Project: Creditworthiness

Business and Data Understanding

- **What decisions needs to be made?**
  Identify the best classification models to figure out the best model and provide a list of creditworthy customers to the manager.
- **What data is needed to inform those decisions?**
  We have many data of customers we could use to build our classification models to reach these decisions. These data relate with customers worthiness as the current length of employment, income, credit score, if the customer carries a credit balance from month to month, age, and customer's current savings.
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**
  Since we must determine creditworthiness of customers or not, we should use binary model.

Building the Training Set

In this set there is no high correlation between numeric fields, the highest correlation is 0.57 between Credit.Amount and Duration.of.Credit.Month.

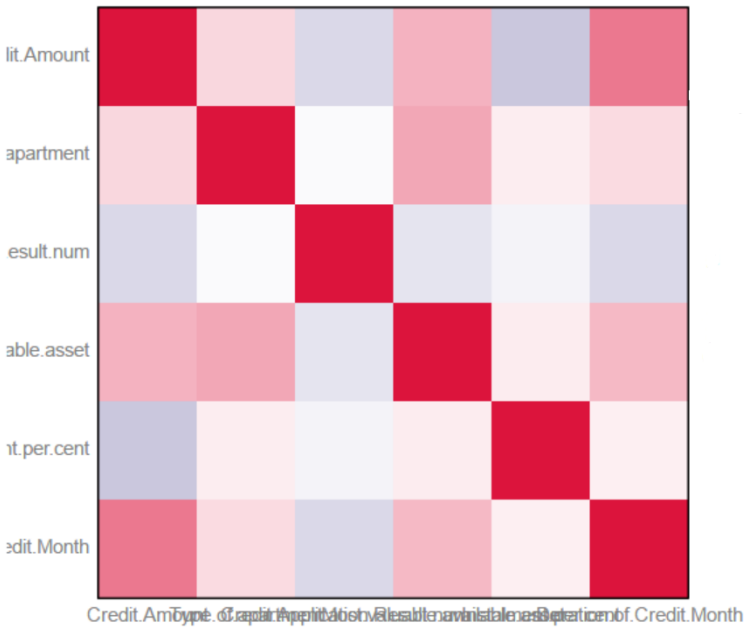

Figure 1: Correlation Matrix with ScatterPlot

In the clean-up process, there are many fields need to be removed. Because it has huge of missing data or will skew the analysis results. I removed Duration In Current Address because of 69% of data are missed. Telephone field should also be removed due to its relevance to the customer creditworthy. In addition, Guarantors, Foreign Worker and No of Dependents, Concurrent Credits and Occupation show low variability where data skewed towards one data.

Age Years has 2% missing data, the missing data imputed with the median number of age. Median age was used because it is much more representative for the data sample.

Figure 2: Fields Summery of the data

Train your Classification Models

1. **Logistic Regression:**

Credit Application Result used as the target variables, Account Balance, Purpose, Credit Amount, credit amount, instalment per cent, are the most significant variables with p-value of less than 0.05.

Report for Logistic Regression Model 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

Figure 3:Report for Logistic Regression Model Stepwise

The accuracy of logistic regression model stepwise is 76.0%. accuracy for creditworthy is 80.0% higher than non-creditworthy at 62.9%. The model is biased towards predicting customers as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Confusion matrix of Stepwise

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

Figure 4:Logistic Regression Comparison Report

2. **Decision Tree:**
   Credit Application Result used as the target variables. Account Balance, Duration of Credit Month, and Credit Amount are the most significant variables.

