Project: Creditworthiness

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?

Decide whether or not to give loans for a list of customers.

- What data is needed to inform those decisions?

Data on past applications and the result of the credit analysis, as well as the list of customers to be processed with the same data available (for example Account Balance and Credit Amount).

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

This kind of problem asks for binary classification models such as logistic regression, decision tree, forest model and boosted.

Step 2: Building the Training Set

Here are some guidelines to help guide your data cleanup:

- 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)

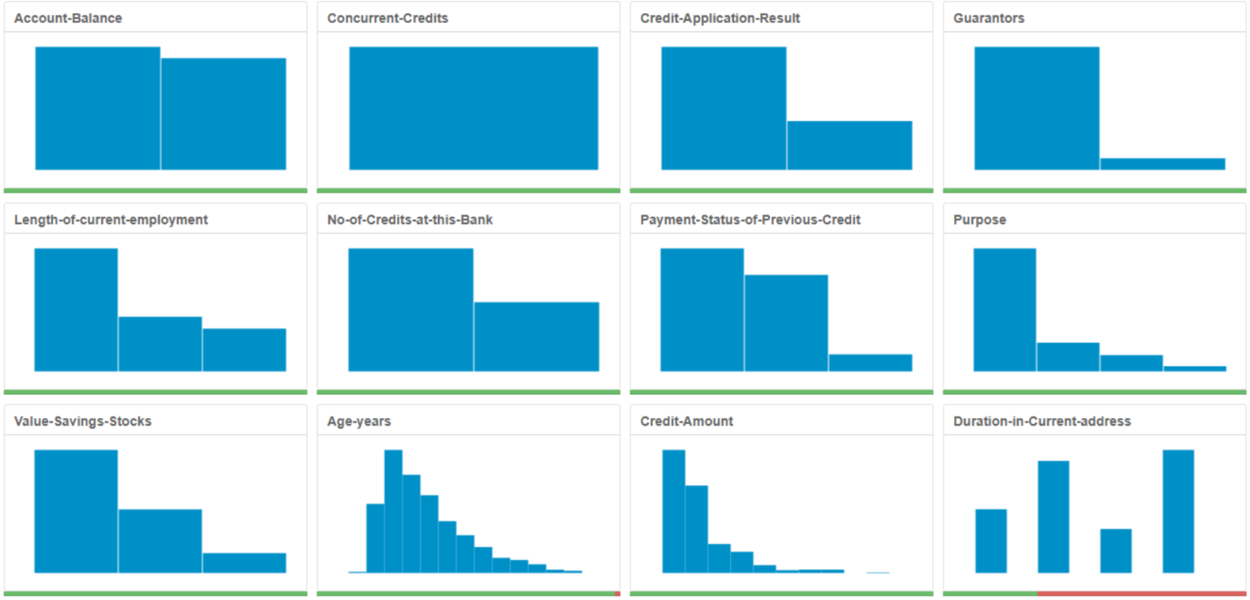
Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

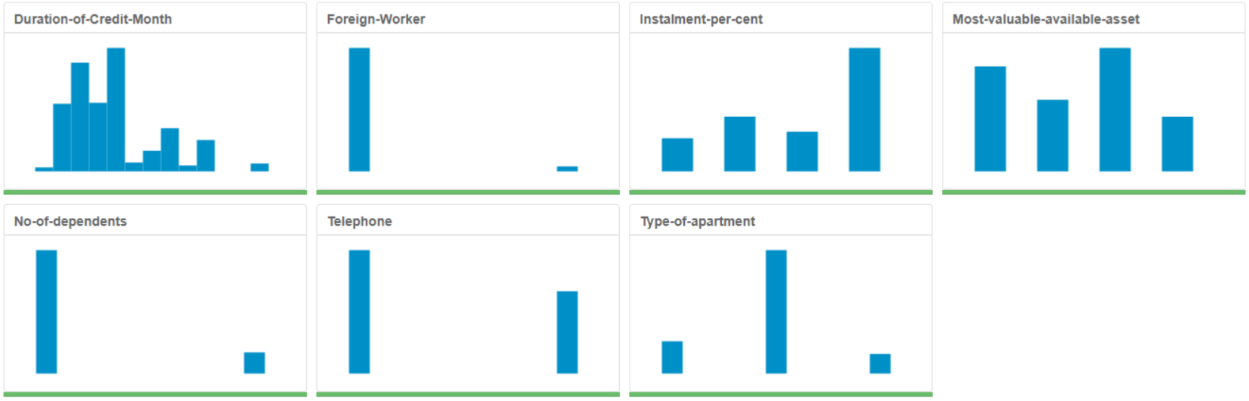
Note: *For students using software other than Alteryx, please format each variable as:*

