Project: Creditworthiness

Step 1: Business and Data Understanding

What decisions needs to be made?

A small bank has the responsibility of determining if a customer is creditworthy to give loan to. Usually the bank receives about 200 loan applications per week which are approved by hand, however due to a financial scandal that hit competitive bank, they have now received nearly 500 loan applications. My manager wants me to figure out how to process all these applications within a week. Based on the courses I have taken in Classification modelling; I would need to systematically evaluate the creditworthiness of all the new applicants and provide my manager with a list of creditworthy customers in the next 2 days.

What data is needed to inform those decisions?

- The data of all past applicants
- The list of new customers who needs to be processed in the next few days

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

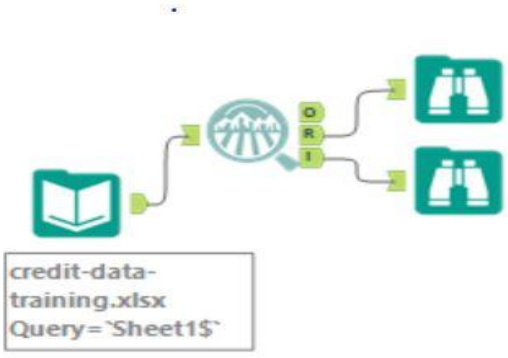
The kind of problem we need to build to solve this problem is a binary model, because we want to identify whether a customer is creditworthy or not creditworthy.

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?
- Are there only a few values in a subset of your data field? Does the data field look very uniform? This is called “low variability” and you should remove fields that have low variability.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

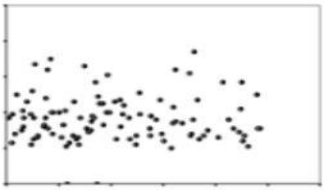
Here is the workflow that I used to Analyze the data:



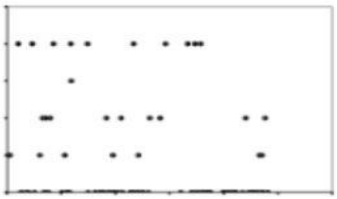
Using the above workflow, I have identified the missing data and fields with low variability. Below are the details from the report output and the interactive output.

Missing Data:

Age(years) – This field as we can see below has 2.4% missing data. Since the no. of missing data is not too much, I will impute the data using the median of the entire age field, because all the data for age is shifted towards the left.

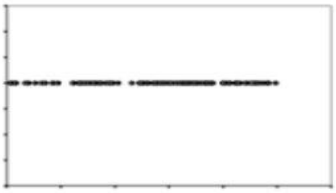
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502

Duration-in-Current—address- This field has 68.8% missing value, and thus it should be removed.

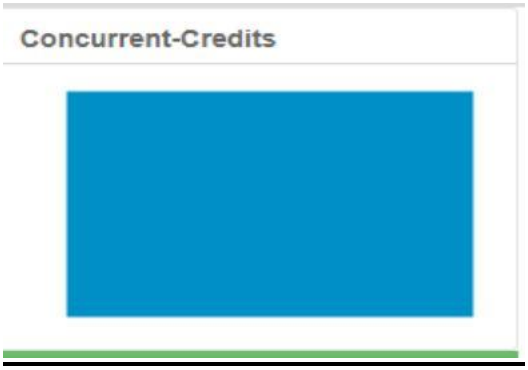
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150

Low Variability: the data is very uniform and there is no variations of the data

Occupation- This field only has 1 value.

