Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

We need to decide which of the 500 loan applications we will approve

- What data is needed to inform those decisions?

The data we need is the historic data on past applications and the data of the customers who applied for a loan

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

We need a binary model as we only have two categories: Creditworthy and Non-Creditworthy

Step 2: Building the Training Set

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

In my clean-up I imputed the Age-Years fields as the missing values were not too many (2.4%) and I completely removed Duration-in-Current-address field as 68.8% were missing values.

Additionally, I removed the fields Occupation and Concurrent-Credits due to low variability. More specifically, the data had just one value making them uniform.

Similarly, due to low variability I removed the fields Guarantors, Foreign-Worker and No-of-dependents. More specifically all of the fields had 2 unique values but the data were heavily skewed towards one.

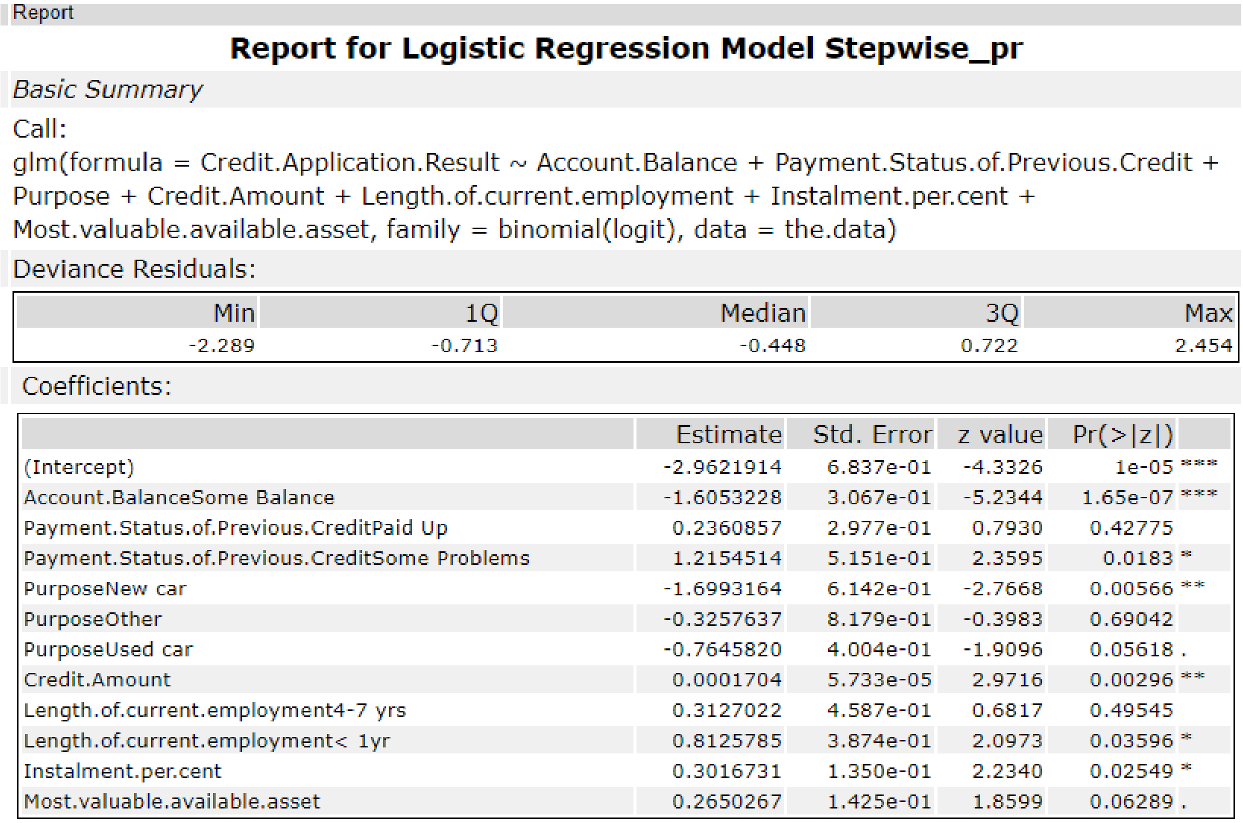
Lastly I removed the field Telephone as it had only two unique values and would not add much information to the model

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/ Stepwise:

In my Log. Regression Stepwise model the significant variables were Account.Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount and Length-of-current-employment as shown below:



Report

Report for Logistic Regression Model Stepwise_pr

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:

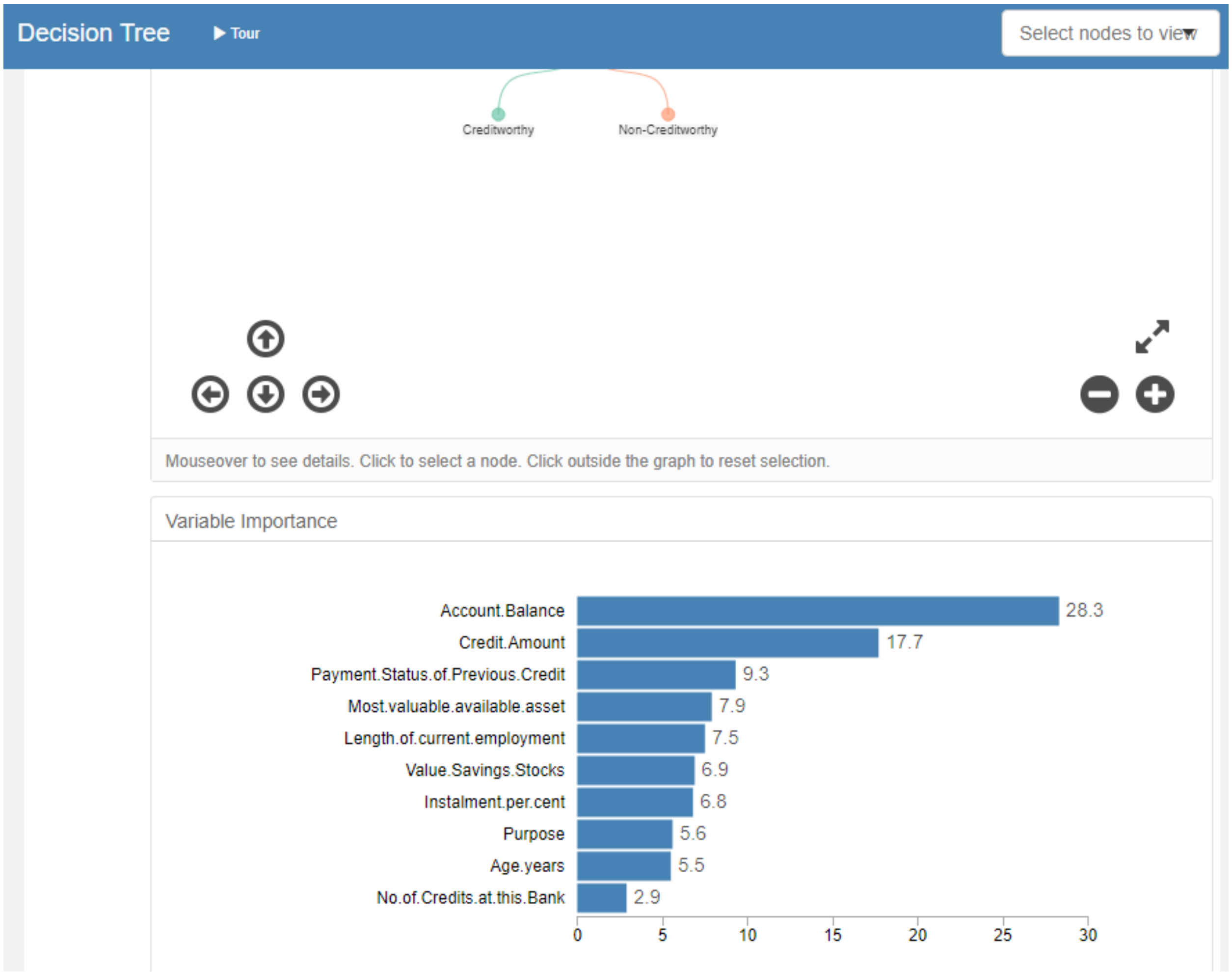
Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

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	.

