

Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

This analysis would determine if the customers are creditworthy enough to be allowed a loan or not. The data needed is past account balances and credit amount and customers. Binary classification models need to be used to analyze this business problem.

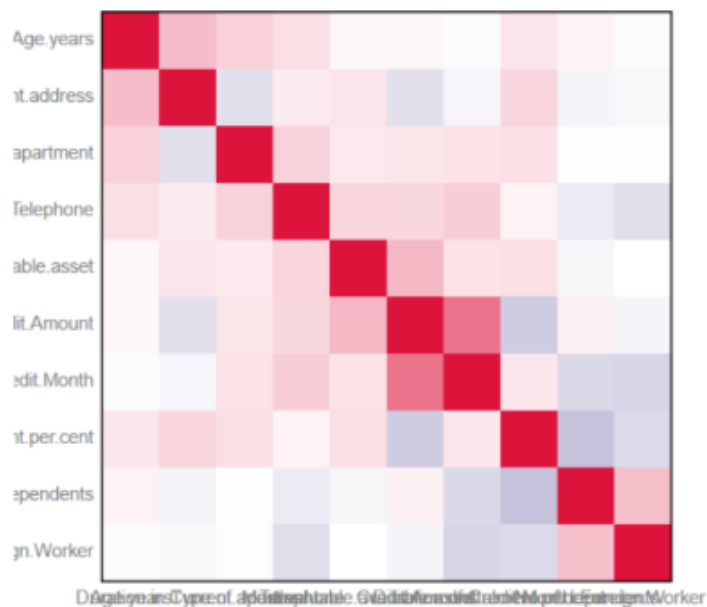
Step 2: Building the Training Set.

After the association analysis, no highly correlated numerical variables are present. 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.

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 creditworthiness.

Correlation Matrix with ScatterPlot



Step 3: Train your Classification Models

1. Logistic Regression

Using Credit Application as the target variable while Account Balance, Purpose and Credit Amount significant variables with p-value of less than 0.05. 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.

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 .

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Logistic	0.7600	0.8364	0.7306	0.8000	0.6286
Confusion matrix of Stepwise_Logistic					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	92	23			
Predicted_Non-Creditworthy	13	22			

2. Decision Tree

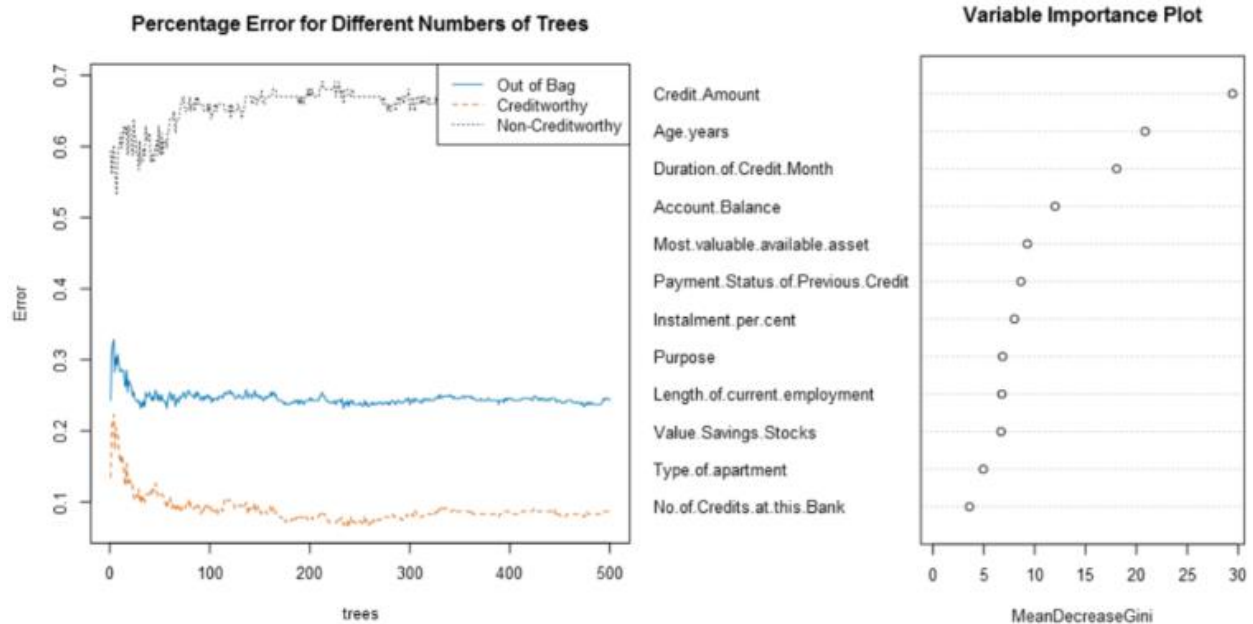
Using Credit Application Result as the target variables, Account Balance, Value Savings Stocks and Duration of Credit Month are the top 3 most important variables. The overall accuracy is 74.7%.

