Project: Creditworthiness

Step 1: Business and Data Understanding

The small bank I work for has had an influx of loan applications and the manager wants the loans to be processed quickly in a new way.

Key Decisions:

- What decisions needs to be made?

Determine whether or not the new loan applicants are creditworthy.

- What data is needed to inform those decisions?

Data on all past loan applications from c*redit-data-training.xlsx*.

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

We need to predict the outcome of a loan applications, from a data-rich spreadsheet, with classification data. Approving the loan or not is binary, therefore I will use a binary classification model.

Step 2: Building the Training Set

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.

Full pearson correlation was run.

There appear to be no data sets highly correlated within the data fields.

- Are there any missing data for each of the data fields?

Age-years and Duration-in-Current address are both missing data.

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asset	Type.of.apartment
Duration.of.Credit.Month	1.0000000	0.5906171	0.1040048	-0.0506493	0.1195555	0.1201070
Credit.Amount	0.5906171	1.0000000	-0.2653537	-0.1580691	0.3012233	0.1069607
Instalment.per.cent	0.1040048	-0.2653537	1.0000000	0.1733930	0.1341344	0.1369001
Duration.in.Current.address	-0.0506493	-0.1580691	0.1733930	1.0000000	0.1092968	-0.1575495
Most.valuable.available.asset	0.1195555	0.3012233	0.1341344	0.1092968	1.0000000	0.0938777
Type.of.apartment	0.1201070	0.1069607	0.1369001	-0.1575495	0.0938777	1.0000000
No.of.dependents	-0.1959091	0.0638629	-0.3127847	-0.0566456	-0.0479319	0.0039290
Telephone	0.2103393	0.1715142	0.0526591	0.0849249	0.1788326	0.1905344
Foreign.Worker	-0.2184723	-0.0563574	-0.1898275	-0.0365874	-0.0013900	-0.0087732
Age_years	-0.0203686	0.0349014	0.1036610	0.2793304	0.0333870	0.1942070
	No.of.dependents	Telephone	Foreign.Worker	Age_years		
Duration.of.Credit.Month	-0.1959091	0.2103393	-0.2184723	-0.0203686		
Credit.Amount	0.0638629	0.1715142	-0.0563574	0.0349014		
Instalment.per.cent	-0.3127847	0.0526591	-0.1898275	0.1036610		
Duration.in.Current.address	-0.0566456	0.0849249	-0.0365874	0.2793304		
Most.valuable.available.asset	-0.0479319	0.1788326	-0.0013900	0.0333870		
Type.of.apartment	0.0039290	0.1905344	-0.0087732	0.1942070		
No.of.dependents	1.0000000	-0.1055013	0.2699280	0.0490067		
Telephone	-0.1055013	1.0000000	-0.1718538	0.1334909		
Foreign.Worker	0.2699280	-0.1718538	1.0000000	-0.0214109		
Age_years	0.0490067	0.1334909	-0.0214109	1.0000000		

For age-years only 2% of the data is missing so I will impute the null values. For Duration-in-Current address 69% of the data is missing so I will exclude this field.

- Are there only a few values in a subset of your data field?

Concurrent credits and Occupation has only one value for the entire field. With such low variability the field will be removed.

Foreign worker, Guarantors and No-of-dependents have only 2 values all dominated by one main value. With low variability these fields will be removed.

Telephone has only 2 values and a low score across the correlation matrix so will be removed.

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.

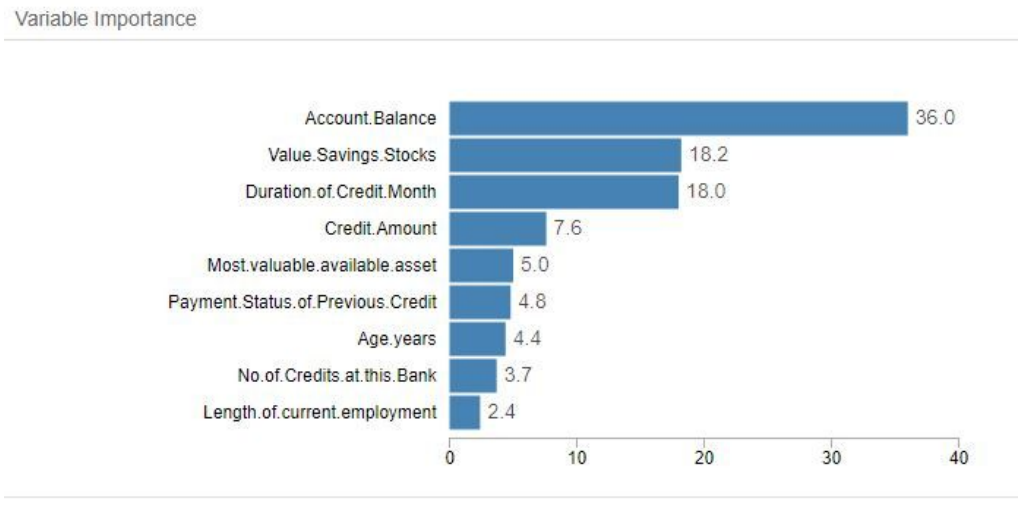
For the Logistic Regression Model the coefficients are:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315	**
Age.years	-0.015092	1.539e-02	-0.9809	0.32666	
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163	
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946	*
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06	***
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412	
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943	
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182	*
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518	**
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848	
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679	.
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966	**
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213	
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046	
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645	*
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949	
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184	
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644	*