Decision Tree:

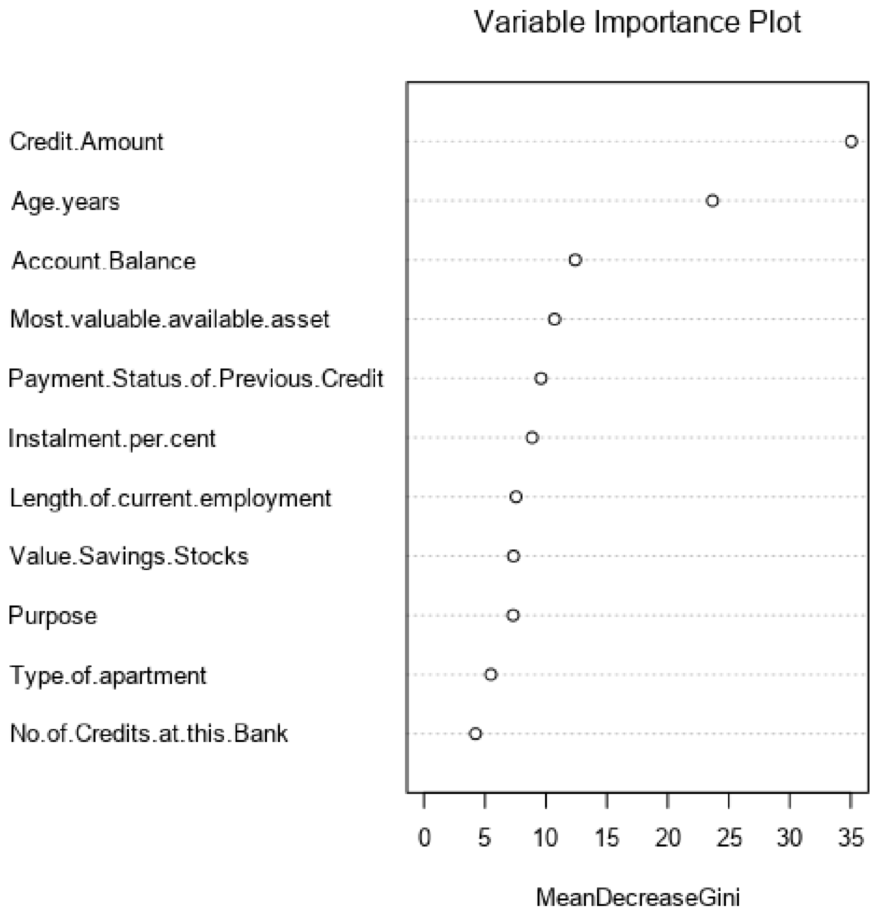
In this model as shown below the most important variables are Account.Balance and Credit.Amount.



Mouseover to see details. Click to select a node. Click outside the graph to reset selection.

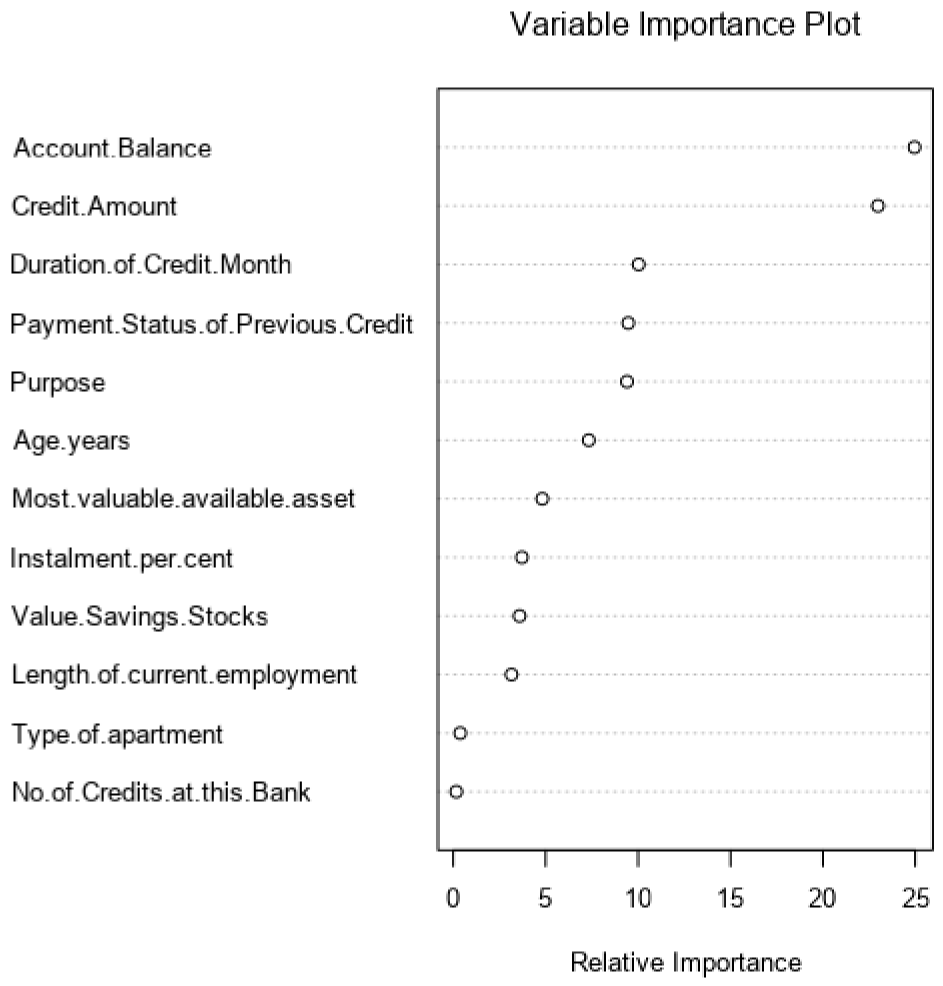
Forest Model:

