Creditworthiness

Business and Data Understanding

1. What decisions needs to be made?

The objective is to identify whether customers who applied for loan are creditworthy to be extended one.

2. What data is needed to inform those decisions?

Data on past applications such as Account Balance and Credit Amount and list of customers to be processed are required in order to inform those decisions.

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

Binary classification models such as logistics regression, decision tree, forest model and boosted tree will be used to analyze and determine creditworthy customers.

Building the Training Set

An association analysis is performed on the numerical variables and there are no variables which are highly correlated with each other, i.e. a correlation of higher than 0.7.

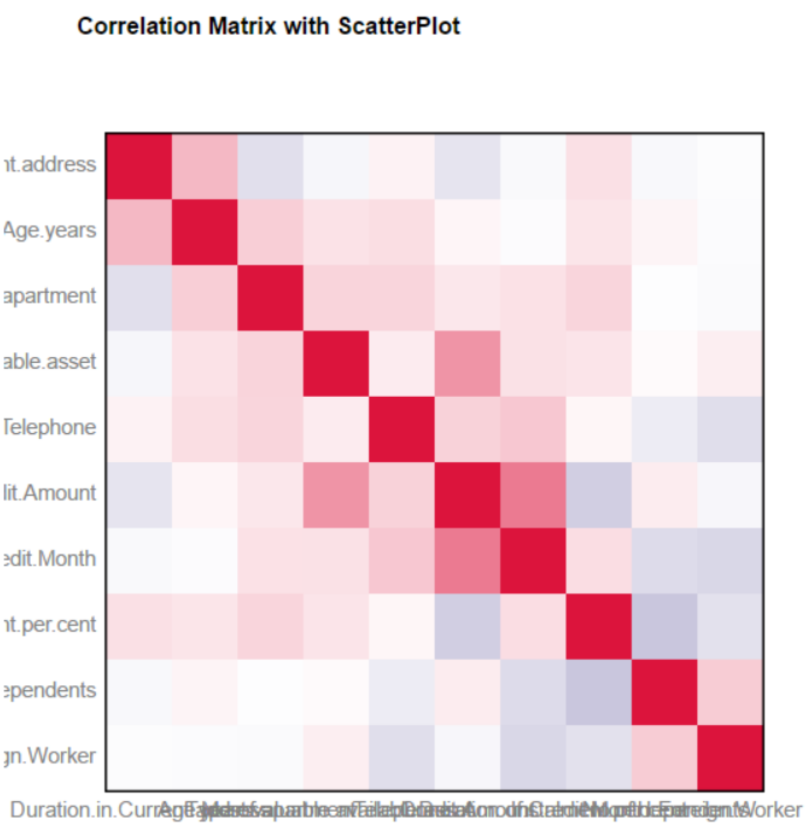

Figure 1: Correlation Matrix of variables

When summarizing all data fields, *Duration in Current Address* has 69% missing data and should be removed.

While *Age Years* has 2% missing data, it is appropriate to impute the missing data with the median age. Median age is used instead of mean as the data is skewed to the left as shown below.

In addition, *Concurrent Credits* and *Occupation* has one value while *Guarantors*, *Foreign Worker* and *No of Dependents* show low variability where more than 80% of the data skewed towards one data. These data should be removed in order not to skew our analysis results.

Telephone field should also be removed due to its irrelevancy to the customer creditworthy.

Figure 2: Field Summary of all variables

Train your Classification Models

a. Logistic Regression (Stepwise)

Using *Credit Application Result* as the target variables, *Account Balance*, *Purpose* and *Credit Amount* are the top 3 most significant variables with p-value of less than 0.05.

1 **Report for Logistic Regression Model Credit_Stepwise_Logistic**

2 *Basic Summary*

3 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(config$Link), data = the.data)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

6 Coefficients:

7

	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)

8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

Figure 3: Summary Report for Stepwise Logistic Regression Model

Overall accuracy is around 76.0% while accuracy for creditworthy is higher than non-creditworthy at 80.0% and 62.9% respectively. The model is biased towards predicting customers as non-creditworthy.

Fit and error measures

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

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Credit_Stepwise_Logistic

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

Figure 4: Model Comparison Report for Stepwise Logistic Regression Model

b. Decision Tree

Using Credit Application Result as the target variables, Account Balance, Duration of Credit Month and Credit Amount are the top 3 most important variables. The overall accuracy is 66.67%

Accuracy for creditworthy is 77.77% while accuracy for non-creditworthy is 43.59%. The model seems to be biased towards predicting customers as non-creditworthy.