So the significant predictor variables are:

- Length of current employment
- Account Balance Some Balance
- Payment status of previous credit some problems
- Purpose new car
- Credit amount
- Most valuable available asset
- Instalment per cent

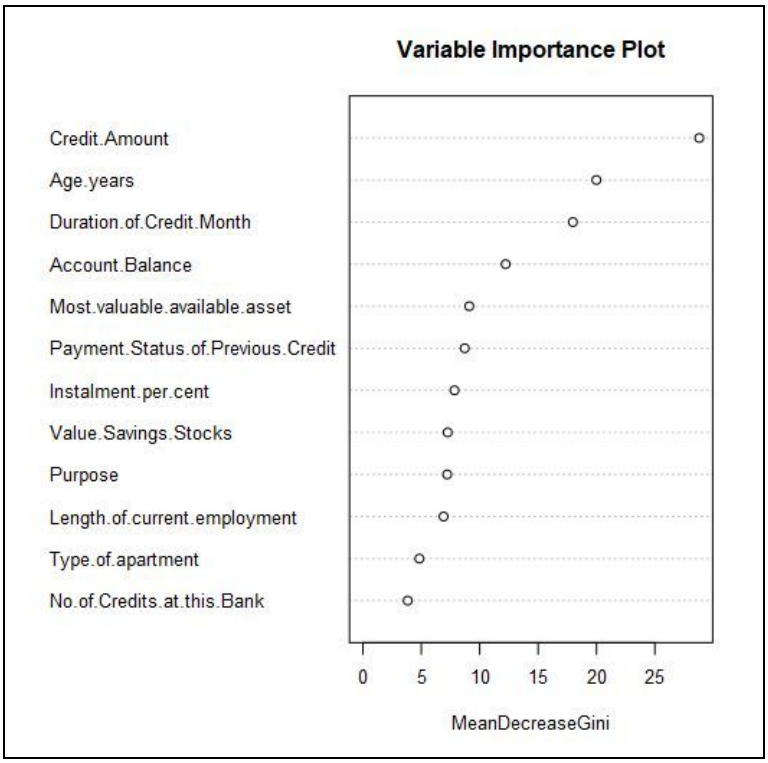
For the decision tree:



The significant predictor variables are:

- Account Balance
- Value Savings Stocks
- Duration of Credit Month

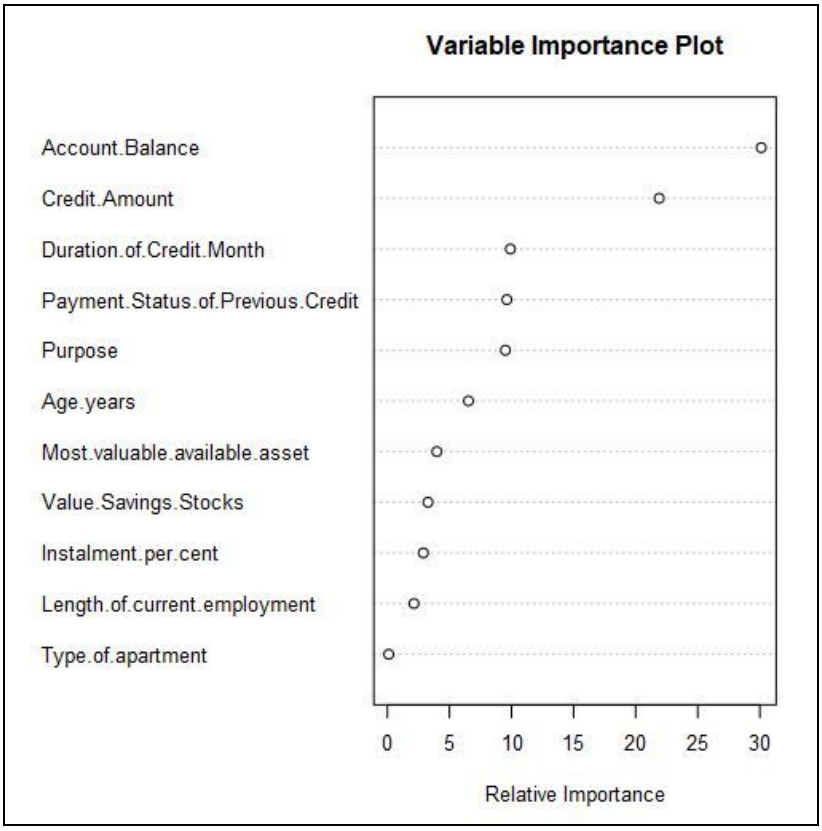
For the Forest Model:



The significant predictor variables are:

- Credit Amount
- Duration of Credit Months
- Age years

For the Boosted Model:



The significant predictor variables are:

- Account Balance
- Credit Amount

- 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Log_creditworthy	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthy	0.7467	0.8273	0.7054	0.8667	0.4667
DF_creditworthy	0.8067	0.8755	0.7455	0.9714	0.4222
Boost_creditworthy	0.7867	0.8632	0.7524	0.9619	0.3778

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as ***recall***.

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

F1: F1 score, 2 * precision * recall / (precision + recall). The ***precision*** measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boost_creditworthy

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

Confusion matrix of DF_creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of DT_creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Log_creditworthy

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

Performance Diagnostic Plots

The model with the highest accuracy is the random forest model with 0.8067.

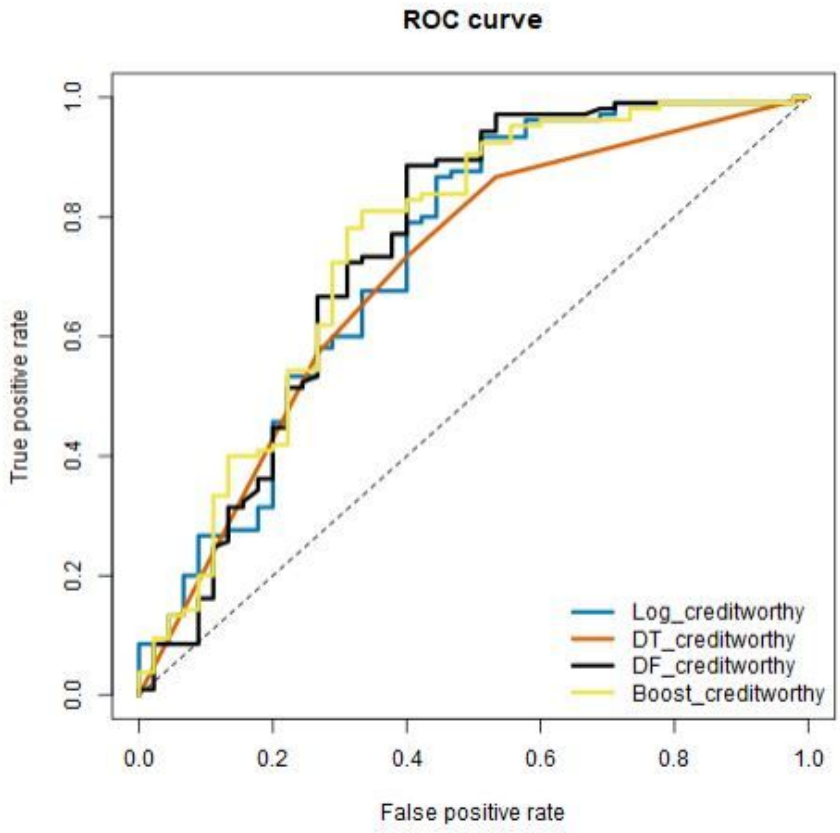
There are more creditworthy data applicants than non-creditworthy applicants in our data set. The model appears to predict creditworthy applicants better than non-creditworthy applicants.

Step 4: Writeup

- Which model did you choose to use?

  The Random Forest Model should be chosen due to its highest accuracy against the other validation models. It has an accuracy of 0.9714 for creditworthy candidates and 0.4222 for non-creditworthy candidates.

  The ROC graph is below and has a AUC of 0.8755 the best in the validation comparisons.



  The model is biased to creditworthy applicants over non-creditworthy applicants due to the accuracy rates

- How many individuals are creditworthy?

  409 individuals are creditworthy.