Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

  The key business decision is to determine the creditworthiness of new loan applicants using a Binary Classification Model.

- What data is needed to inform those decisions?

  We will use the data on all past applicants (*credit-data-training.xlsx*) to create and train the model. Afterwards, the data on new customers will be scored using the created classification model to determine which applicants are creditworthy (*customers-to-score.xlsx*).

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

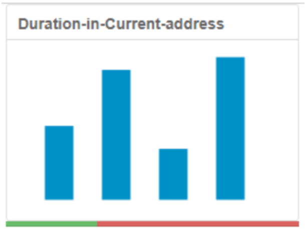
  We will use a Binary Classification Model since we will predict if a new customer is *Creditworthy* or *Non-Creditworthy*.

Step 2: Building the Training Set

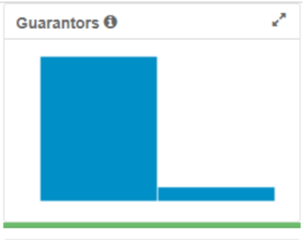
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

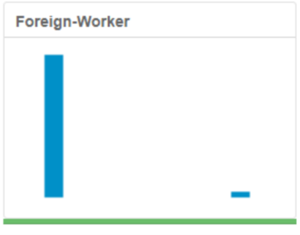
Removed Fields:

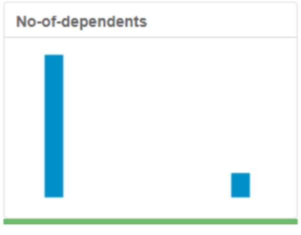
- *Duration-in-Current-address*: 69% of values are missing data.



- *Guarantors, Foreign-Worker, No-of-dependents*: The following fields are heavily skewed towards one type of data (low variability).

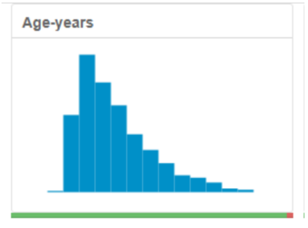






- *Concurrent-Credits, Occupation*: The following fieds have data that is entirely uniform and there is no other variations of the data.
- *Telephone*: There is no logical reason for including this field as advised.

Imputed Field: *Age-years* has 2% missing data and has been imputed using the median.