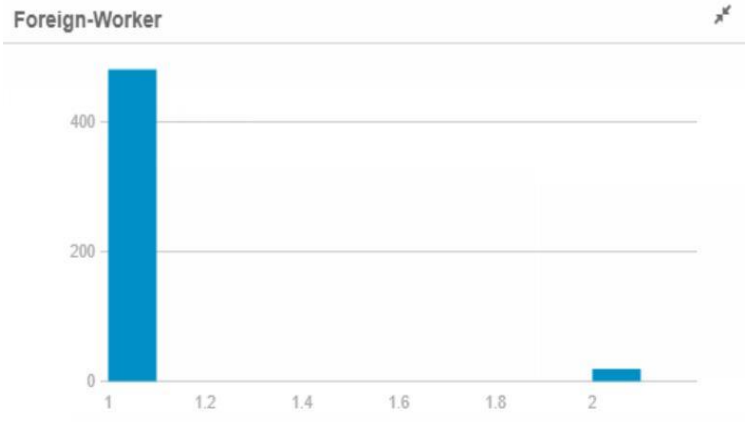
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000

Concurrent-Credits- This field also has only 1 value i.e., “Other Banks/Debts”(500 instances)

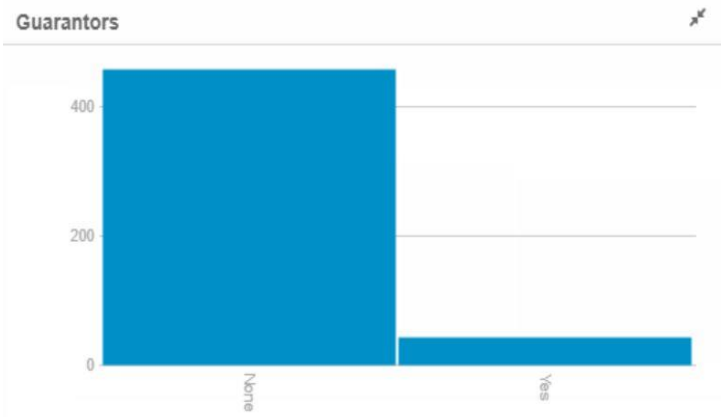


Low Variability: the data is heavily skewed towards one type of data

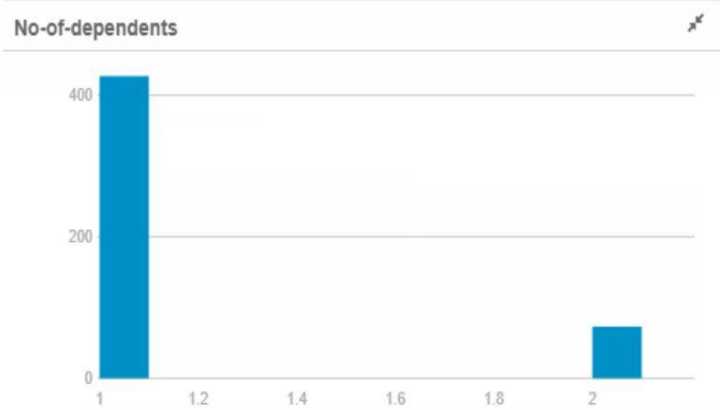
Foreign-Worker- The histogram below shows that most of the data is skewed towards 1.



Guarantors- This variable has most of its data skewed towards “None”



No. of dependents- The variable shows that most of the data is skewed towards 1.



Telephone- This should also be removed as it cannot tell us anything about the creditworthiness of an applicant.

Therefore, to summarize:

- **Impute** data with the median value of the entire set- **Age- years**
- **Remove variables** of No. of dependents, Duration-in-Current-address, Occupation, Concurrent-Credits, Foreign-Worker, Guarantors and Telephone