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.

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



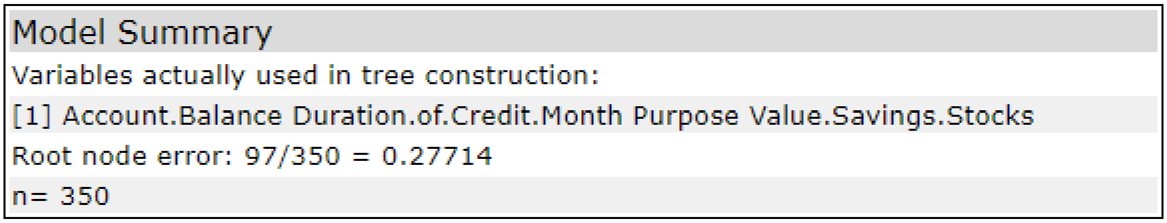


I removed the Duration-in-Current-address field since it has 69% of null values. Age years has 2% missing data, but this is not that much and it could be useful to the model, so I decided to impute missing values with the median age.
Concurrent Credits, Occupation, Guarantors, Foreign Worker and No of Dependents show low variability with either one single value or more than 80% of the data skewed towards one option. Telephone was also removed since the correlation analysis showed that this field is irrelevant to assess creditworthiness, with a p-value of 0.71.

Step 3: Train your Classification Models

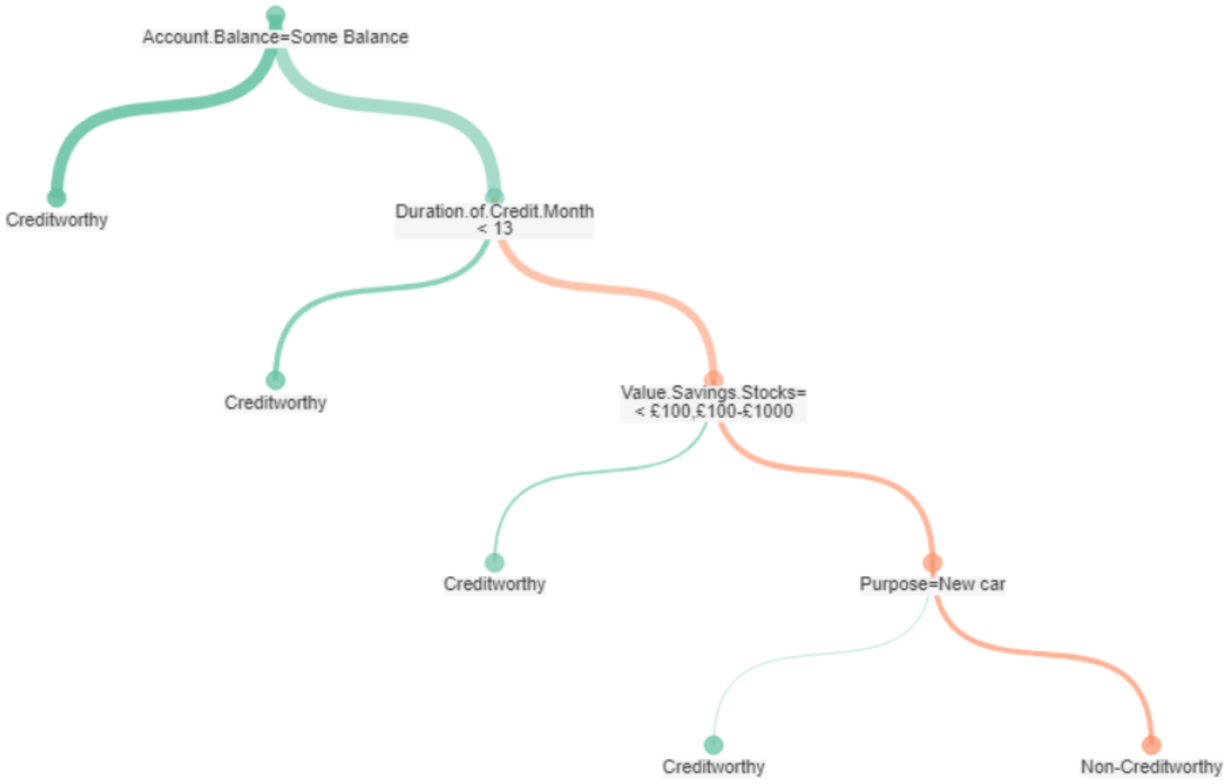
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