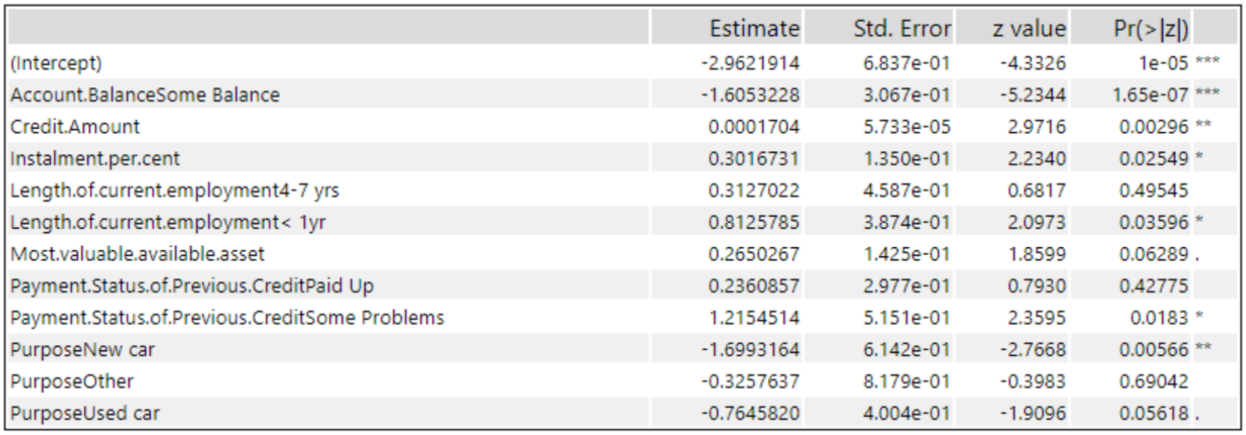


Step 3: Train your Classification Models

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

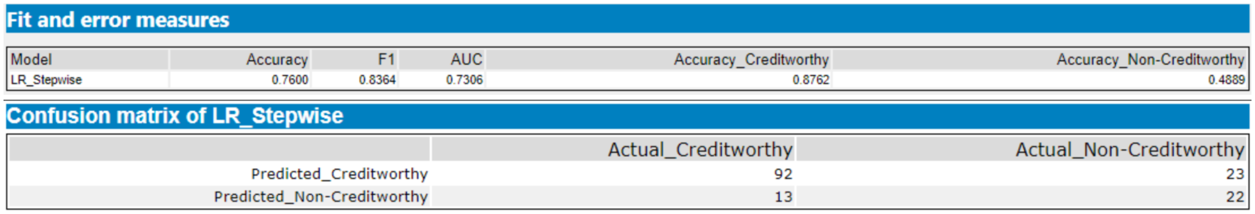
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

Logistic Regression Model



	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	***
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
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	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.
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	.

The significant predictor variables for the Logistic Regression Model are [*Account.Balance*], [*Credit.Amount*], [*Instalment.per.cent*], [*Length.of.current.employment*], [*Most.valuable.available.asset*], [*Payment.Status.of.Previous.Credit*], and [*Purpose*].



Fit and error measures

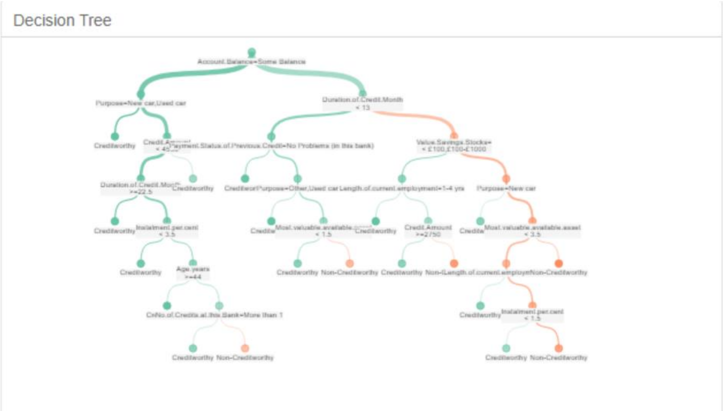
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

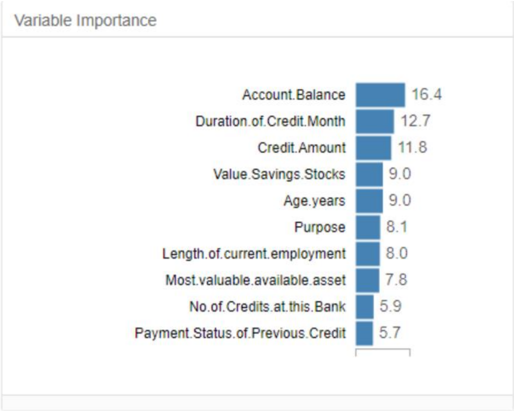
Confusion matrix of LR_Stepwise

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

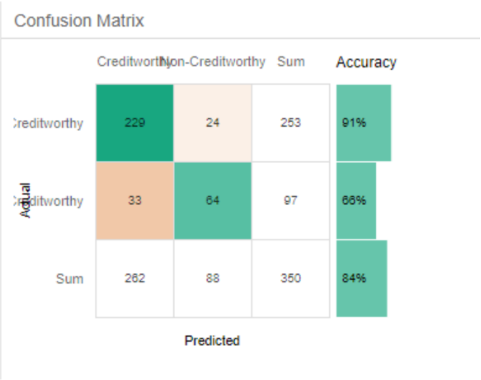
The Logistic Regression Model's overall percent accuracy is 76%. It leans towards predicting new applicants as creditworthy than not. The model's Positive Predictive Value (PPV) is 0.80 while the Negative Predictive Value (NPV) is 0.63.

Decision Tree


Decision Tree
Account.Balance=Some Balance
Purpose=New car,Used car
Duration.of.Credit.Month < 13
Creditworthy
Credit.Amount < 45...
Payment.Status.of.Previous.Credit=No Problems (in this bank)
Value.Savings.Stocks= < £100,£100-£1000
Duration.of.Credit.Month >=22.5
Creditworthy
Creditworthy
Purpose=Other,Used car
Length.of.current.employment=1-4 yrs
Purpose=New car
Creditworthy
Instalment.per.cent < 3.5
Creditw
Most.valuable.available.asset < 1.5
Creditworthy
Credit.Amount >=2750
Creditw
Most.valuable.available.asset < 3.5
Creditworthy
Age.years >=44
Creditworthy
Non-Creditworthy
Creditworthy
Non-Creditworthy
Length.of.current.employ
Non-Creditworthy
No.of.Credits.at.this.Bank=More than 1
Creditworthy
Instalment.per.cent < 1.5
Creditworthy
Non-Creditworthy
Creditworthy
Non-Creditworthy



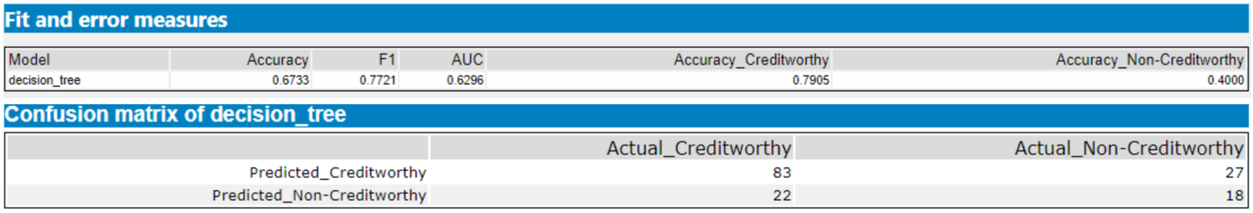
The significant predictor variables for the Decision Tree Model are [*Account.Balance*], [*Duration.of.Credit.Month*], and [*Credit.Amount*].



Confusion Matrix

Actual	Creditworth	Non-Creditworthy	Sum	Accuracy
reditworthy	229	24	253	91%
editworthy	33	64	97	66%
Sum	262	88	350	84%

Predicted



Fit and error measures

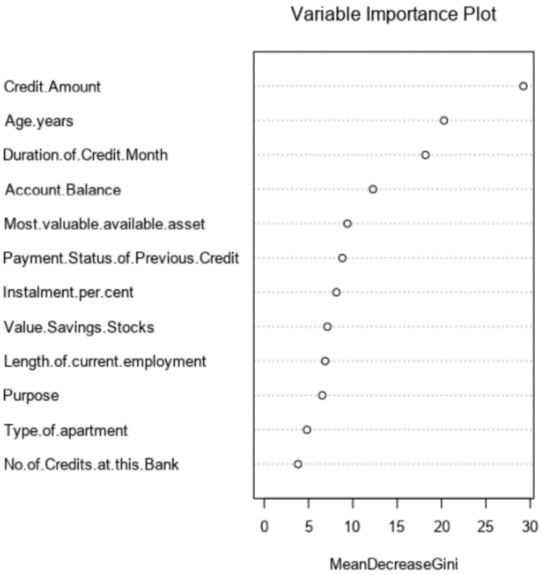
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
decision_tree	0.6733	0.7721	0.6296	0.7905	0.4000

Confusion matrix of decision_tree

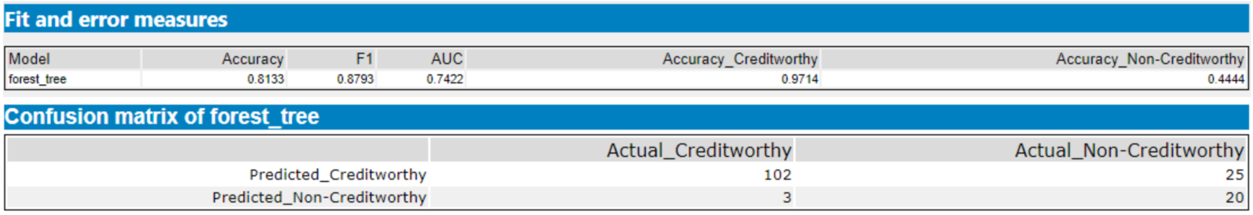
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

The Decision Tree Model's overall accuracy is 67%. There are more actual non-creditworthy that are predicted creditworthy compared to actual creditworthy predicted non-creditworthy. The model's Positive Predictive Value (PPV) is 0.75 while the Negative Predictive Value (NPV) is 0.45.

Forest Model



Based on the Forest Model's Variable Importance Plot, the significant predictor variables are *[Credit.Amount]*, *[Age.years]*, and *[Duration.of.Credit.Month]*.



Fit and error measures

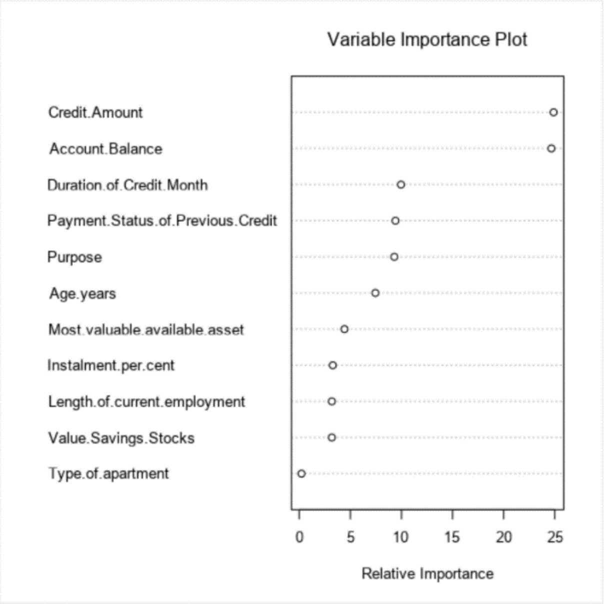
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
forest_tree	0.8133	0.8793	0.7422	0.9714	0.4444

Confusion matrix of forest_tree

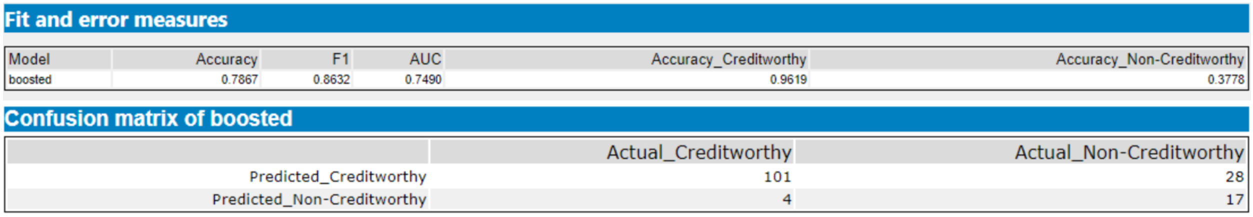
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

The Forest Model's overall accuracy is 81% and is higher to all the tested models' accuracy. Its Positive Predictive Value (PPV) is 0.80 while the Negative Predictive Value (NPV) is 0.87. This model has little to no bias in its predictions as it has one of the least differences in NPV and PPV values.

Boosted Model



Based on the Boosted Model's Variable Importance Plot, the significant predictor variables are *[Credit.Amount]* and *[Account.Balance]*.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
boosted	0.7867	0.8632	0.7490	0.9619	0.3778

Confusion matrix of boosted

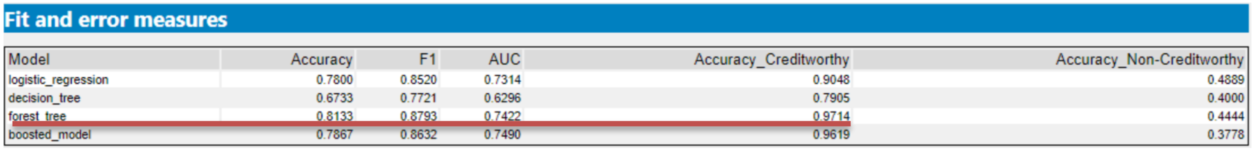
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

The Boosted Model's overall accuracy is 79%. There are more actual non-creditworthy that are predicted creditworthy compared to actual creditworthy predicted non-creditworthy. The model's Positive Predictive Value (PPV) is 0.78 while the Negative Predictive Value (NPV) is 0.81.

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

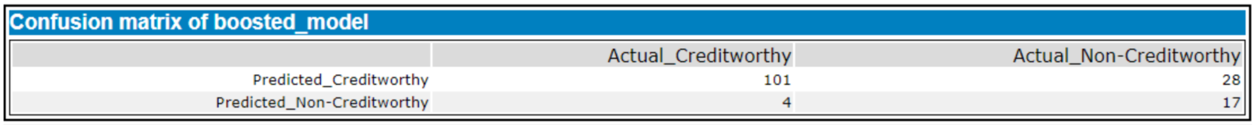


Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_regression	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.6733	0.7721	0.6296	0.7905	0.4000
forest_tree	0.8133	0.8793	0.7422	0.9714	0.4444
boosted_model	0.7867	0.8632	0.7490	0.9619	0.3778

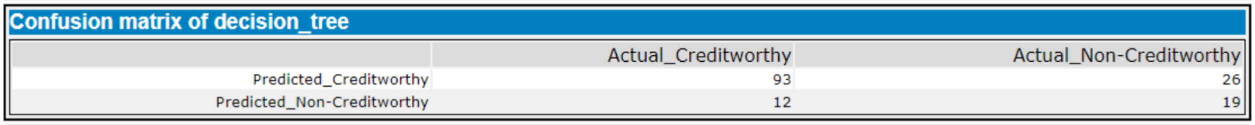
Among all the models evaluated with the estimation and validation data, the Forest Tree Model has the highest overall accuracy (81%). This is followed by the Boosted Model (79%), Logistic Regression Model (78%), and Decision Tree Model (67%). Further, the Forest Tree Model's accuracy for predicting creditworthy applicants is 97%.

    ○ Accuracies within "Creditworthy" and "Non-Creditworthy" segments



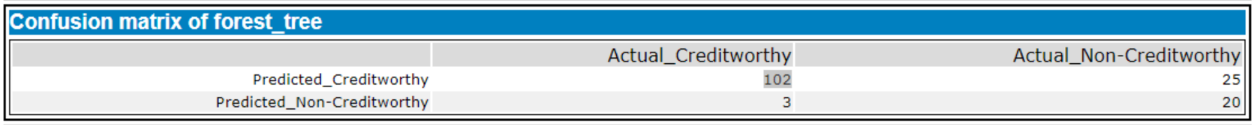
Confusion matrix of boosted_model

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



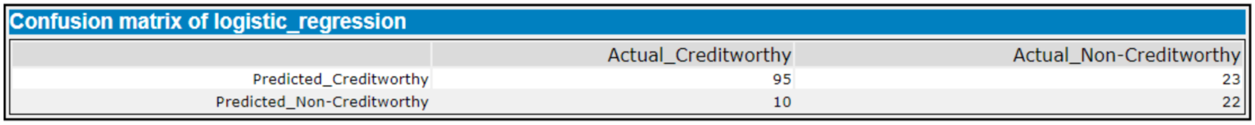
Confusion matrix of decision_tree

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



Confusion matrix of forest_tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

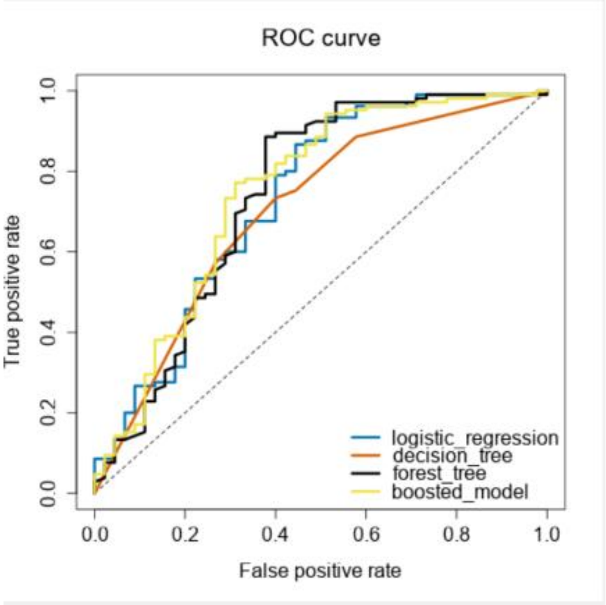


Confusion matrix of logistic_regression

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

The Forest Tree Model correctly predicted the most creditworthy applicants (102) compared to the other models. It also predicted the least non-creditworthy applicants (3) that are actual creditworthy.

- ROC graph



The Forrest Tree Model has the highest true positive rates among all models.

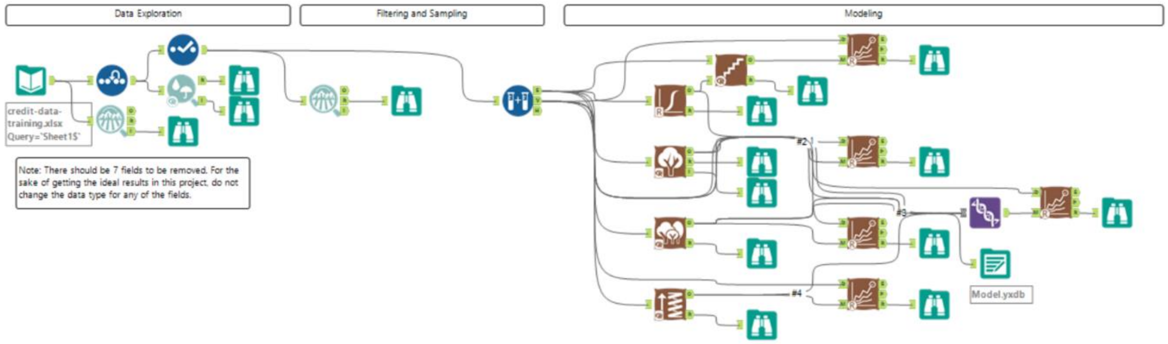
- Bias in the Confusion Matrices

The Forest Tree Model reduces the possibility or risk of over fitting data since it uses multiple decision trees at random compared to a single decision tree.

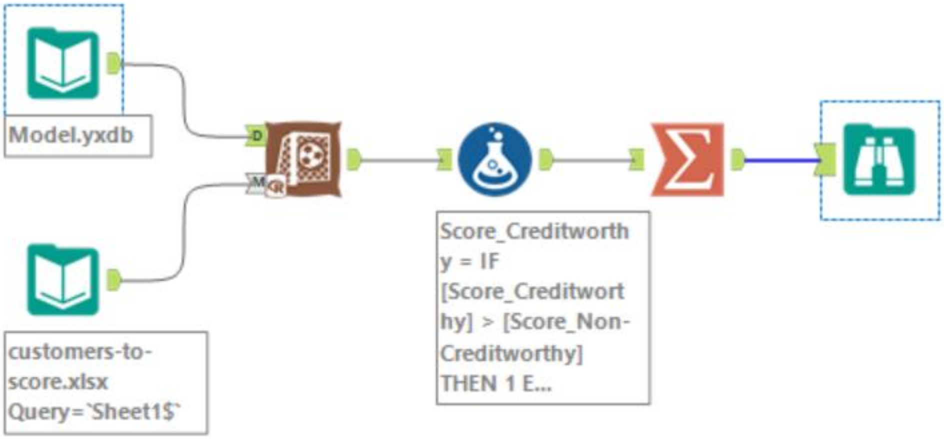
- How many individuals are creditworthy?

Based on the score of new customers, there are **412** creditworthy loan applicants.

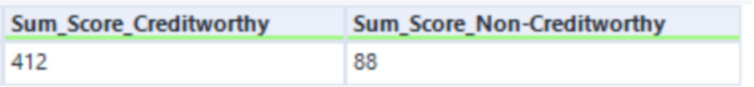
Alteryx Workflow



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 E...
```



Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
412	88