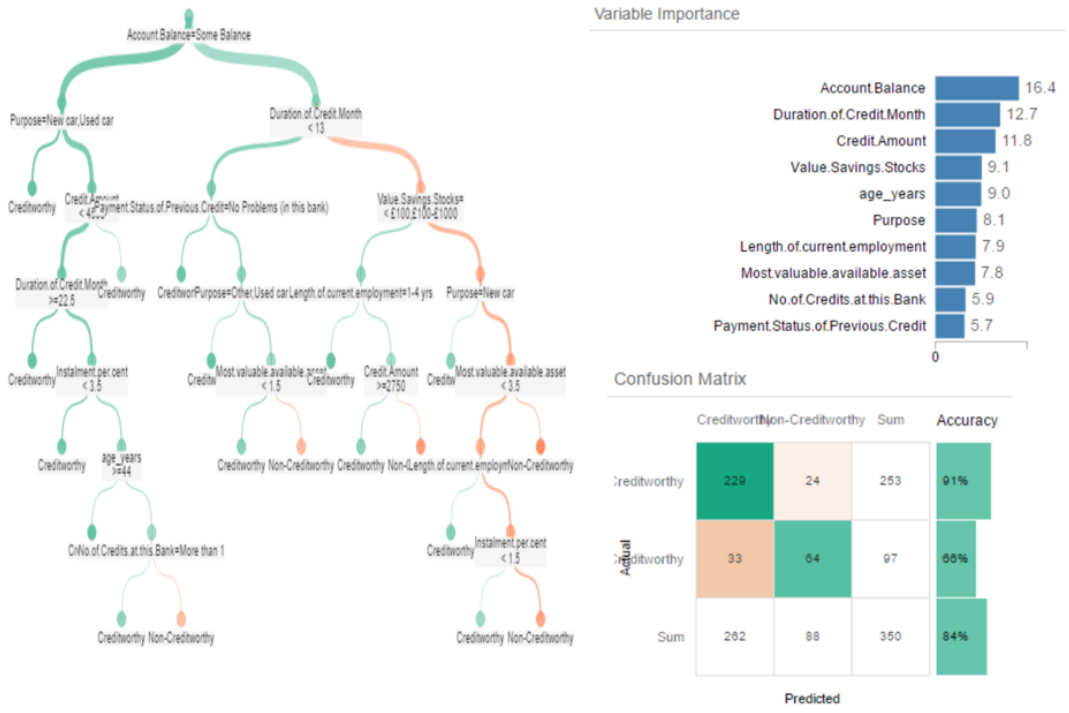

Figure 5:Report for Decision Tree

The accuracy of Decision Tree model is 67.3%. accuracy for creditworthy is 75.5% higher than non-creditworthy at 45.0%. The model is biased towards predicting customers as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree	0.6733	0.7721	0.6296	0.7545	0.4500

Confusion matrix of DecisionTree

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

Figure 6: Decision Tree Model Comparison Report

3. Forest Model:

Credit Application Result used as the target variables. Credit Amount, Age Years, Duration of Credit Month, Account Balance are the most significant variables.

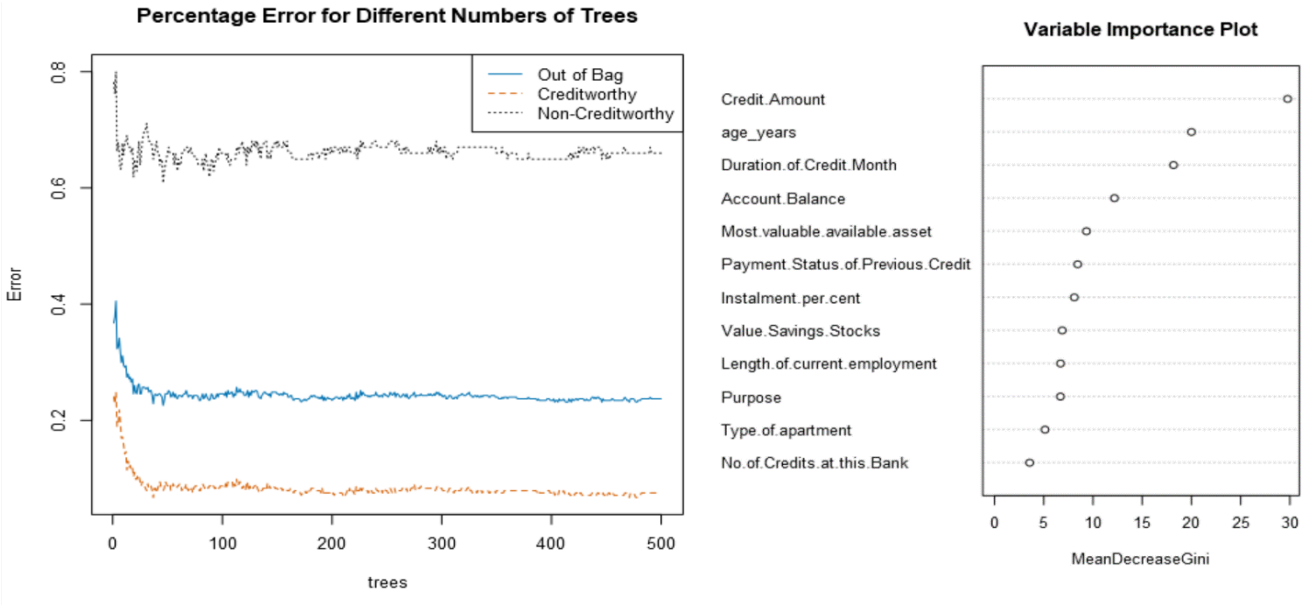

Figure 7: Percentage Error for Different Number of Trees and Variable Importance Plot

The accuracy of Forest Model is 80.0%. accuracy for creditworthy is 79.5% less than non-creditworthy at 82.6%. The difference between those accuracies is very small, so the model is almost not biased at all

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	0.8000	0.8707	0.7419	0.7953	0.8261

Confusion matrix of Forest

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

Figure 8: Forest Model Comparison Report

4. **Boosted Model:**

Credit Application Result used as the target variables. Account Balance, Credit Amount are the most significant variables.