Decision Tree:
The three most important variables for Decision Tree were Account Balance (Some Balance), Duration of Credit Month and Value Savings Stocks.

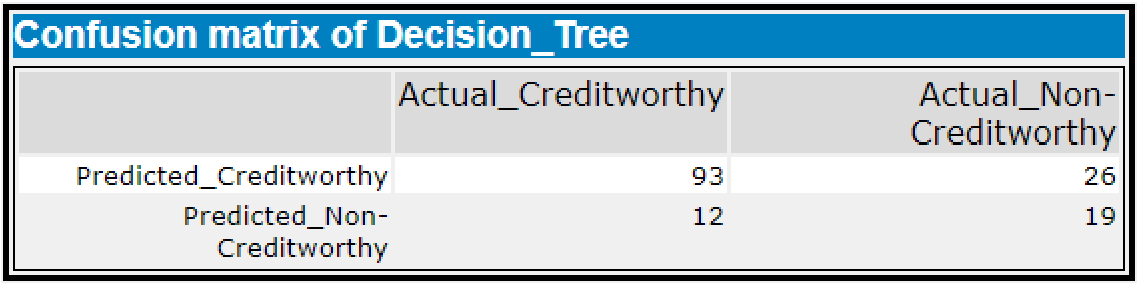


Model Summary

```
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
```



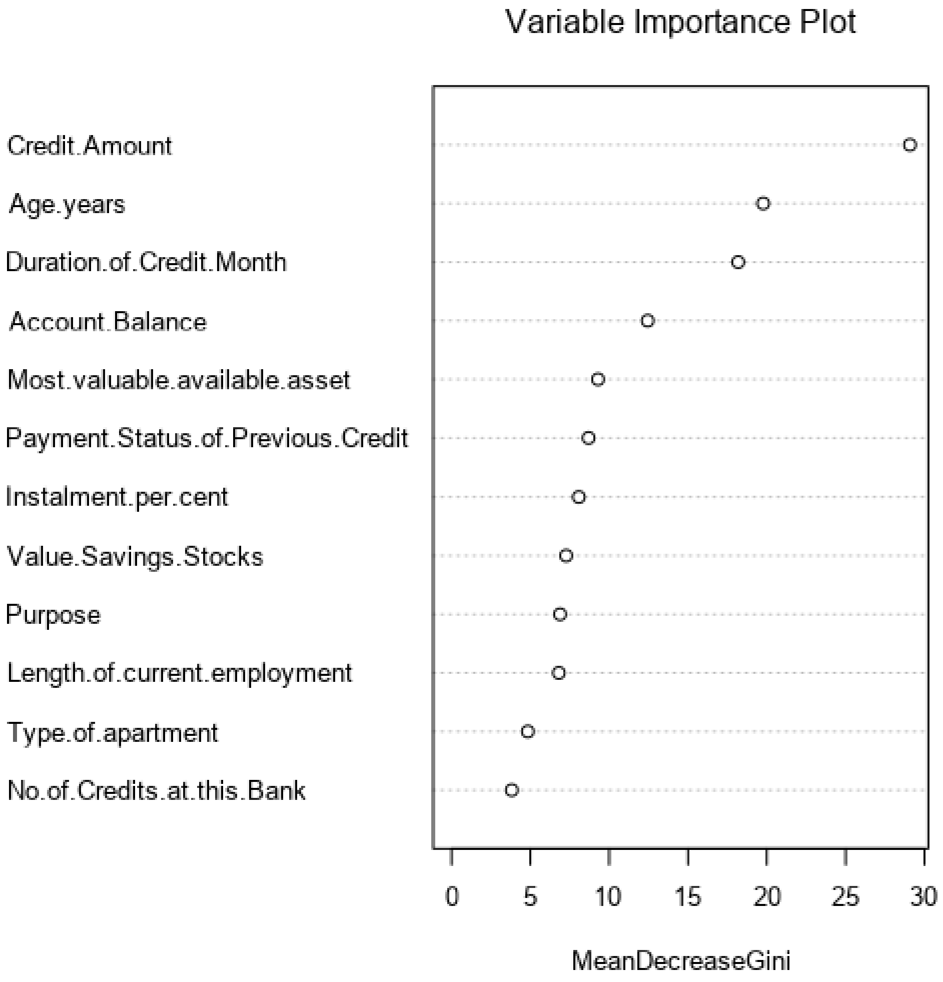
Validating the model, we get an overall accuracy of 70.35%, and the model shows a bias towards scoring customers as creditworthy.