Step 3: Train your Classification Models

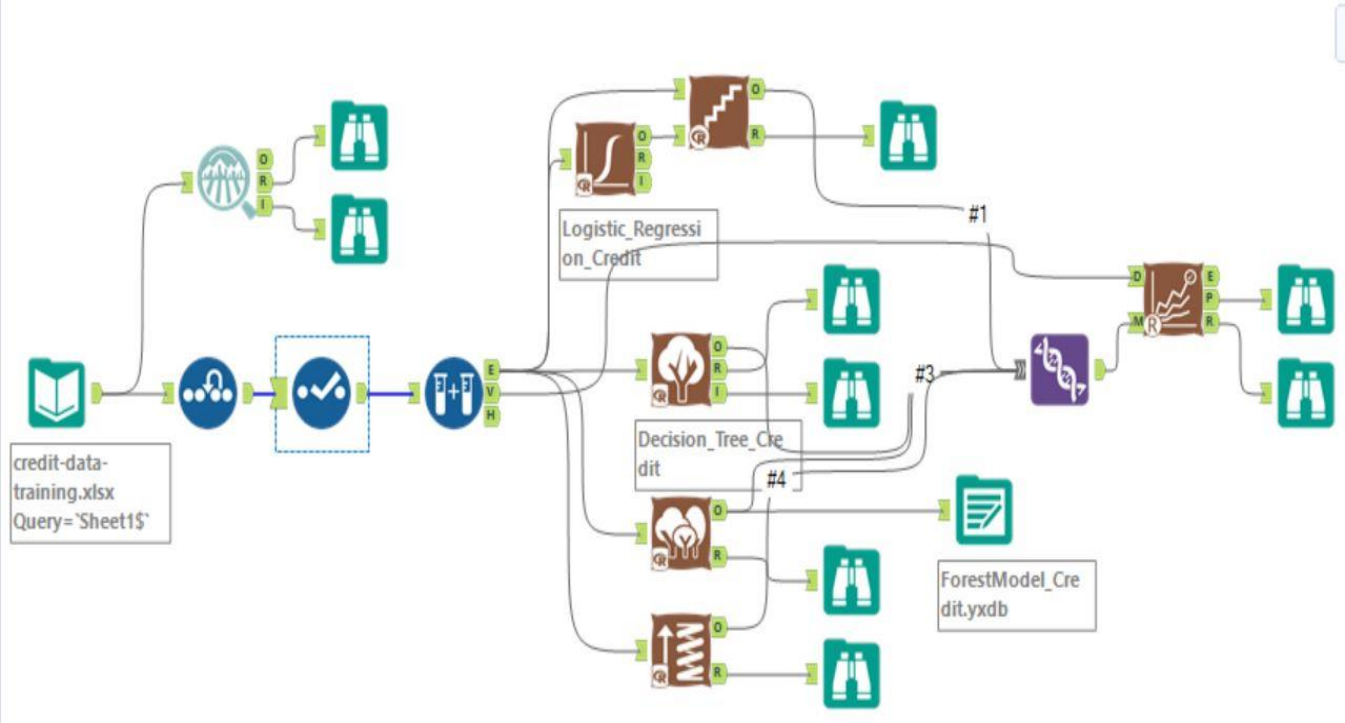
I will be answering to the following questions for each of my 4 models:

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.
- Validate your model against the Validation set. What was the overall percent accuracy?

Firstly, I built the 4 models in Alteryx and then used the model comparison tool to validate these models.

- Logistic Stepwise
- Decision Tree
- Random Forest Model
- Boosted Model

Below, is the complete workflow that I created in Alteryx. I will discuss the results of each models separately.



Logistic Stepwise: As per the report below**,** the significant predictive variables (based on their p values) are Account Balance, Payment Status, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment and Installment per cent.

Report for Logistic Regression Model Stepwise_Credit

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:

