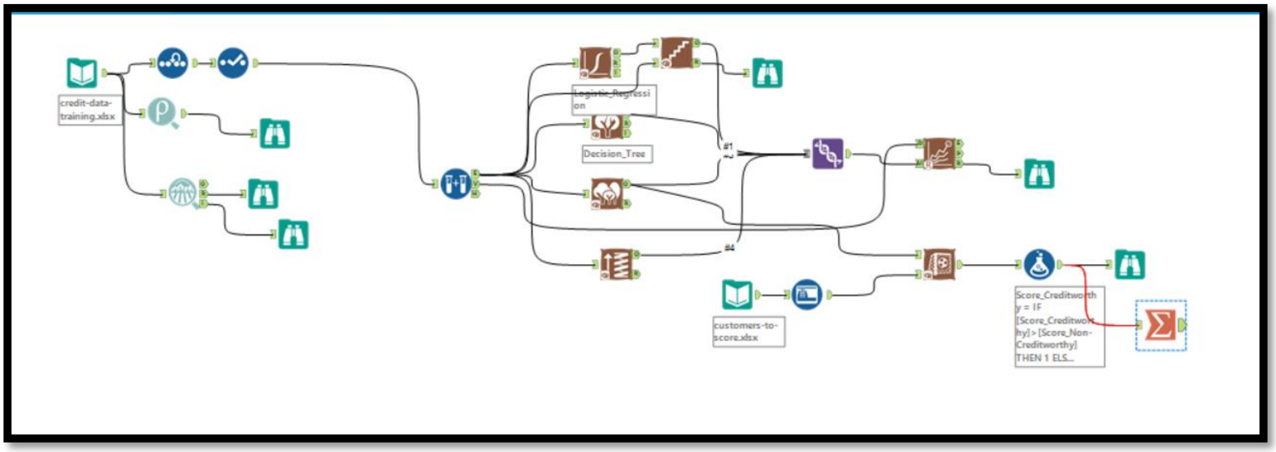


Therefore, from these 4 parameters, **Forest Model** turns to be the best for our predictive analysis.

1. How many individuals are creditworthy?

<u>Answer</u>:

After scoring the Forest model with the data of new customers, **408** out of the 500 customers turned out to be creditworthy for loan.


credit-data-training.xlsx
Logistic_Regression
Decision_Tree
#1
#3
#4
customers-to-score.xlsx
Score_Creditworthy = IF [Score_Creditworthy] > [Score_Non-Creditworthy] THEN 1 ELS...