In the Forest Model as shown below, the most important variables are Credit.Amount and Age.Years.



Boosted Model:

In the Boosted Model the most important variables are Account.Balance and Credit.Amount as shown below:



- 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?

Bias & Accuracy:

Logistic Regression/ Stepwise:

The overall percent accuracy of the Logistic model is 76% which is strong.
PPV= true positives \ (true positives + false positives) = 92 / (92+23) =.80
NPV= true negatives\ (true negatives + false negatives) =22/ (22+13) = .63
So, after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

Decision Tree:

The overall percent accuracy of the Decision Tree model is 73% which is strong.
PPV= true positives \ (true positives + false positives) = 87 / (87+22) =.80
NPV= true negatives\ (true negatives + false negatives) =23/ (23+18) = .56
So, after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

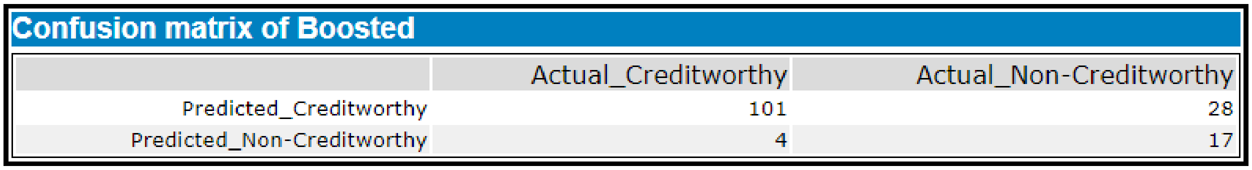
Boosted Model:

The overall percent accuracy of the Boosted model is 79% which is strong.
PPV= true positives \ (true positives + false positives) = 101 / (101+28) =.78
NPV= true negatives\ (true negatives + false negatives) =17/ (17+3) = .57
So, after checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.

Forest Model:

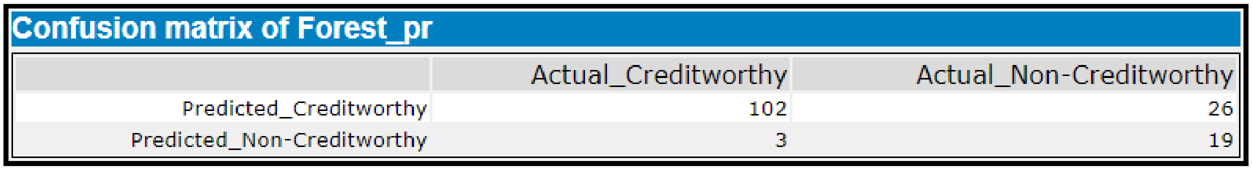
The accuracy of the Forest model is 80% which is strong
PPV= true positives \ (true positives + false positives) = 102/ (102+26) =.80
NPV= true negatives \ (true negatives + false negatives) = 19/ (19+3) = .86
So, after checking the confusion matrix there is no bias seen in the model's prediction.

The confusion matrices for the four models:



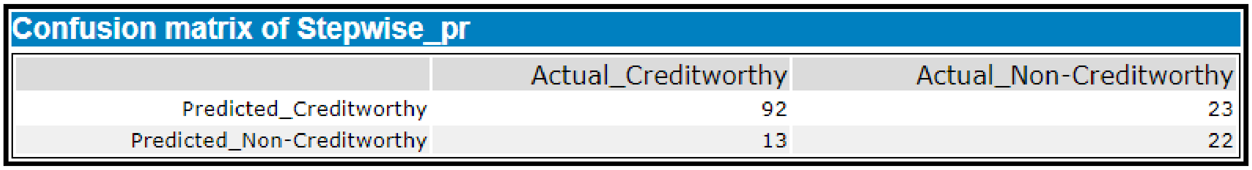
Confusion matrix of Boosted

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



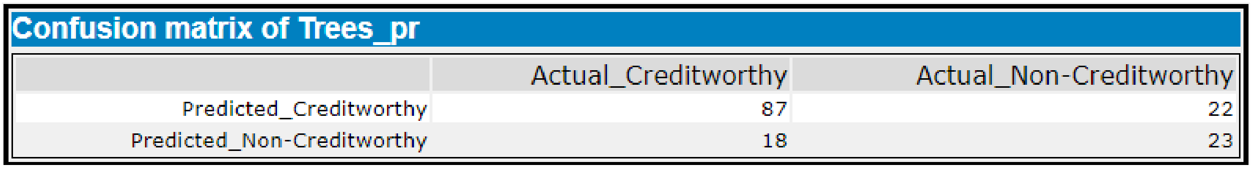
Confusion matrix of Forest_pr

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



Confusion matrix of Stepwise_pr

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



Confusion matrix of Trees_pr

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	87	22
Predicted_Non-Creditworthy	18	23

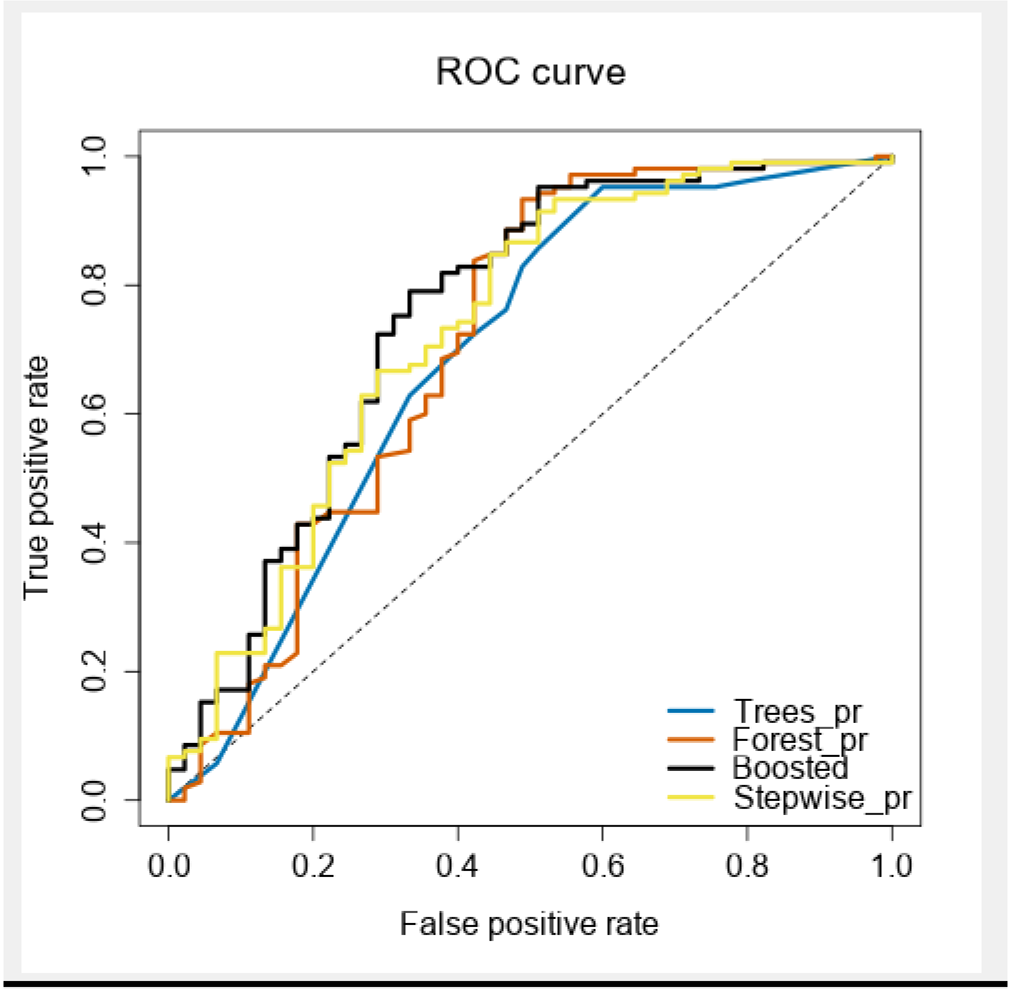
Step 4: Writeup

- 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 used the Forest Model as it had the highest overall accuracy against the validation set, the highest F1 (0.8755) score and the highest accuracy within the “Creditworthy” segment (0.9714). Overall, the above values show the model with the most predictive power.