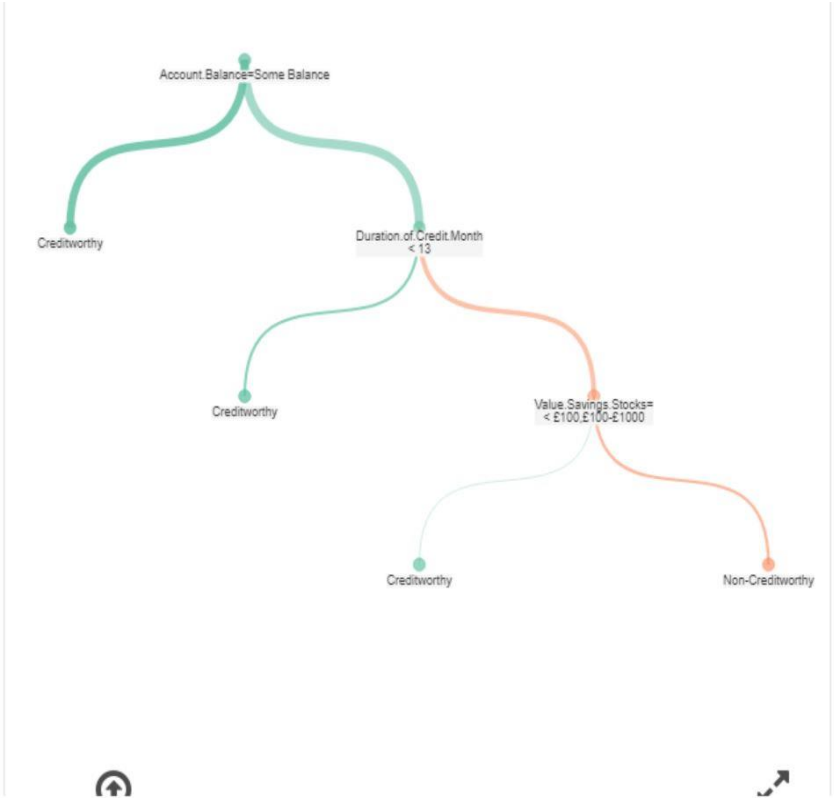
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01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

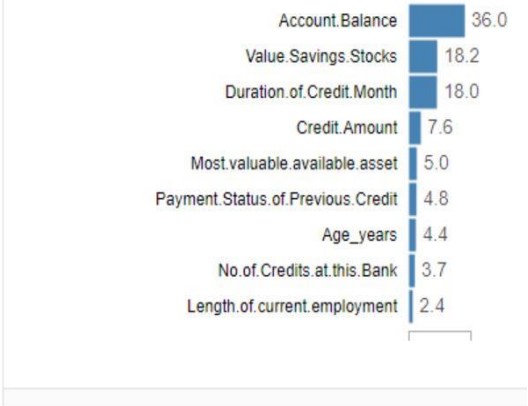
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

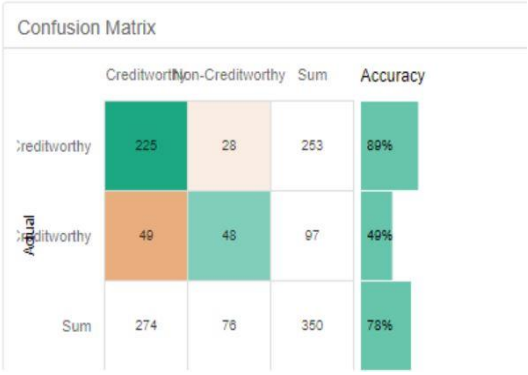
(Dispersion parameter for binomial taken to be 1)

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

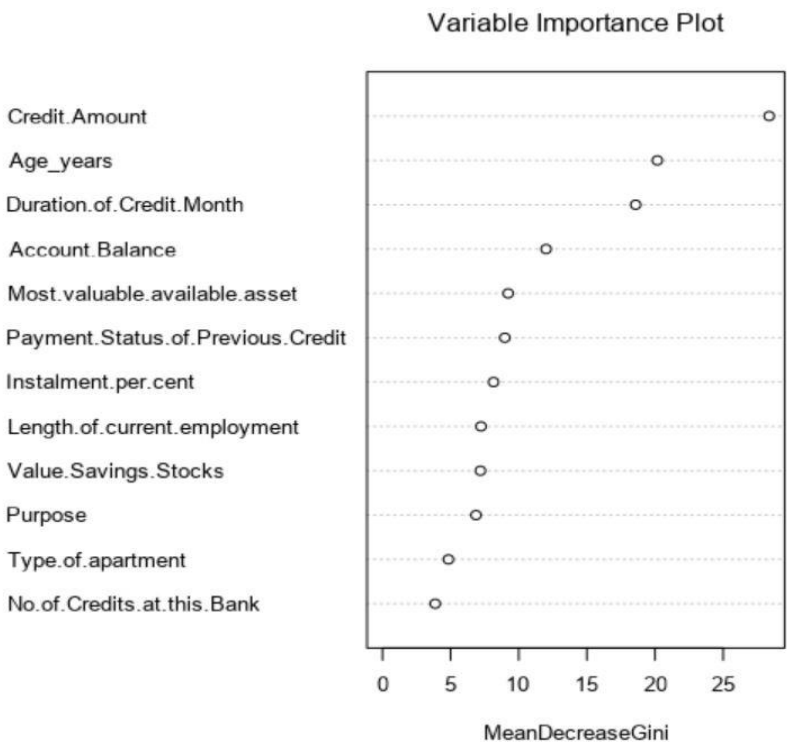
Decision Tree: As we can see below from the decision tree and the variable importance report (on the right of the diagram) the top 3 predictive variable based on which the splits have been made are Account Balance, Duration of Credit Month and Value Savings Stocks.