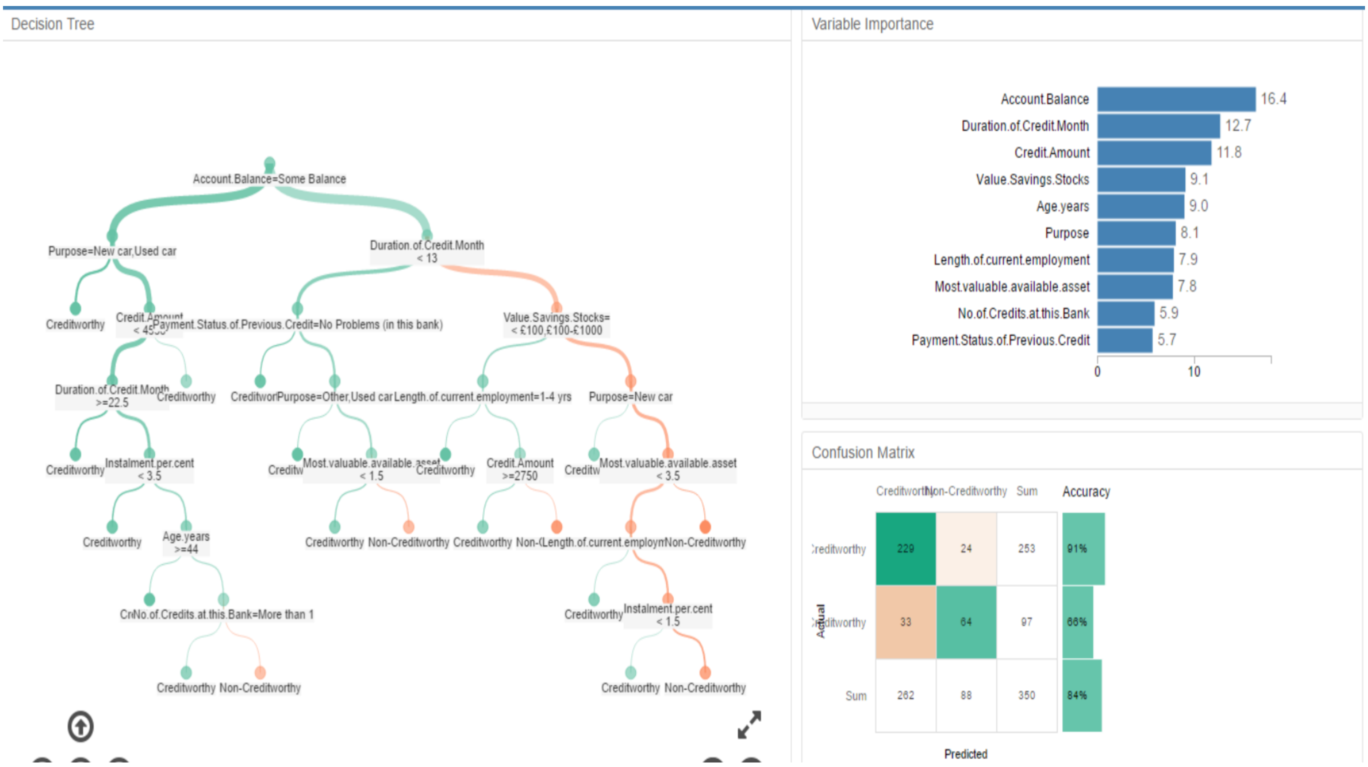

Figure 5: Decision Tree, Variable Importance and Confusion Matrix

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Credit_Decision_Tree	0.6667	0.7685	0.6272	0.7477	0.4359

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Credit_Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Figure 6: Model Comparison Report for Decision Tree

c. Forest Model

Using Credit Application Result as the target variables, Credit Amount, Age Years and Duration of Credit Month are the 3 most important variables.

Overall accuracy is 80.0%. The model isn't biased as the accuracies for creditworthy and non-creditworthy are 79.53% and 82.61% respectively, which are comparable.