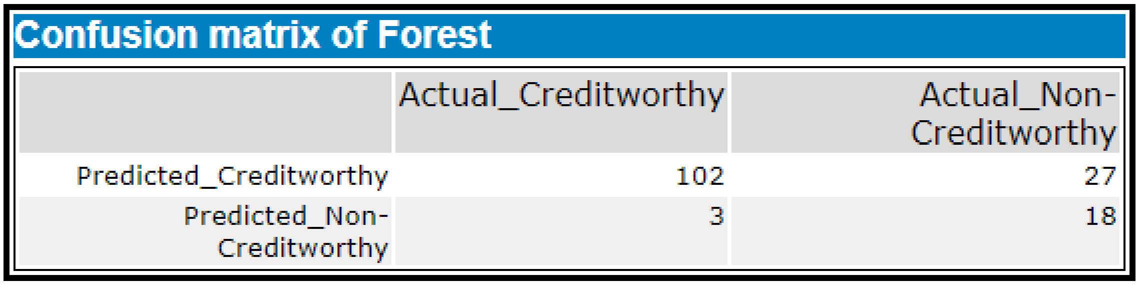
Confusion matrix of Decision_Tree

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

Forest:
The three most important variables for the Forest Model were Credit Amount, Age Years and Duration of Credit Month.



Validating the model, we get an overall accuracy of 80%, and the model shows a bias towards scoring customers as creditworthy.