Accuracy for creditworthy is 79.1% while accuracy for non-creditworthy is 60.0%. The model seems to be biased towards predicting customers as non-creditworthy.



3. 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 79.1% and 85.7% respectively, which are comparable.



Model Comparison Report						
Fit and error measures						
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy	
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571	
Confusion matrix of FM_Credit						
			Actual_Creditworthy	Actual_Non-Creditworthy		
Predicted_Creditworthy			102	27		
Predicted_Non-Creditworthy			3	18		

4. Boosted Model

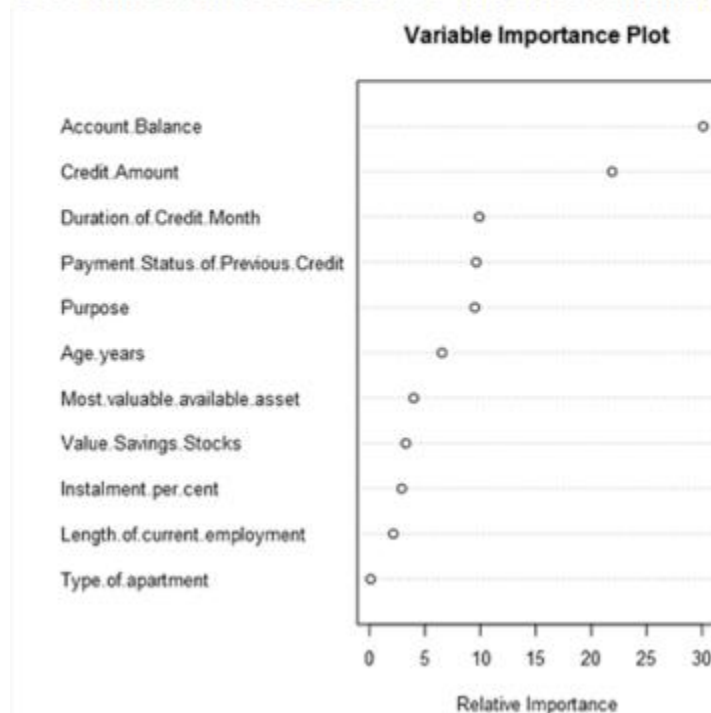
Account Balance and Credit Amount are the most significant variables from figure 10. Overall accuracy for is 76.7%. Accuracies for creditworthy and non-creditworthy are 76.7% and 78.3% respectively which indicates a lack of bias in predicting credit-worthiness of customers.

Basic Summary:

Loss function distribution: Bernoulli

Total number of trees used: 4000

Best number of trees based on 5-fold cross validation: 2377



Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826
Confusion matrix of BM_Credit					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	100	27			
Predicted_Non-Creditworthy	5	18			

Step 4: Writeup

There are 408 creditworthy customers using forest models to score new customers. 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.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826
Stepwise_Logistic	0.7600	0.8364	0.7306	0.8000	0.6286
Confusion matrix of BM_Credit					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	100		27		
Predicted_Non-Creditworthy	5		18		
Confusion matrix of DT_Credit					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	91		24		
Predicted_Non-Creditworthy	14		21		
Confusion matrix of FM_Credit					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	102		27		
Predicted_Non-Creditworthy	3		18		
Confusion matrix of Stepwise_Logistic					
	Actual_Creditworthy	Actual_Non-Creditworthy			
Predicted_Creditworthy	92		23		
Predicted_Non-Creditworthy	13		22		