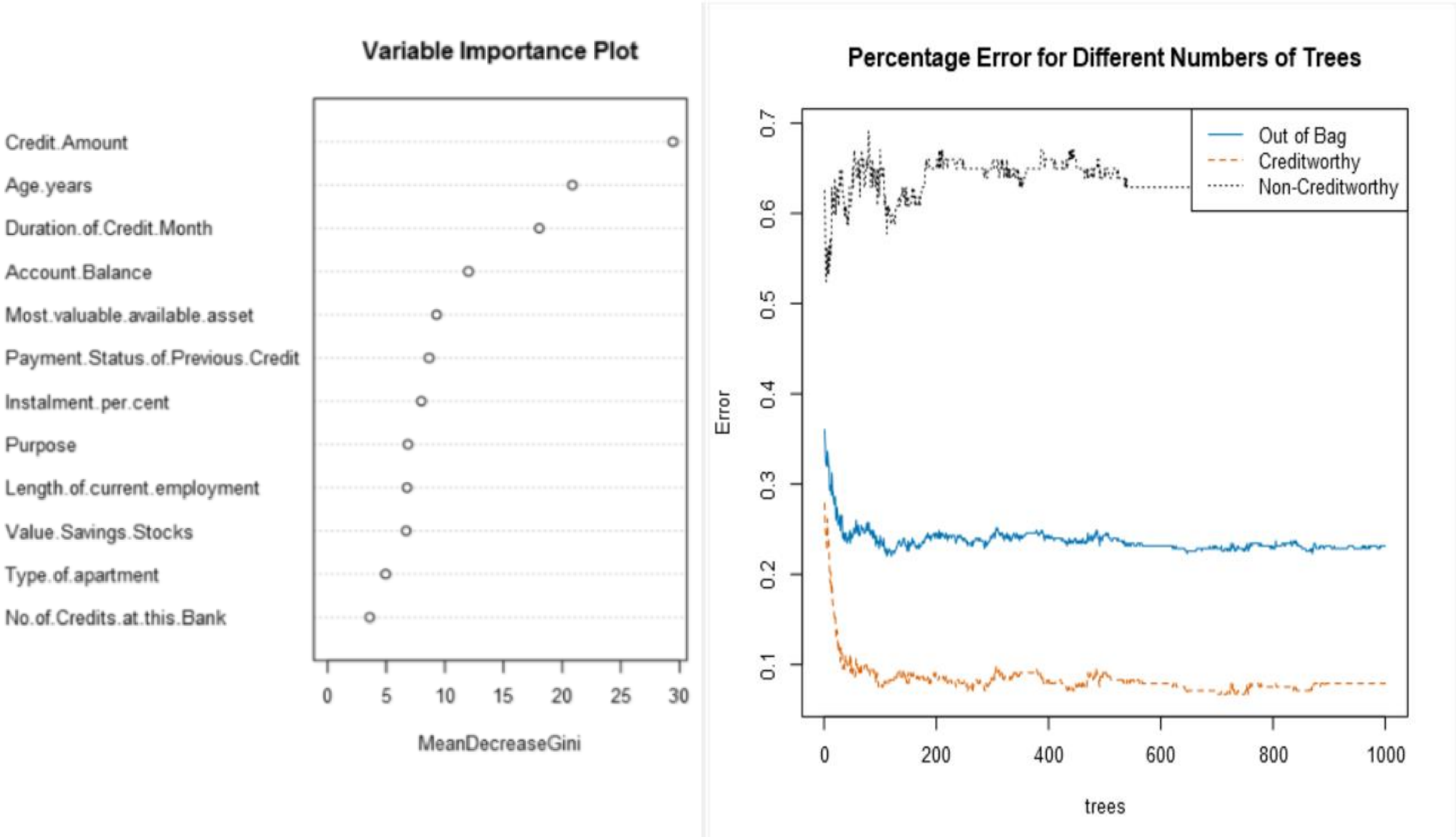

Figure 7: Variable Importance Plot and Percentage Error for different number of trees

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Credit_Forest_Model	0.8000	0.8707	0.7382	0.7953	0.8261

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Credit_Forest_Model

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

Figure 8: Model Comparison Report for Forest Model

d. Boosted Model

Account Balance and Credit Amount are the most significant variables from figure below. Overall accuracy for is 78.67%. Accuracies for creditworthy and non-creditworthy are 78.29% and 80.95% respectively which indicates a lack of bias in predicting credit-worthiness of customers.

