Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. **What decisions needs to be made?**

   The Decision that needs to be taken here is to classify the customer either as credit worthy or non-credit worthy and based on that we need classify the customer into two categories.

2. **What data is needed to inform those decisions?**

   We need to know information that are:

   - **'credit-data-training.xlsx'**: This file contains the data of the customers to whom bank has provided the loan to and based on that data we can make predictive model to analyse the 'credit-data-training.xlsx' data set and can categorize the customer into credit worthy or non-credit worthy.
   - The variables which will be useful in deciding the credit worthiness of the customer will be: Account Balance, Credit Amount, Payment Status of Previous Credit, Purpose New Car, Value Saving Stocks, Age-Year, Duration of Credit Month.

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

   The model that we need to build for this problem is Binary model as we need to decide whether the customer is credit worthy or non-credit worthy.

Step 2: Building the Training Set

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

   There are several fields which are removed and one field which is imputed and they are:

   - The Histogram of the variable **Guarantors, Foreign-Worker** and **No of dependents** have shown that majority of data is heavily skewed towards one type of data. These variables are removed due to heavily skewed towards one data.
   - **Concurrent Credits** and **Occupation** both of them have entirely uniform data and there are no other variations in the data that's why both of them are removed due to low variability.
   - **Duration in Current address** has 69% of the missing data. That's why this field is removed.
   - **Telephone,** this field does not have any predictive ability to credit application results, so this field is also removed.
   - **Age-Year** this field has 2% of missing value. The missing data of this variable has been imputed using the median, 33 of the entire data field.

Step 3: Train your Classification Models

1. **Logistic Regression (Stepwise):**

   From below Chart it can be observer that the most important predictor variables for the logistic regression(stepwise) model are Account Balance Some Balance, Purpose New car and Credit Amount. The p-value of all these variables can be observed from below chart.

Report for Logistic Regression Model LR_Stepwise

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

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Below is the model comparison report for Logistic Regression (Stepwise) which shows that this model has an accuracy of 76%. Using Confusion Matrix,

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 92/ (92+23) = 0.8 = 80%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 22/ (13+22) = 0.6286 = 62.86%

The models seem to be slightly biased towards predicting customer as non-creditworthy.

Model Comparison Report

Fit and error measures

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

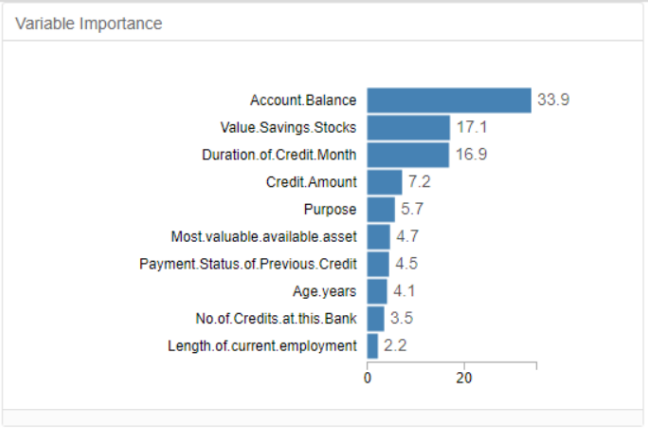
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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.

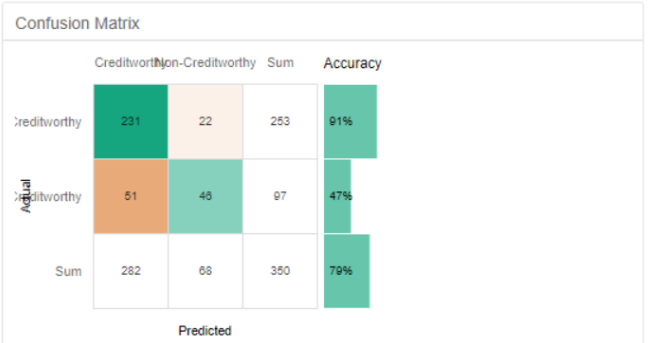
Confusion matrix of LR_Stepwise

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

2. Decision Tree

The significant predictor variable for Decision Tree model can be observed from the variable important graph and these are: Account Balance, Value Savings Stocks and Duration of Credit Month.





Below is the model comparison report for Decision Tree which shows that this model has an accuracy of 74.67%. Using Confusion Matrix,

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 93/ (93+26) = 0.7815 = 78.15%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 19/ (12+19) = 0.6129 = 61.29%

The models seem to be slightly biased towards predicting customer as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222

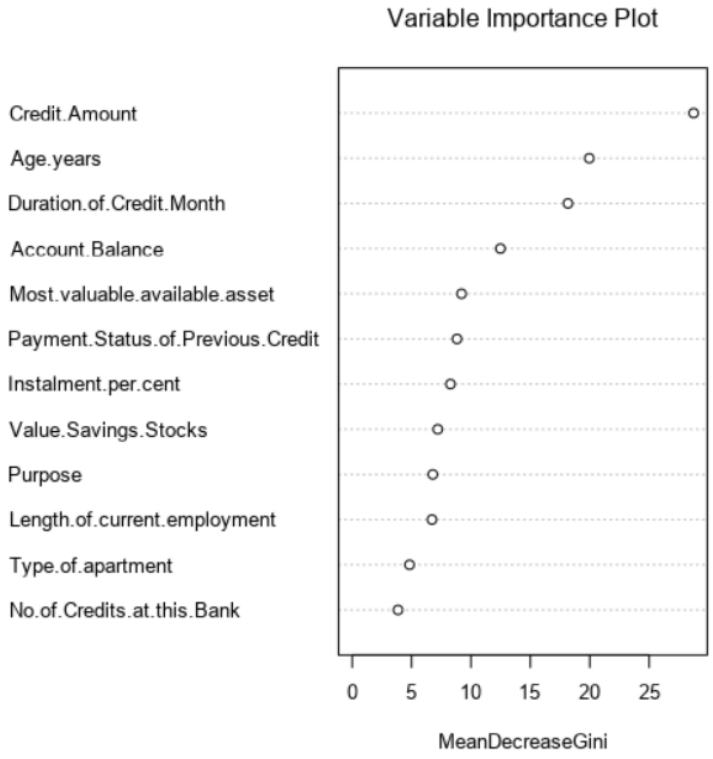
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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 DT

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

3. Forest Model:

From the Variable importance plot of the forest model, we can infer that the most important predictor variables for this model are Credit Amount, Age Years and Duration of Credit Month.



Below is the model comparison report for Forest Model which shows that this model has an accuracy of 79.33%. Using Confusion Matrix,

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 102/ (102+28) = 0.7846 = 78.46%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 17/ (3+17) = 0.85 = 85%

Since accuracies for creditworthy and non-creditworthy are comparable 78.46% and 85% respectively, this model isn't biased.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RF	0.7933	0.8681	0.7368	0.9714	0.3778

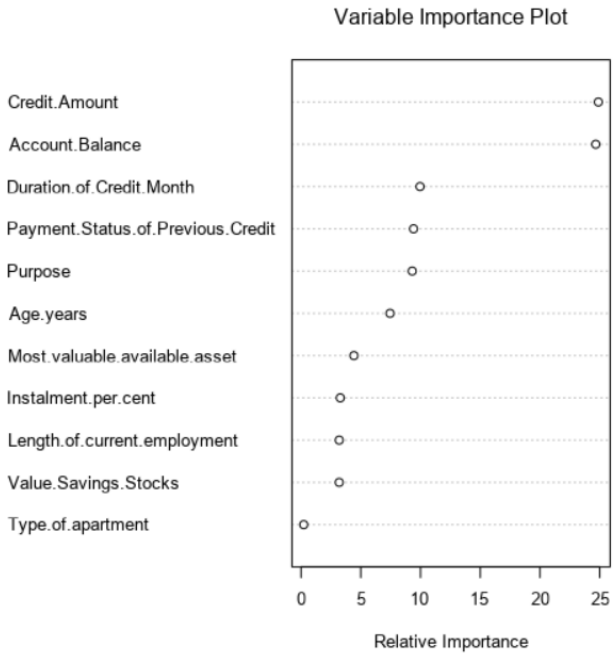
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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.

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

4. Boosted Model:

From the Variable importance plot of the boosted model, we can infer that the most important predictor variables for this model are Credit Amount, Account Balance and Duration of Credit Month.



Below is the model comparison report for Boosted Model which shows that this model has an accuracy of 78.67%. Using Confusion Matrix,

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 101/ (101+28) = 0.7829 = 78.29%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 17/ (4+17) = 0.6129 = 80.95%

Since accuracies for creditworthy and non-creditworthy are comparable 78.29% and 80.95% respectively, this model isn't biased.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BT	0.7867	0.8632	0.7490	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.

Confusion matrix of BT

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

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**

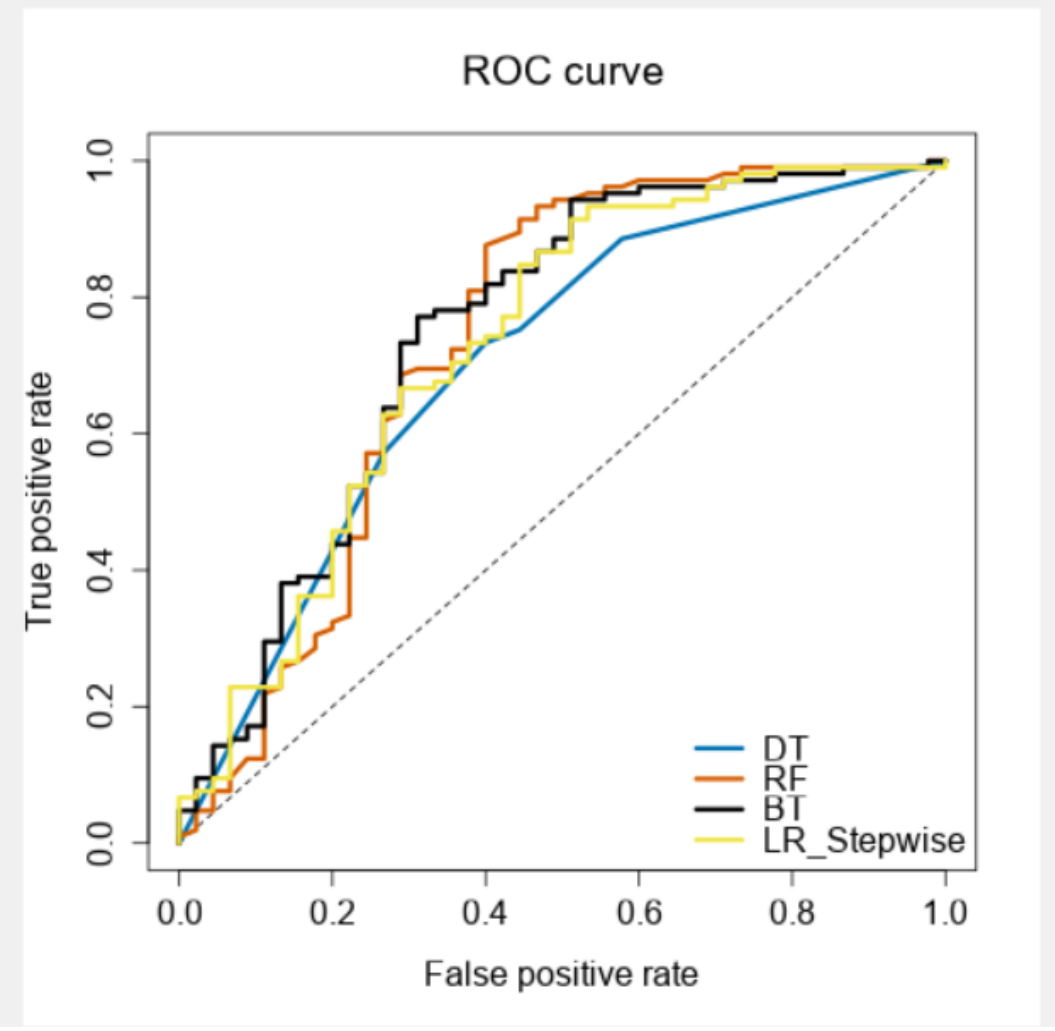
    Forest Model has been chosen since it has the highest accuracy of 79.33%among all four classification models. Also, the accuracies for creditworthy and non-creditworthy are among the highest of all.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222
RF	0.7933	0.8681	0.7368	0.9714	0.3778
BT	0.7867	0.8632	0.7490	0.9619	0.3778
LR_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Forest reaches the top true positive rate the quickest and overall, the highest the most.



Using the confusion matrix of the Forest Model:

Accuracy for creditworthy = (actual creditworthy) / (predicted creditworthy)
= 102/ (102+28) = 0.7846 = 78.46%

Accuracy for non-creditworthy = (actual non-creditworthy) / (predicted non-creditworthy)
= 17/ (3+17) = 0.85 = 85%

Since accuracies for creditworthy and non-creditworthy are comparable 78.46% and 85% respectively, this model isn't biased.

Confusion matrix of BT

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

Confusion matrix of DT

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

Confusion matrix of LR_Stepwise

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

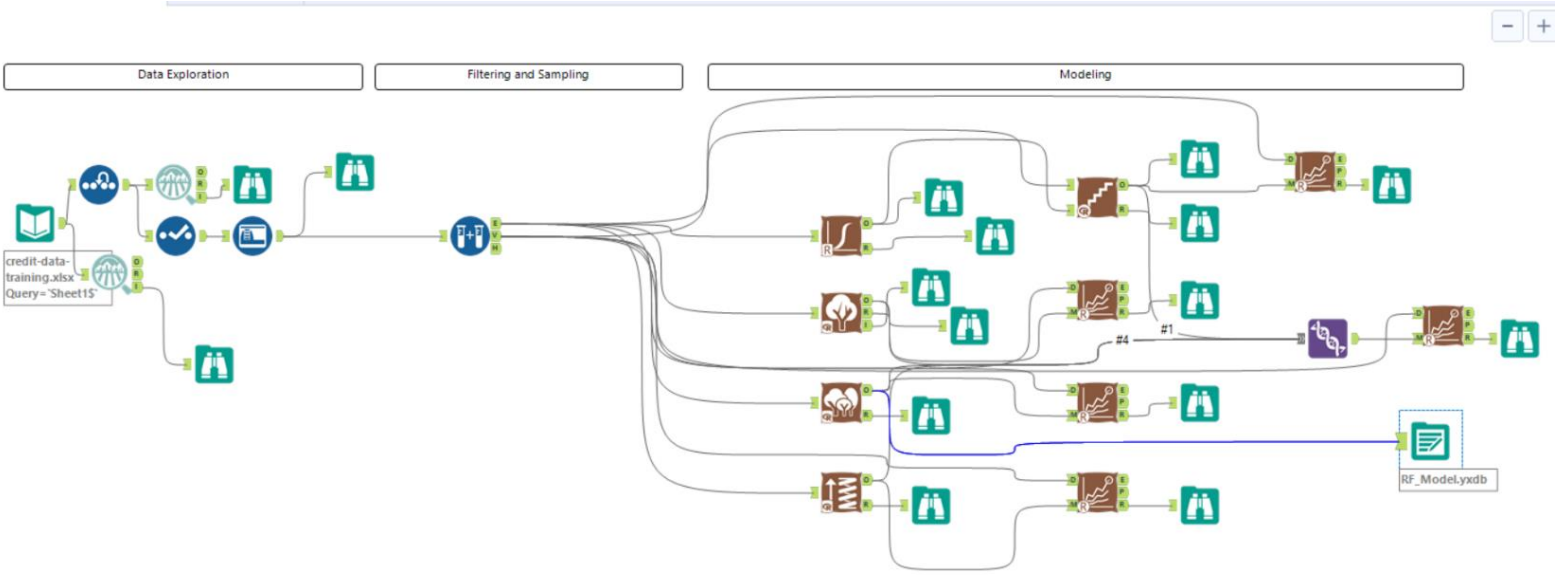
Confusion matrix of RF

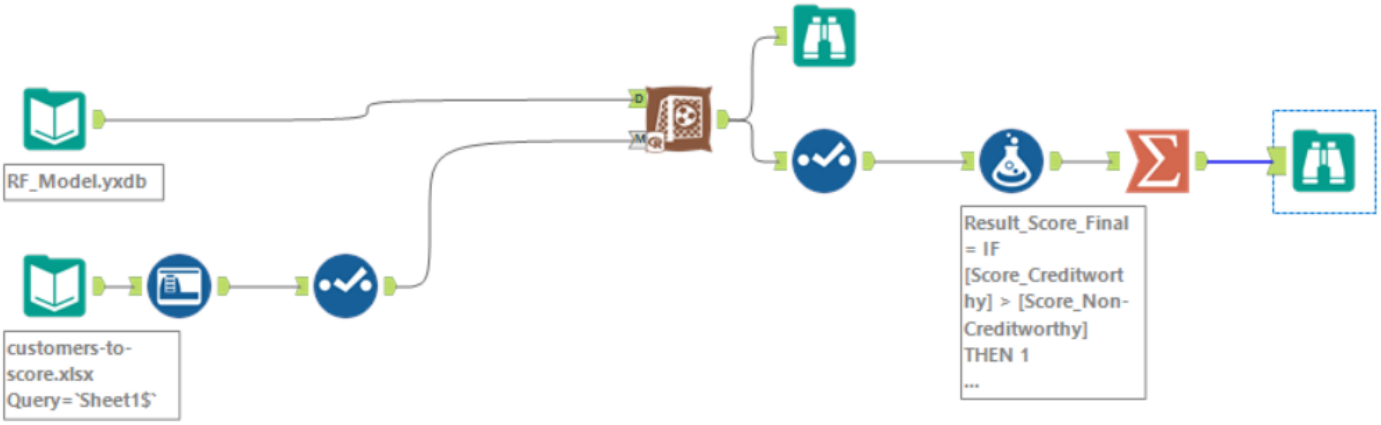
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

2. How many individuals are creditworthy?

There are 410 creditworthy new customers that we could approve for a loan and 92 non-creditworthy customers that should not be approved for a loan.

My Alteryx Workflow




RF_Model.yxdb
customers-to-score.xlsx Query=`Sheet1$`
Result_Score_Final = IF [Score_Creditworthy] > [Score_Non-Creditworthy] THEN 1
...