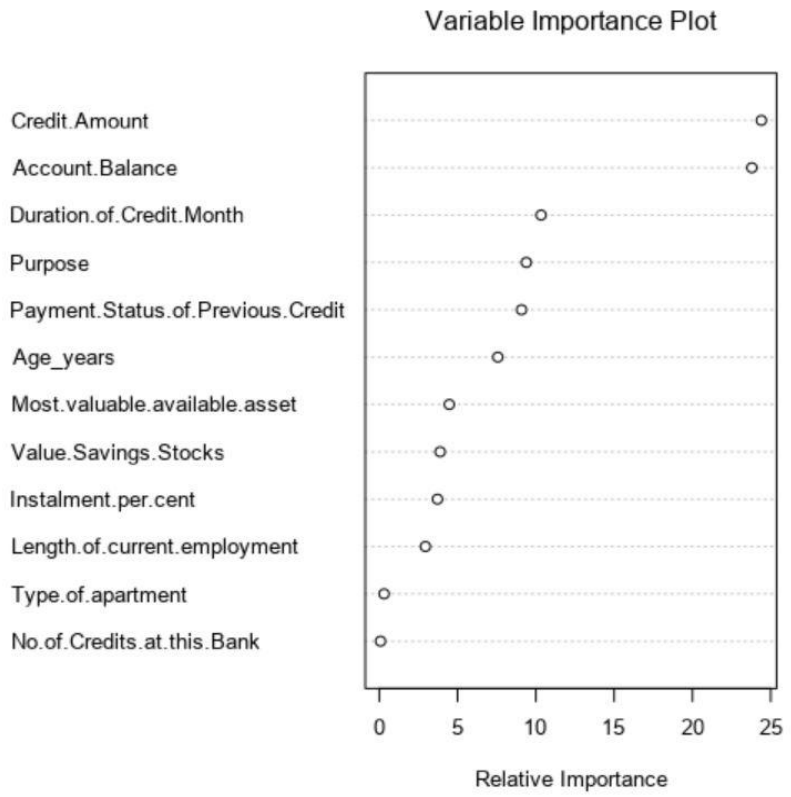
Random Forest Model:



The Variable Importance plot on the left from the random forest model tells us that the top 3 predictive variables are

- Credit Amount
- Age Years
- Duration of Credit month.

Boosted Model:



Based on the Variable Importance plot on the left from the Boosted model tells us that the top 3 predictive variables are

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

Validate and Compare Models

Using the model comparison tool to validate and compare the four models the output is as follows:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Credit	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
ForestModel_Credit	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Credit	0.7867	0.8632	0.7524	0.9619	0.3778

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.

As per the report, we can see that the overall accuracy for
Logistic Stepwise is 0.7600, Decision Tree is 0.7467, Forest Model is 0.8000 and Boosted Model is 0.7867.