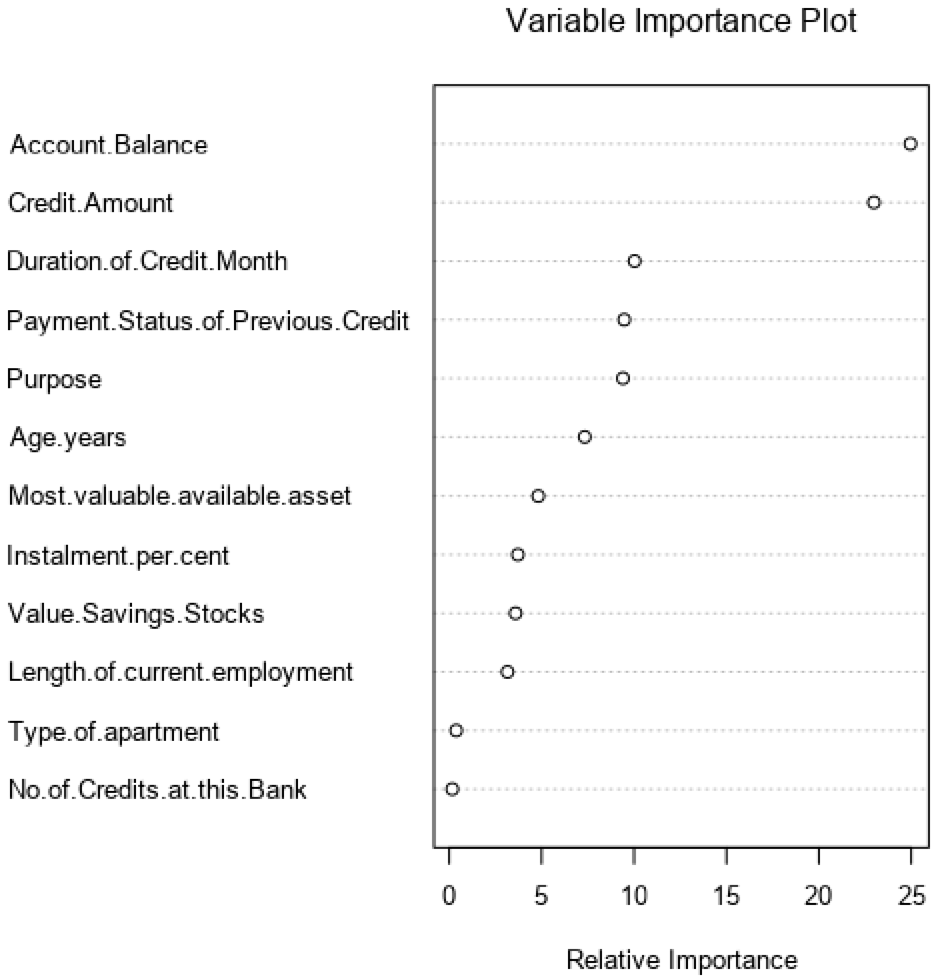


Confusion matrix of Forest

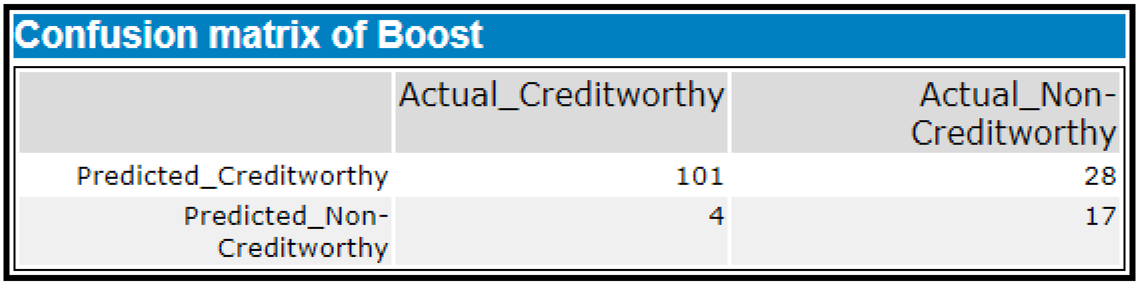
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Boost:

The three most important variables for the Boost Model were Account Balance, Credit Amount and Duration of Credit Month.