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:

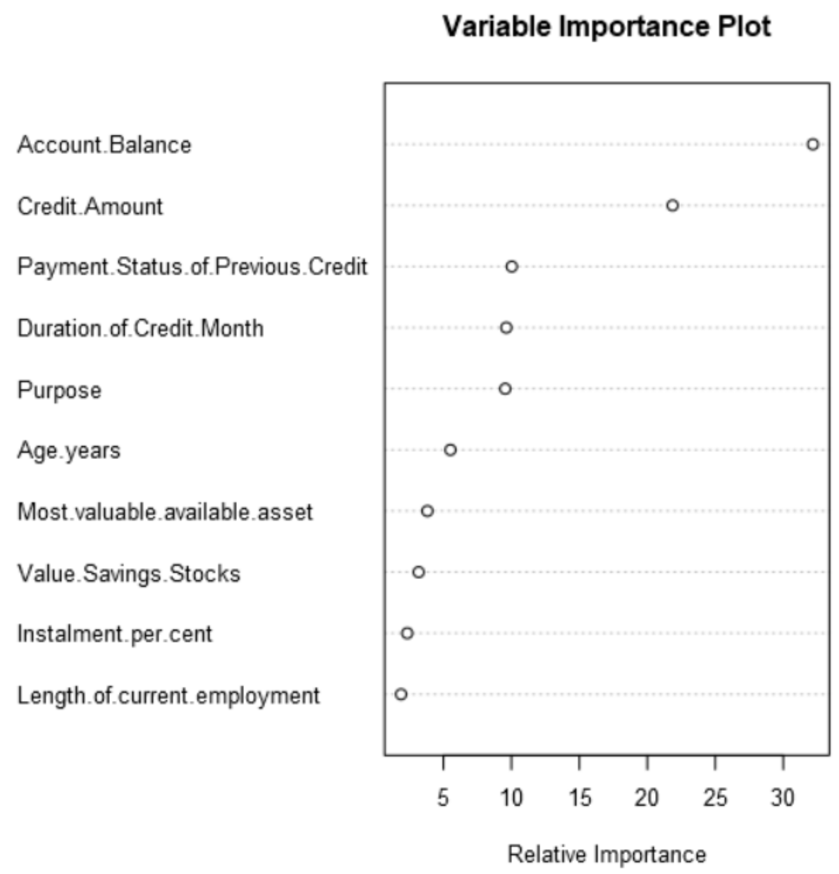

Figure 9: Variable Importance Plot for Boosted Model

Fit and error measures

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

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Credit_Boosted_Model

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

Figure 10: Model Comparison Report for Boosted Model

Write-Up

Forest model is chosen as it offers the highest accuracy at 80% against validation set. Its accuracies for creditworthy and non-creditworthy are among the highest of all.

Forest model reaches the true positive rate at the fastest rate. The accuracy difference between creditworthy and non-creditworthy are also comparable which makes it least bias towards any decisions. This is crucial in avoiding lending money to customers with high probability of defaulting while ensuring opportunities are not overlooked by not loaning to creditworthy customers.

There are **414 creditworthy customers** using forest models to score new customers.

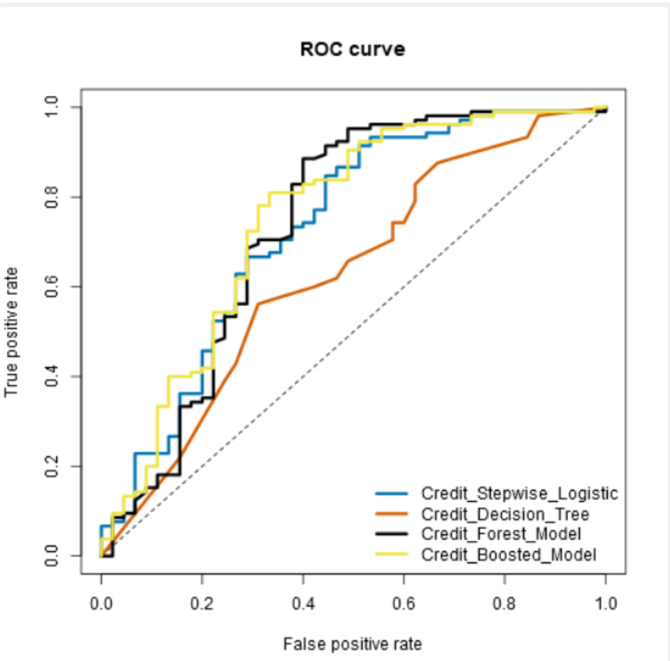

Figure 11: ROC curve for all 4 classification models

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Credit_Stepwise_Logistic	0.7600	0.8364	0.7306	0.8000	0.6286
Credit_Decision_Tree	0.6667	0.7685	0.6272	0.7477	0.4359
Credit_Forest_Model	0.8000	0.8707	0.7382	0.7953	0.8261
Credit_Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

Confusion matrix of Credit_Boosted_Model

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

Confusion matrix of Credit_Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of Credit_Forest_Model

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

Confusion matrix of Credit_Stepwise_Logistic

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

Figure 12: Model Comparison Report for all 4 classification Models

Alteryx Flow