Now, let's take a look at the confusion matrix for all these models:

Confusion matrix of Boosted_Credit

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

Confusion matrix of Decision_Tree_Credit

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

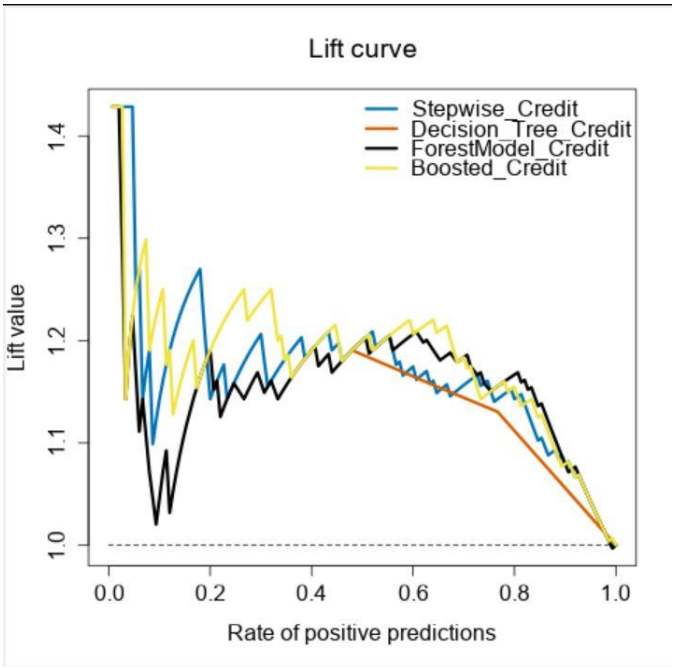
Confusion matrix of ForestModel_Credit

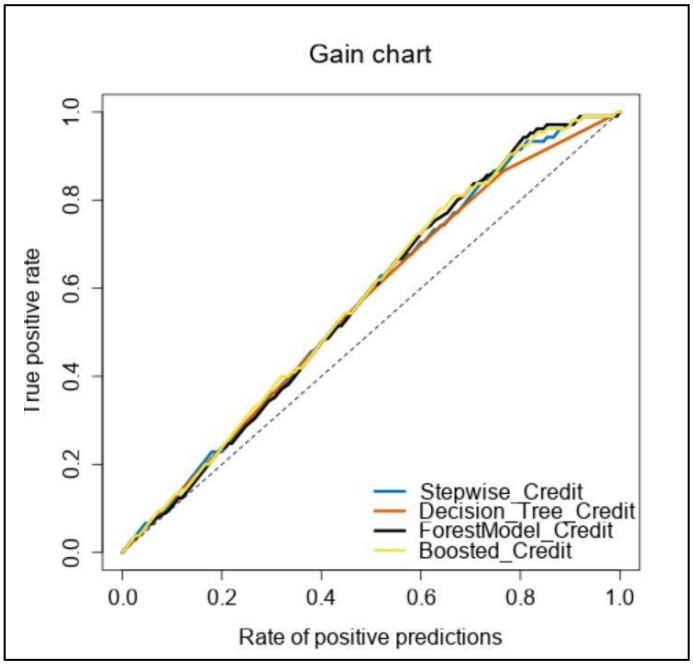
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Stepwise_Credit

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

We can see that we have an imbalanced dataset here i.e., there are more creditworthy applicants than non-creditworthy applicants.





Step 4: Writeup

It is important to remember that my boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.
Therefore, in order to select a model for prediction we will look at the following techniques

- Overall Accuracy against your Validation set - As we mentioned earlier Forest Model has higher overall accuracy followed by Boosted Model, Stepwise and lastly Decision Tree.
- Accuracies within “Creditworthy” and “Non-Creditworthy” segments – When we see the overall Accuracy within the Creditworthy and Non-Creditworthy, there is no difference between Forest and Boosted for Accuracy Creditworthy which is 0.9619 for both models. The Forest model however, has slightly better Accuracy for Non- Creditworthy which is 0.4222, whereas Boosted has 0.3778.
- Bias in the Confusion Matrices- Let’s look at the Confusion Matrix again:

Confusion matrix of Boosted_Credit

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

Confusion matrix of Decision_Tree_Credit

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

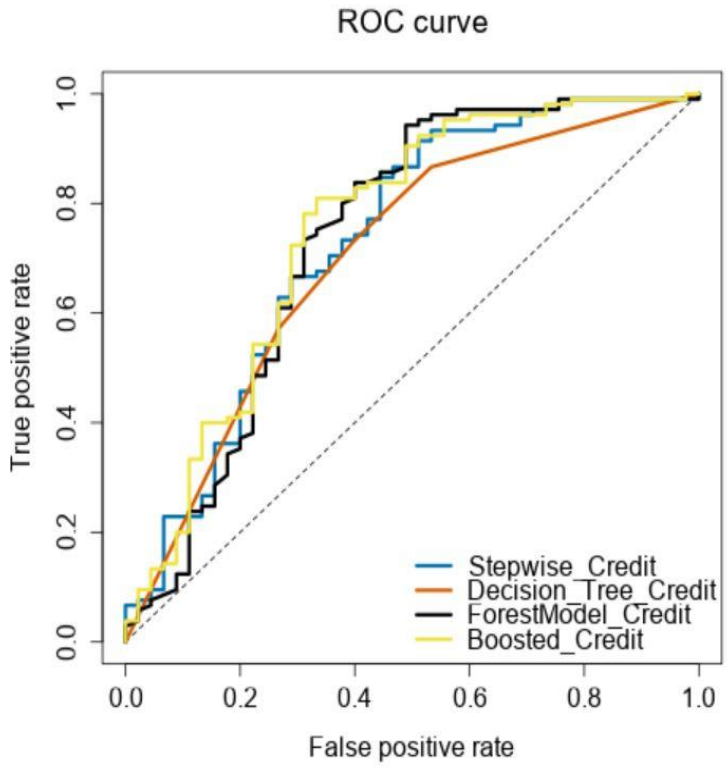
Confusion matrix of ForestModel_Credit

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

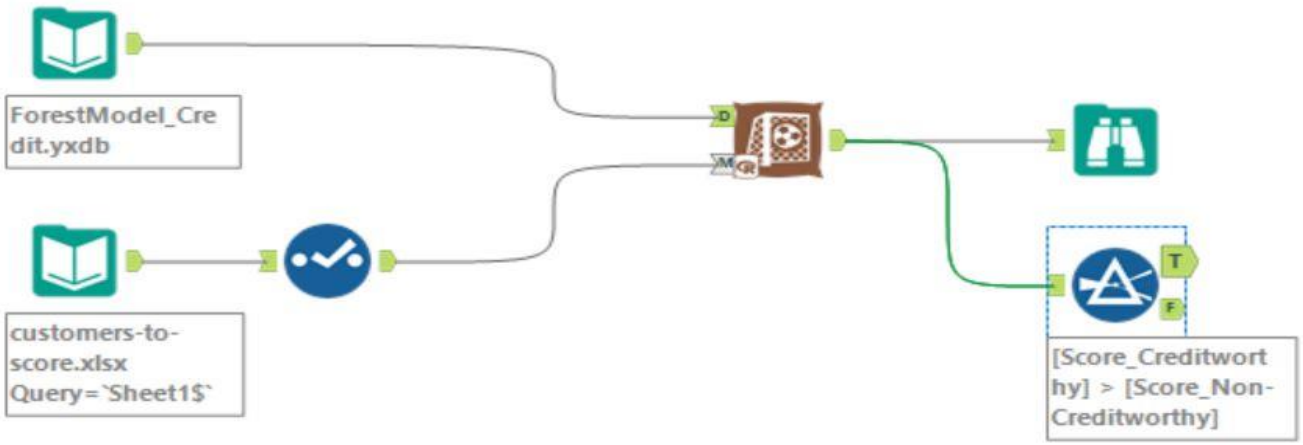
Confusion matrix of Stepwise_Credit

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

- ROC graph: We can say that the models are biased towards predicting individuals who are creditworthy, as they are not predicting individuals who are not creditworthy nearly at the same level as those who are.



Considering all the overall accuracy, PPV, NPV, F1 score and the ROC graph, I have decided to use Random Forest Model. The higher the values, the more accurate the model is. Next to find out the number of applicants who are creditworthy, I saved the Random Forest model and then applied the Score tool. Here is the following workflow:



- How many individuals are creditworthy?
  My result shows that **406** individuals are Creditworthy.