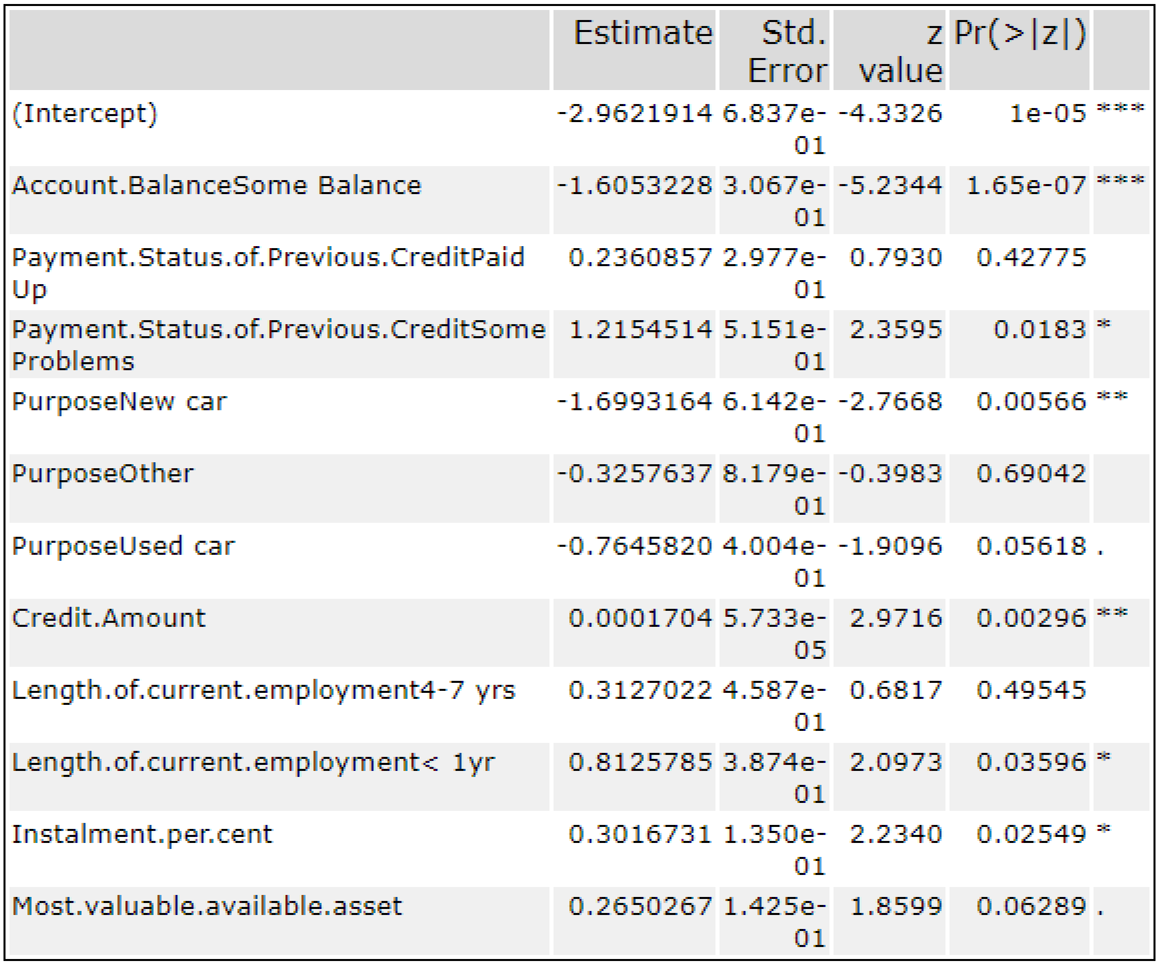
Validating the model, we get an overall accuracy of 78.67%, and the model shows a bias towards scoring customers as creditworthy.



Confusion matrix of Boost

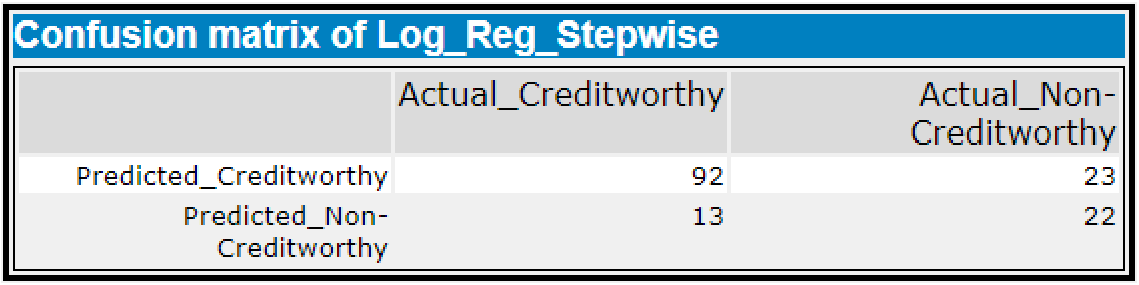
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Logistic Regression Stepwise:
The three most important variables for the Logistic Regression Stepwise model were Account Balance (Some Balance), Credit Amount and Purpose (New Car).