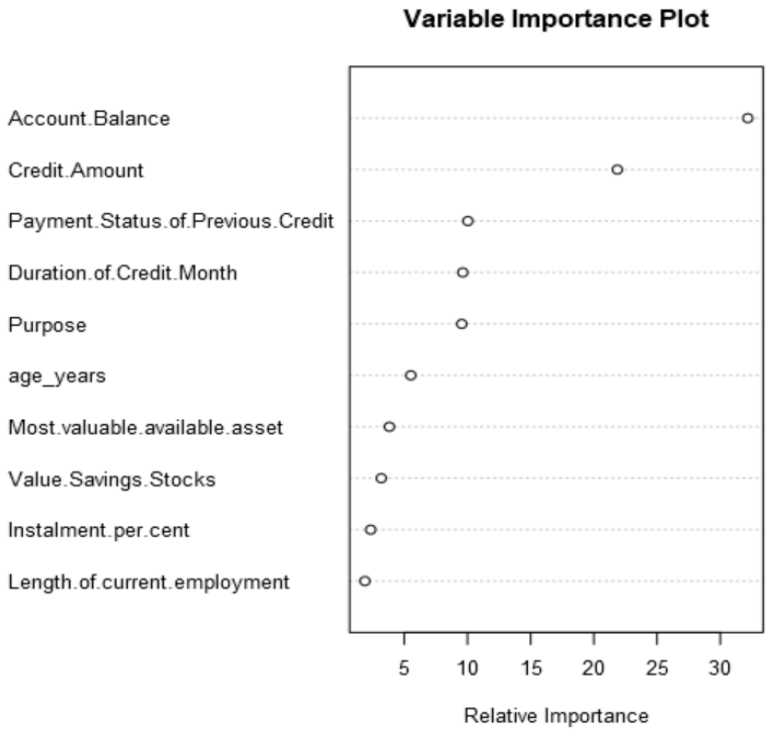

Figure 9 Variable Importance Plot Of Boosted Model

The accuracy of Boosted Model is 78.7%. accuracy for creditworthy is 78.3% less than non-creditworthy at 80.9%. The difference between those accuracies is very small, so the model is almost not biased at all

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BoostedModel	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of BoostedModel

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

Figure 10: Boosted Model Comparison Report

Writeup

Forest model was chosen, because it has the highest accuracy between validation set at 80%. There isn't any bias at the accuracies for creditworthy is 80.8% and non-creditworthy is 84%. Which are comparable.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree	0.6933	0.7890	0.6303	0.7611	0.4865
Forest	0.8133	0.8783	0.7342	0.8080	0.8400
BoostedModel	0.7867	0.8632	0.7526	0.7829	0.8095
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Figure 11:Classification models Model Comparison Report

The confusion matrix presents that Forest Model predicts best creditworthy and Non-creditworthy among all “Creditworthy” and “Non-Creditworthy” values.

Confusion matrix of BoostedModel

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

Confusion matrix of DecisionTree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	27
Predicted_Non-Creditworthy	19	18

Confusion matrix of Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	24
Predicted_Non-Creditworthy	4	21

Confusion matrix of Stepwise

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

Figure 12:all classification models Model Comparison Report

ROC curve presents the forest model true positive rate against other models:

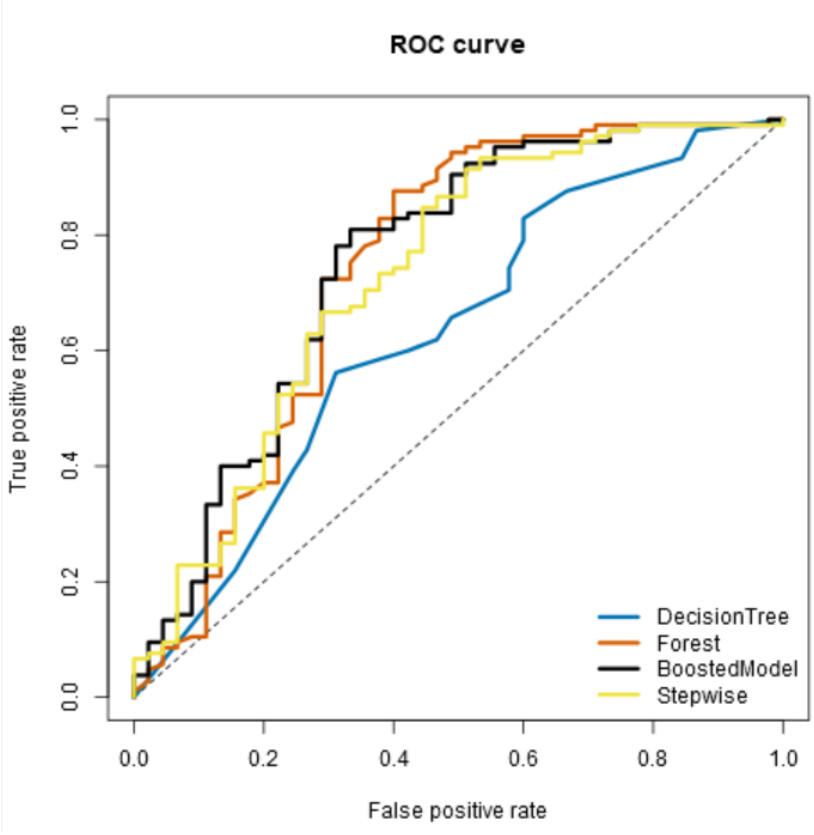

Figure 13:ROC Curve For All Classification Models

After scoring new customers, there are 409 individuals are qualifying for a loan (Creditworthy).

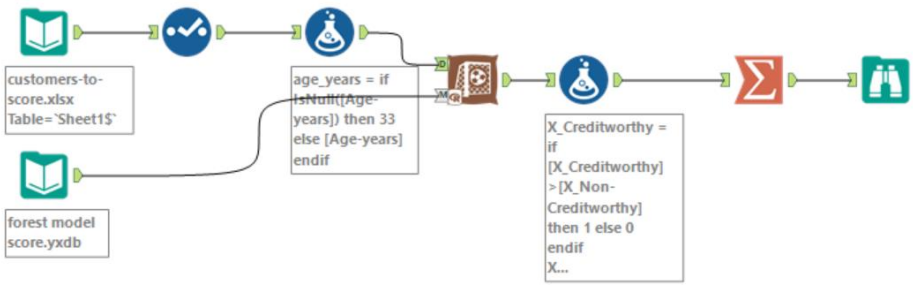

Figure 14: Customers To Score list - Workflow

Record #	Sum_X_Creditworthy	Sum_X_Non-Creditworthy
1	409	91

Figure 15:Sum of The Customers for each situation

Alteryx workflow:

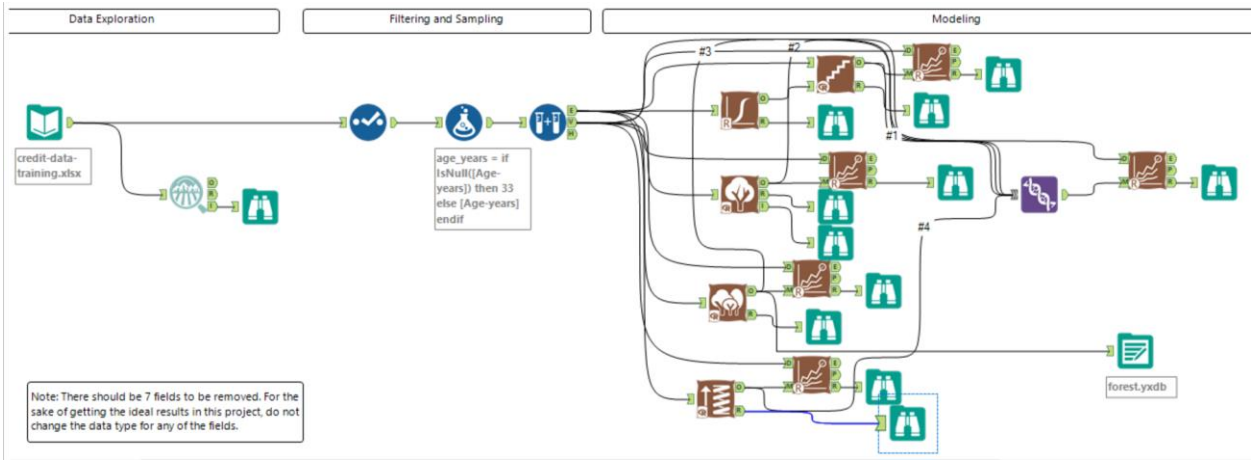

Figure 16:Alteryx workflow