The ROC curve also shows that, as the Forest model reaches the top the quickest compared to the other models.

Forest model also has the highest AUC number with 0.74 which is a good indicator combined with all the rest.



Lastly, as explained in the previous section, the Forest model is the only model out of the four that does not show bias in its predictions.

The numbers again for reference:

Logistic Regression/ Stepwise:

PPV=.80

NPV= .63

Decision Tree:

PPV =.80

NPV = .56

Boosted Model:

PPV =.78

NPV = .57

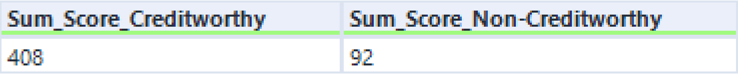
Forest Model:

PPV =.80

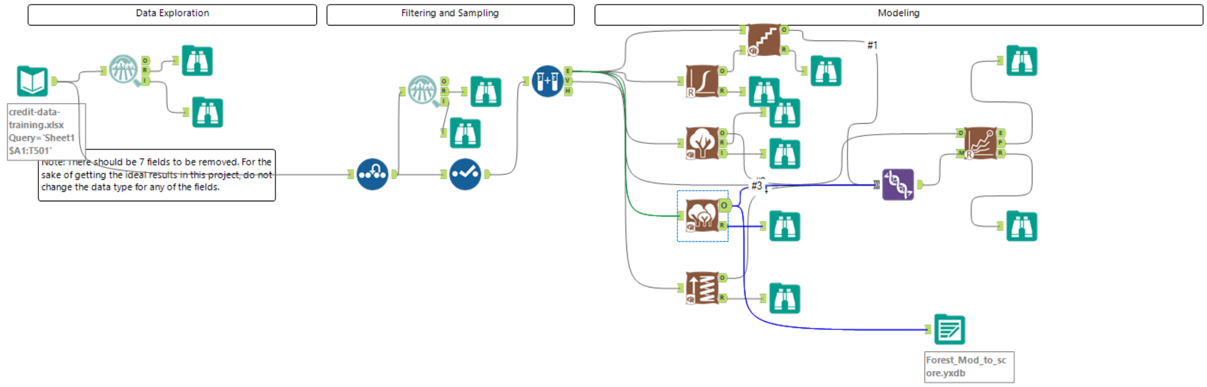
NPV = .86

- How many individuals are creditworthy?

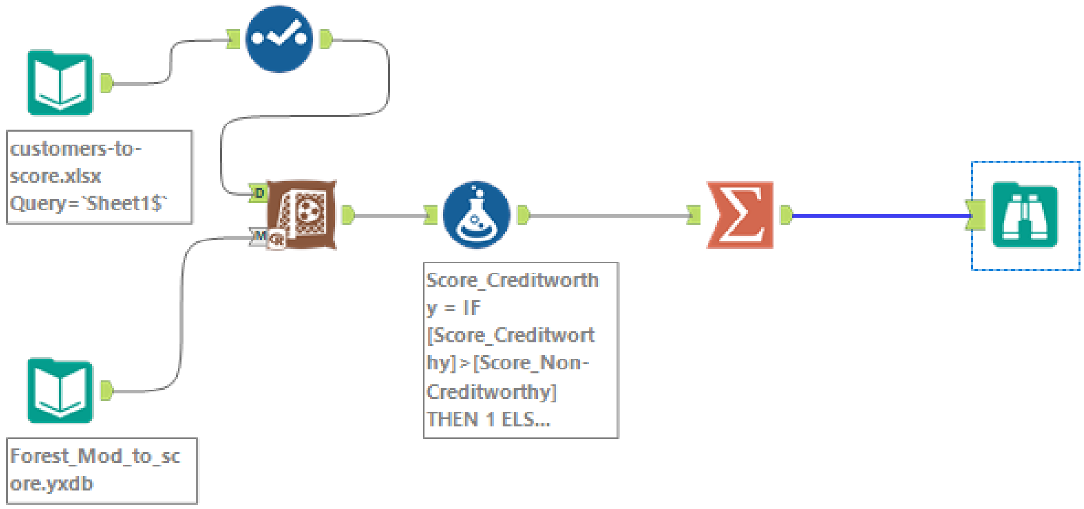
408



Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
408	92



Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.



Score_Creditworthy = IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN 1 ELS...