Project: Creditworthiness

Sarah Alomran

Step 1: Business and Data Understanding

Key Decisions:

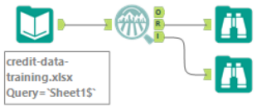
- What decisions needs to be made?
  Determine whether the new customers who applied for the loan are creditworthy.

- What data is needed to inform those decisions?
  -Data on past applications.
  -New customers list.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Since we are trying to determine if the loan application got either accepted or not, then this model consider to be a binary classification model.

Step 2: Building the Training Set

To build the training set, I used a Field Summary tool and browsing the report output and interactive output.



In result I got table below and from it i identified the missing data and low variability and in result I will do the following:

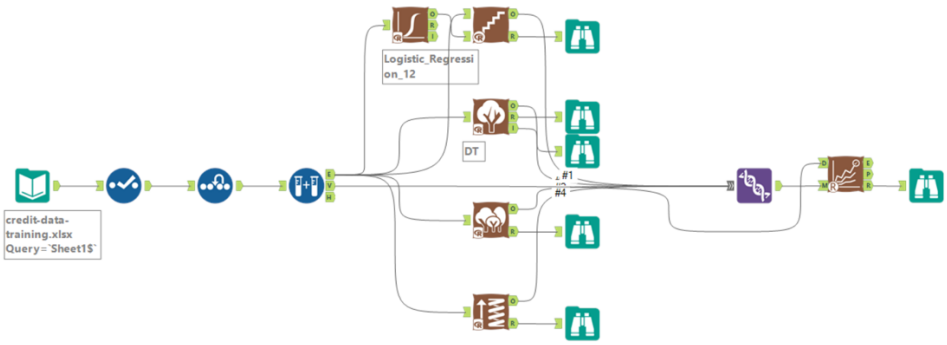
1-Impute Age-Years with median.

2-Remove Duration-in-Current-address (due to missing data), Occupation and Concurrent (uniform), Guarantors, Foreign-Worker, No-of-dependents, and Telephone (low variability).

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Credit-Amount		0.0%	464	276.000	3,199.980	2,236.500	18,424.000	2,831.387	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Duration-of-Credit-Month		0.0%	30	4.000	21.434	18.000	60.000	12.307	
Foreign-Worker		0.0%	2	1.000	1.038	1.000	2.000	0.191	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Instalment-per-cent		0.0%	4	1.000	3.010	3.000	4.000	1.114	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Most-valuable-available-asset		0.0%	4	1.000	2.360	3.000	4.000	1.064	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
No-of-dependents		0.0%	2	1.000	1.146	1.000	2.000	0.353	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Telephone		0.0%	2	1.000	1.400	1.000	2.000	0.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Type-of-apartment		0.0%	3	1.000	1.928	2.000	3.000	0.540	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Step 3: Train your Classification Models

First, create Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% should be reserved for Validation and set the Random Seed to 1. Then create all of the following models: Logistic Regression and Logistic Stepwise, Decision Tree, Forest Model, Boosted Model. Target variable is Credit-Application-Result, and everything else in the cleaned data sheet is predictor variables.



For each model created I will answer the following equations:

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

- 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?

Logistic Stepwise:

Based on the report below, the top 3 significant predictive variables are Account-Balance, Purpose and Credit-Amount with p-value of less than 0.05.

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

(Dispersion parameter for binomial taken to be 1)

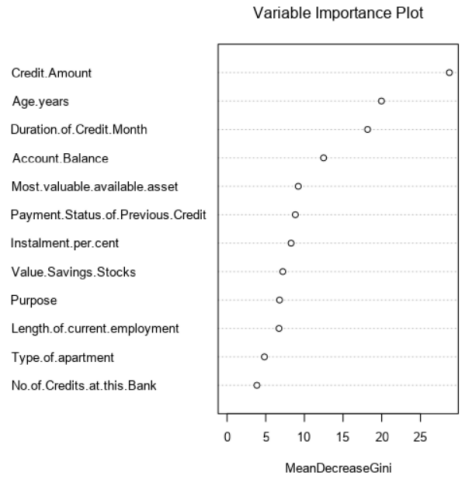
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Decision Tree:

Based on the tree below, the top 3 significant predictive variables are Account-Balance, Duration-of-Credit-Month and Value-Savings-Stocks.

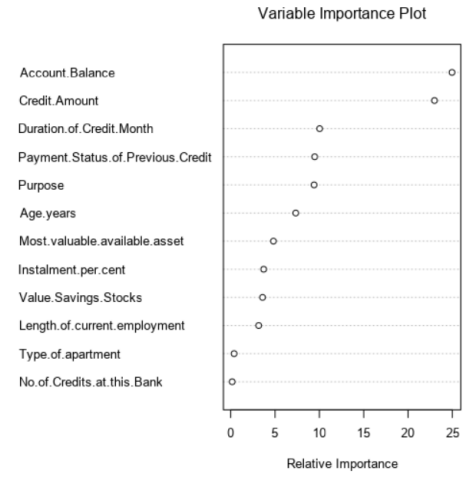
Forest Model:

Based on the plot below, the top 3 significant predictive variables are Credit-Amount, Age and Duration-of-Credit-Month.



Boosted Model:

Based on the plot below, Account-Balance and Credit-Amount are more important.



Step 4: Writeup

As shown in the Model Comparison report, forest model offers the highest accuracy at 79% against validation set. Also, its accuracies for creditworthy and non-creditworthy are among the highest of all.
There are 408 creditworthy customers using the forest models to score new customers.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW	0.7600	0.8364	0.7306	0.8762	0.4889
DT	0.7467	0.8304	0.7035	0.8857	0.4222
FM	0.7933	0.8681	0.7368	0.9714	0.3778
BM	0.7867	0.8632	0.7515	0.9619	0.3778

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of FM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of SW

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

Confusion matrix of Test_Model

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

Performance Diagnostic Plots

ROC Curve: