Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

  To determine if a customer is credit worthy to be given a loan and to figure out how to process all the new loans.

- What data is needed to inform those decisions?

  1. Data of the old applications
  2. The new list of customers that should be processed in the next few days for scoring

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

  I will be testing the business problem as a classification binary model thus, I will be using logistic regression, decision tree, forest model, and boosted model to find out the best suitable model to figure out If we could give a loan or not.

Step 2: Building the Training Set

*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't **need to convert any data fields to the appropriate data types.***

Since the data has already been cleaned up, I have started to explore the coloration between the variables by using a field summary tool.

I will remove **Duration-in-Current-address** as it is missing 69% of the data and if we impute it will skew our results

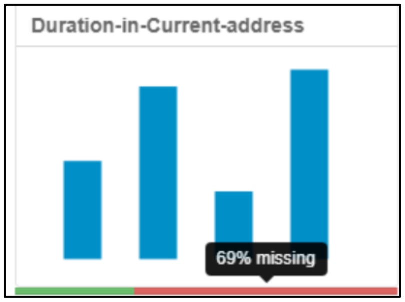
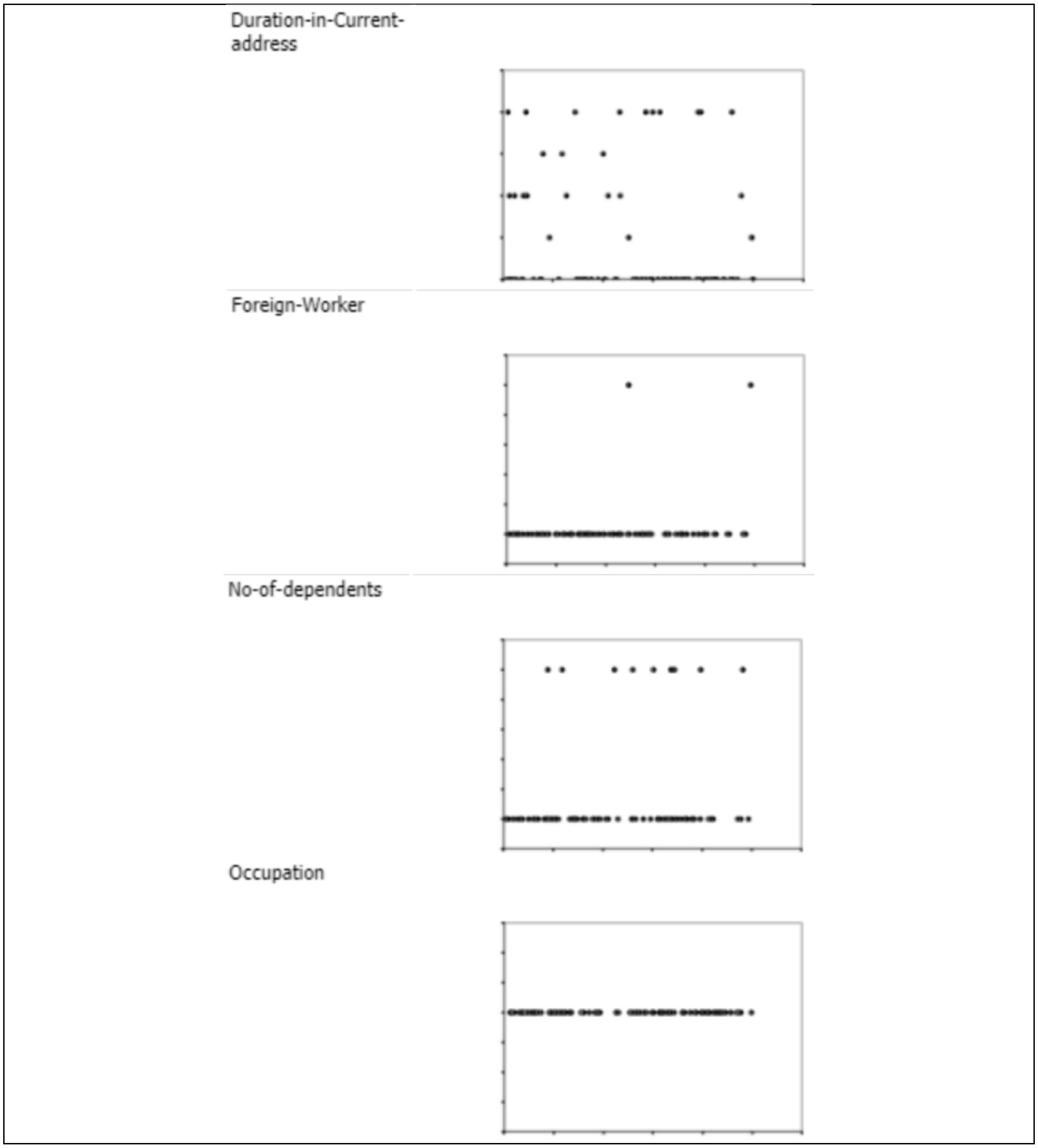


Figure 1: summary tool results of duration-in-current-address

I will also be removing the following fields as they are showing small number of unique values

1. Telephone
2. Occupation
3. No-of-dependents (low variability)
4. Foreign-worker (low variability)
5. Guarantors (low variability)
6. Concurrent-Credits (low variability)



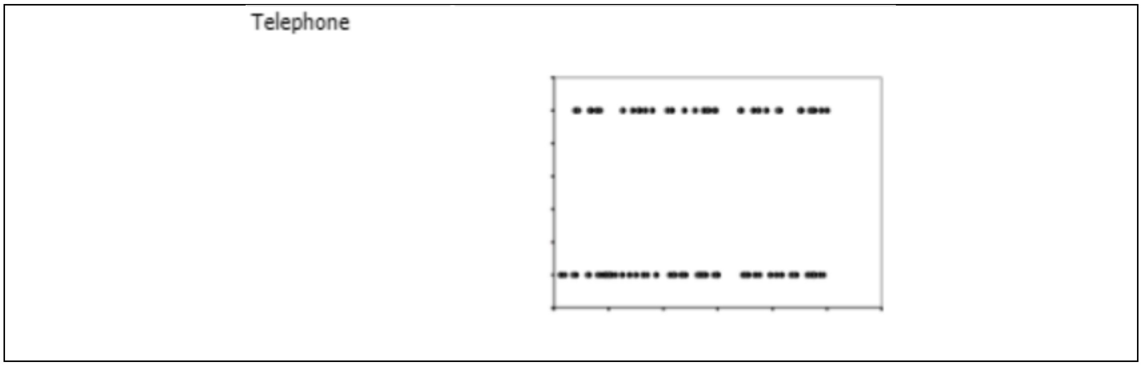


Figure 2: field summery tool results of Telephone, Occupation, number of dependents, foreign workers, and duration in current address

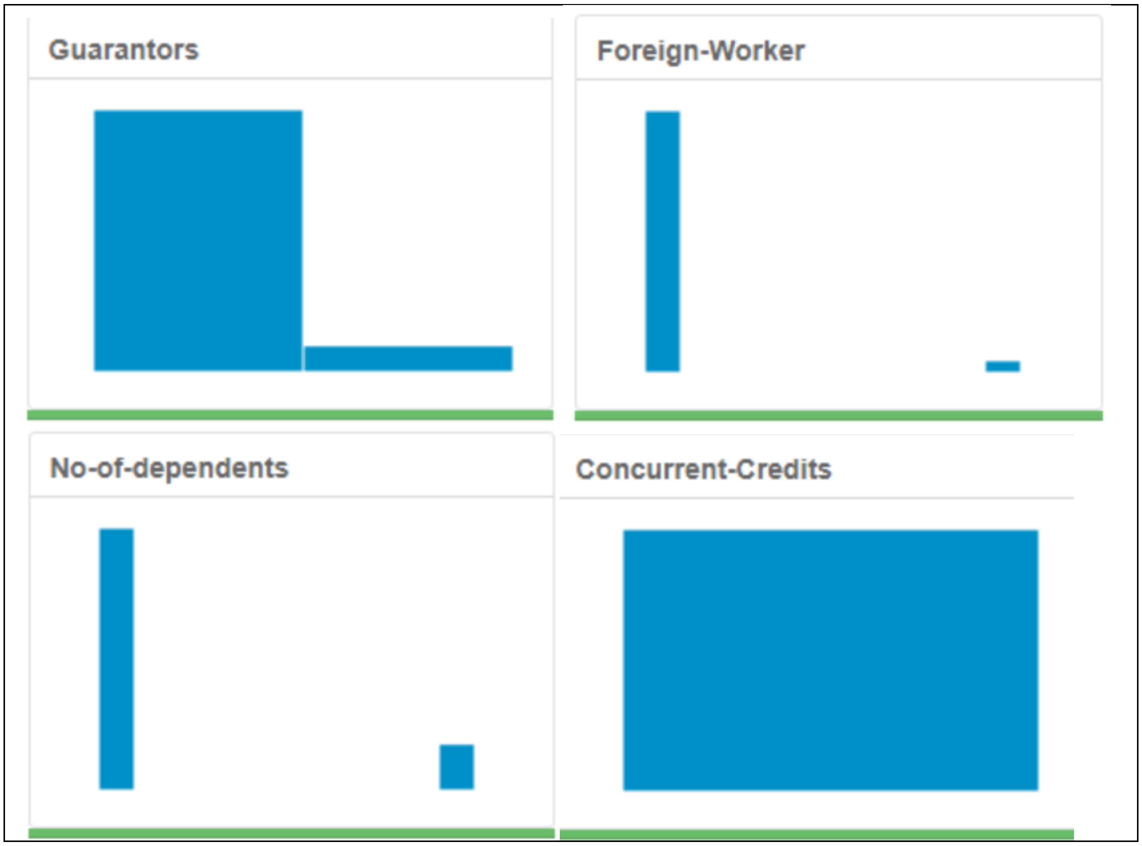


Figure 3: field summery tool report that shows low variability of guarantors, foreign workers, & no of independents

In age-years I have noticed a 2% missing data and instead of removing the whole column because it is a small percentage and losing the whole column could cause us to lose much more than we need to. I will only be imputing the missing values and replace it with the 33 medians by using the imputation tool

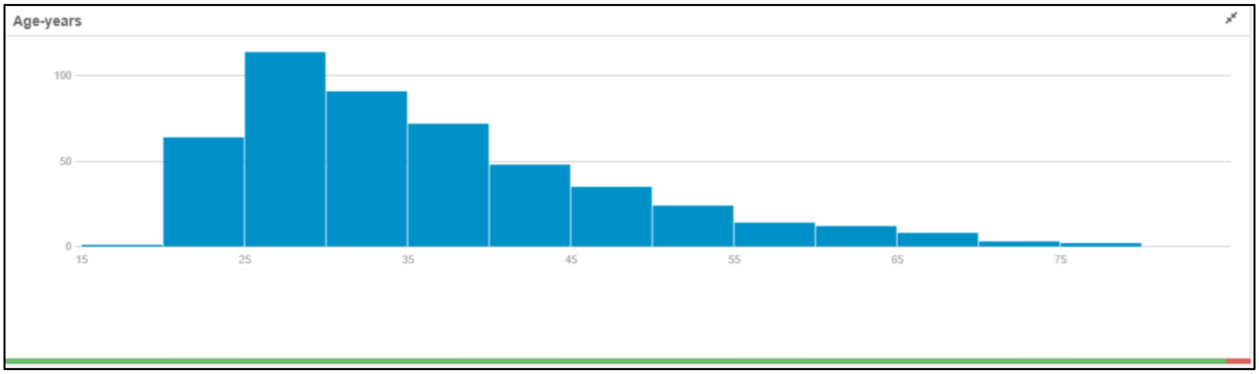


Figure 4: field summary tool showing the 2% of missing data

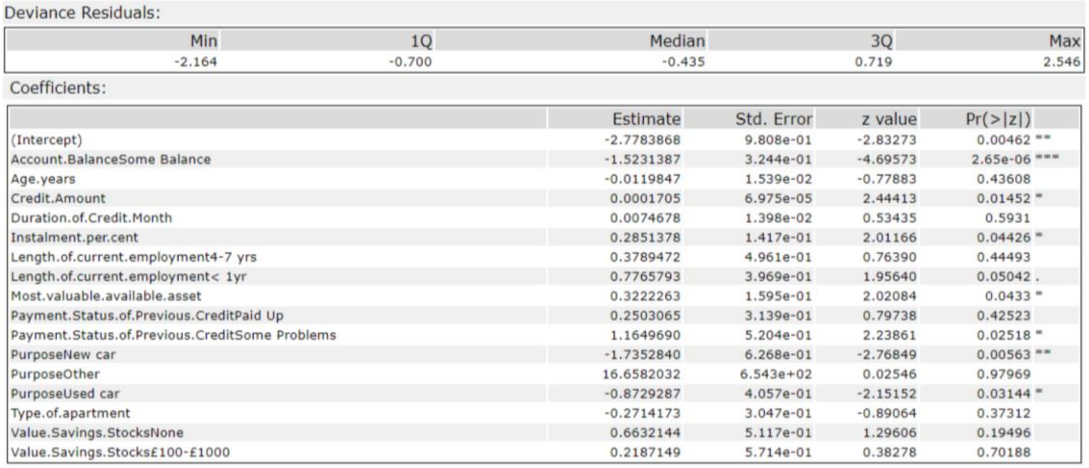
I will choose the following columns to be used in testing the classification models

1) **Account-Balance**
2) **Age-years** (average = 35.6372950819672= 36)
3) **Credit-amount**
4) **Credit-application-result**
5) **Duration-of-credit-month**
6) **Instalments-per-cent**
7) **Length-of-current-employment**
8) **Most-valuable-available-asset**
9) **Payment-status-of-previous-credit**
10) **Purpose**
11) **Type-of-apartment**
12) **Value-saving-stocks**
13) **No-of-credits-at-this-bank**

Step 3: Train your Classification Models

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

Logistic Regression



Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.164	-0.700	-0.435	0.719	2.546

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.7783868	9.808e-01	-2.83273	0.00462	**
Account.BalanceSome Balance	-1.5231387	3.244e-01	-4.69573	2.65e-06	***
Age.years	-0.0119847	1.539e-02	-0.77883	0.43608	
Credit.Amount	0.0001705	6.975e-05	2.44413	0.01452	*
Duration.of.Credit.Month	0.0074678	1.398e-02	0.53435	0.5931	
Instalment.per.cent	0.2851378	1.417e-01	2.01166	0.04426	*
Length.of.current.employment4-7 yrs	0.3789472	4.961e-01	0.76390	0.44493	
Length.of.current.employment< 1yr	0.7765793	3.969e-01	1.95640	0.05042	.
Most.valuable.available.asset	0.3222263	1.595e-01	2.02084	0.0433	*
Payment.Status.of.Previous.CreditPaid Up	0.2503065	3.139e-01	0.79738	0.42523	
Payment.Status.of.Previous.CreditSome Problems	1.1649690	5.204e-01	2.23861	0.02518	*
PurposeNew car	-1.7352840	6.268e-01	-2.76849	0.00563	**
PurposeOther	16.6582032	6.543e+02	0.02546	0.97969	
PurposeUsed car	-0.8729287	4.057e-01	-2.15152	0.03144	*
Type.of.apartment	-0.2714173	3.047e-01	-0.89064	0.37312	
Value.Savings.StocksNone	0.6632144	5.117e-01	1.29606	0.19496	
Value.Savings.Stocks£100-£1000	0.2187149	5.714e-01	0.38278	0.70188	

Figure 5: logistic regression report

Account.balance has showed high significance in logistic regression report, bias in predicting creditworthy

Decision Tree

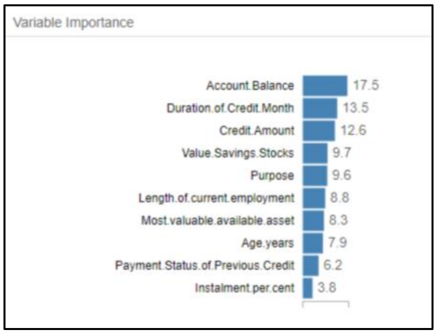


Figure 6: Decision tree report of variable importance

Account.balance has showed high importance in decision tree report

Forest Model

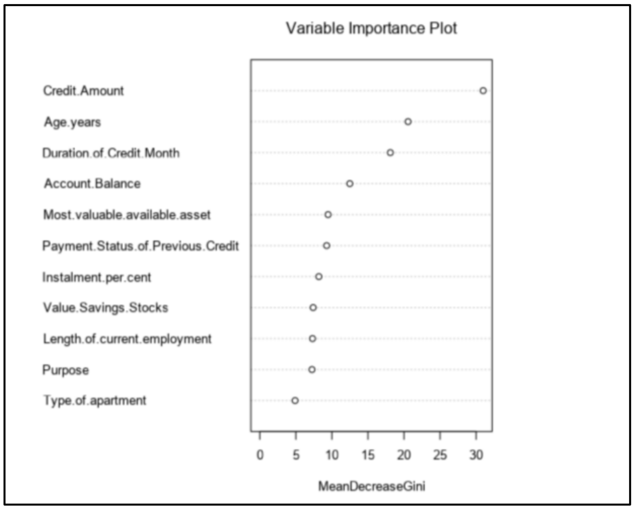


Figure 7: Forest Model report of variable importance

Credit amount is the variable with the highest importance in forest model , No bias in predicting creditworthy or non-credit worthy

Boosted Model

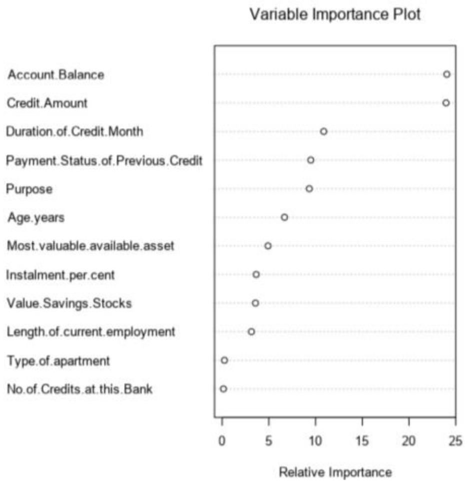


Figure 8: Boosted Model report of variable importance

Account balance and credit amount are both showing variable importance in the boosted model

Step 4: Writeup

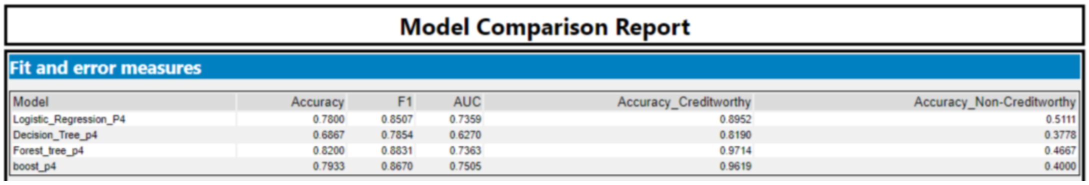
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Answer these questions:

- 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

As per my previous comparisons between models, forest tree has showed the highest results with 81% accuracy thus I'm going to conclude with it

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



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_P4	0.7800	0.8507	0.7359	0.8952	0.5111
Decision_Tree_p4	0.6867	0.7854	0.6270	0.8190	0.3778
Forest_tree_p4	0.8200	0.8831	0.7363	0.9714	0.4667
boost_p4	0.7933	0.8670	0.7505	0.9619	0.4000

Figure 9: model comparison report of all models

Accuracies of non-creditworthy are a bit low which could be concerning

- ○ ROC graph

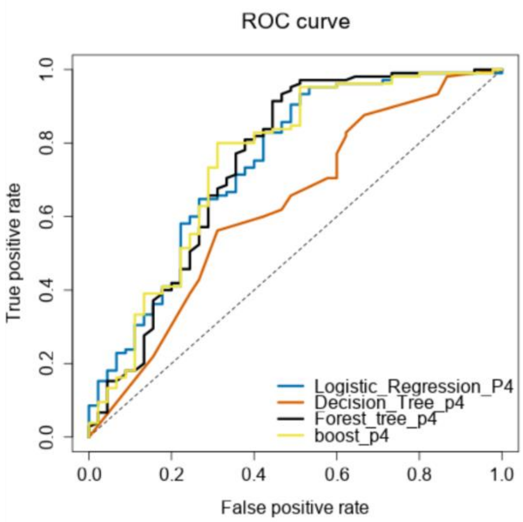
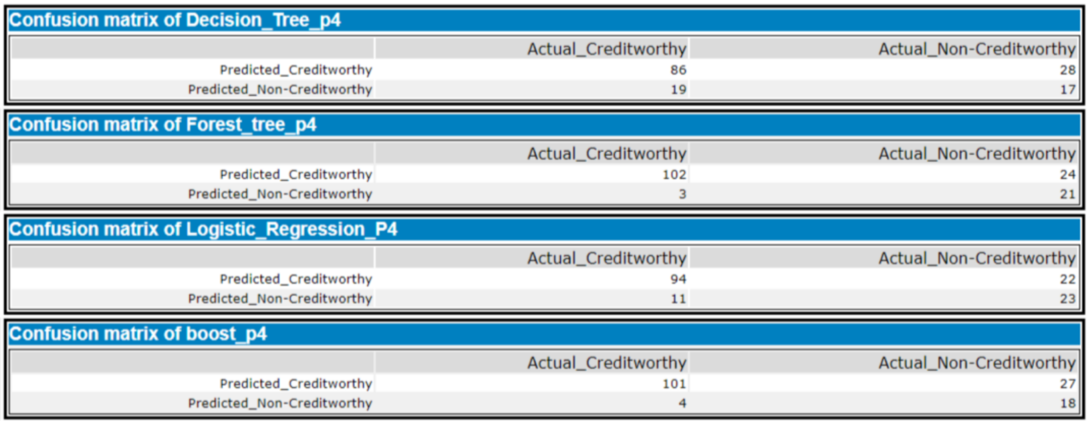


Figure 10: ROC curve of models

ROC has proven that the forest tree has the highest curve and highest true positive rate. It has also reached the top the quickest which means this model will give the best number of true positive predictions

- Bias in the Confusion Matrices

Forest tree shows the highest accuracy rate and decision tree shows the lowest. The model predicted lower numbers in the accuracy of non-credit worthy which shows bias thus, it means the forest tree will be showing the highest true variables for this business problem based on the confusion matrix.



Confusion matrix of Decision_Tree_p4

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	28
Predicted_Non-Creditworthy	19	17

Confusion matrix of Forest_tree_p4

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Confusion matrix of Logistic_Regression_P4

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

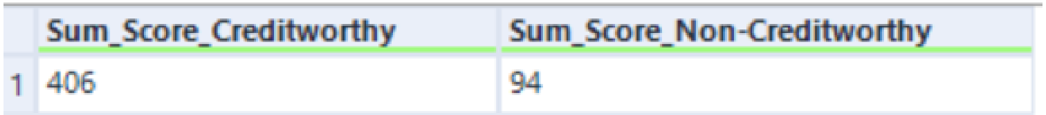
Confusion matrix of boost_p4

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Figure 11: confusion matrix report of all models

- How many individuals are creditworthy?

My model summed that 406 customers are credit worthy



	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	406	94

Figure 12: sum of credit worthy and non-credit worthy

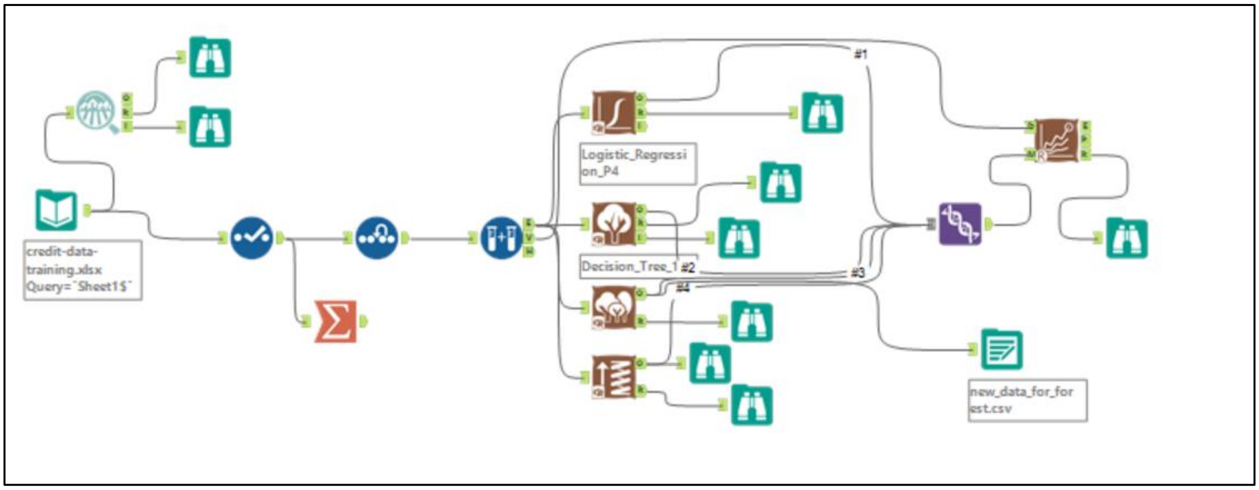


Figure 13: Alteryx workflow 1

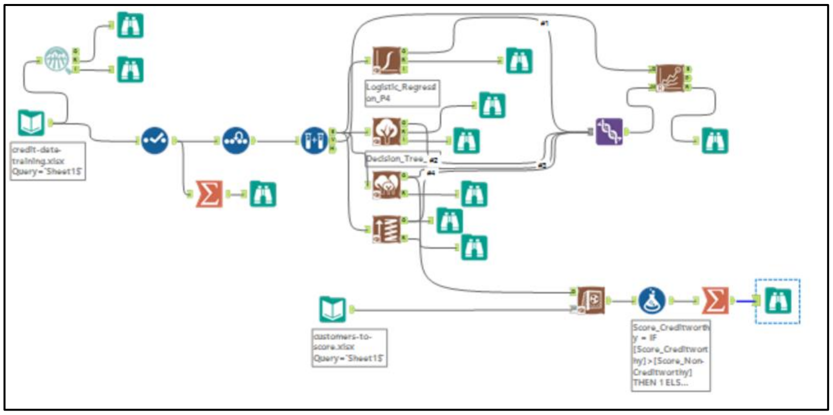

credit-data-training.xlsx
Query=`Sheet1$`
Logistic_Regression_P4
Decision_Tree_
#1
#2
#3
#4
customers-to-score.xlsx
Query=`Sheet1$`
Score_Creditworthy = IF [Score_Creditworthy] > [Score_Non-Creditworthy] THEN 1 ELS...

Figure 14: Alteryx workflow 2