	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	.

Validating the model, we get an overall accuracy of 76%, and the model shows a bias towards scoring customers as creditworthy.



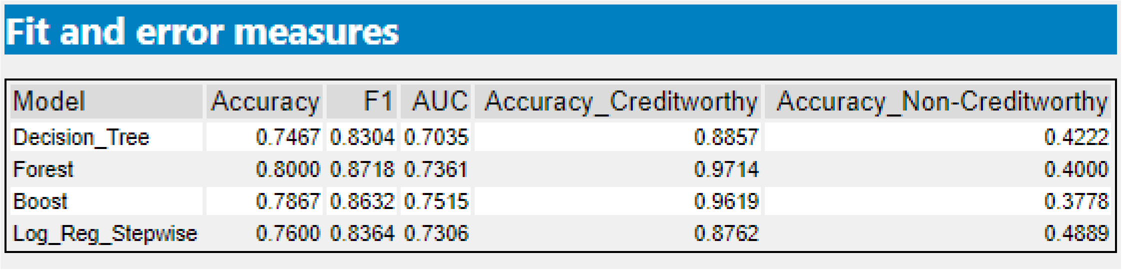
Confusion matrix of Log_Reg_Stepwise

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

- 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

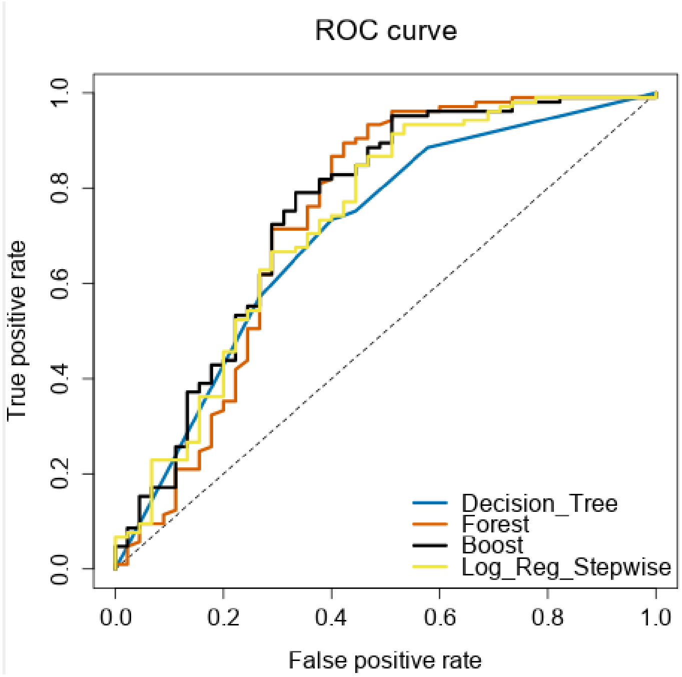
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest	0.8000	0.8718	0.7361	0.9714	0.4000
Boost	0.7867	0.8632	0.7515	0.9619	0.3778
Log_Reg_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889



Forest model has the highest accuracy at 80% and reaches the true positive rate at the fastest rate, while all models have similar behaviours in relation to biasing towards creditworthiness. Choosing the Forest model gives us the best option in terms of giving loans to the right customers (thus not losing money) while having a moderate risk of giving it to non-creditworthy customers.

There are 409 creditworthy customers using forest models to score new customers.