Project: Creditworthiness

Step 1: Business and Data Understanding

A bank needs a model to easely classify new customers that are appling for a loan. Using a data set containing the last two years's customers we can create a model that will predict if a new customer will be creditworthy or not.

Key Decisions:

The bank needs to decide if a new customer is creditworthy or not. This problem could be solved using the data of their customers to create a model for the prediction of the “worthyness” of new clients. We need a Binary model to tell us if a customer is creditworthy or if it is not.

Step 2: Building the Training Set

In the cleanup process I have removed the”Duration-in-current-address” and “Concurrent-Credits” fields because the first one had to many missing values and the second one had only one value for the whole field. I have also replaced null values in the “Age-years” field with the median value of the field (33).

Step 3: Train your Classification Models

Logistic Regression model:
These are the predictors I used and their P-values. I have removed the less significant variables for this model.

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.6561796	1.051e+00	-1.5756	0.11513	
Account.BalanceSome Balance	-1.6879156	3.150e-01	-5.3583	8.40e-08	***
PurposeNew car	-1.7384447	6.147e-01	-2.8281	0.00468	**
PurposeOther	-0.5073510	8.093e-01	-0.6269	0.53074	
PurposeUsed car	-0.7730537	4.000e-01	-1.9324	0.05331	.
Credit.Amount	0.0001849	5.729e-05	3.2277	0.00125	**
Length.of.current.employment4-7 yrs	0.4018327	4.717e-01	0.8520	0.39424	
Length.of.current.employment< 1yr	0.7470200	3.786e-01	1.9732	0.04848	*
Instalment.per.cent	0.3036306	1.363e-01	2.2269	0.02595	*
GuarantorsYes	0.2197896	4.447e-01	0.4942	0.62116	
Most.valuable.available.asset	0.3405007	1.503e-01	2.2659	0.02346	*
Age.years	-0.0149895	1.481e-02	-1.0123	0.3114	
Type.of.apartment	-0.2629291	2.907e-01	-0.9045	0.36571	
No.of.Credits.at.this.BankMore than 1	0.1905012	2.978e-01	0.6397	0.52238	
Foreign.Worker	-0.3032746	6.757e-01	-0.4488	0.65357	

The overall accuracy of thi model is 0.72 and this is the confusion matrix of the logistic regression model.

Confusion matrix of LR

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	89	26
Predicted_Non-Creditworthy	16	19

Decision Tree model :

These are the variables used for the Tree.

Model Summary

Variables actually used in tree construction:

[1] Account.Balance Age.years

[3] Credit.Amount Duration.of.Credit.Month

[5] Instalment.per.cent Length.of.current.employment

[7] Most.valuable.available.asset No.of.Credits.at.this.Bank

[9] Payment.Status.of.Previous.Credit Purpose

[11] Value.Savings.Stocks

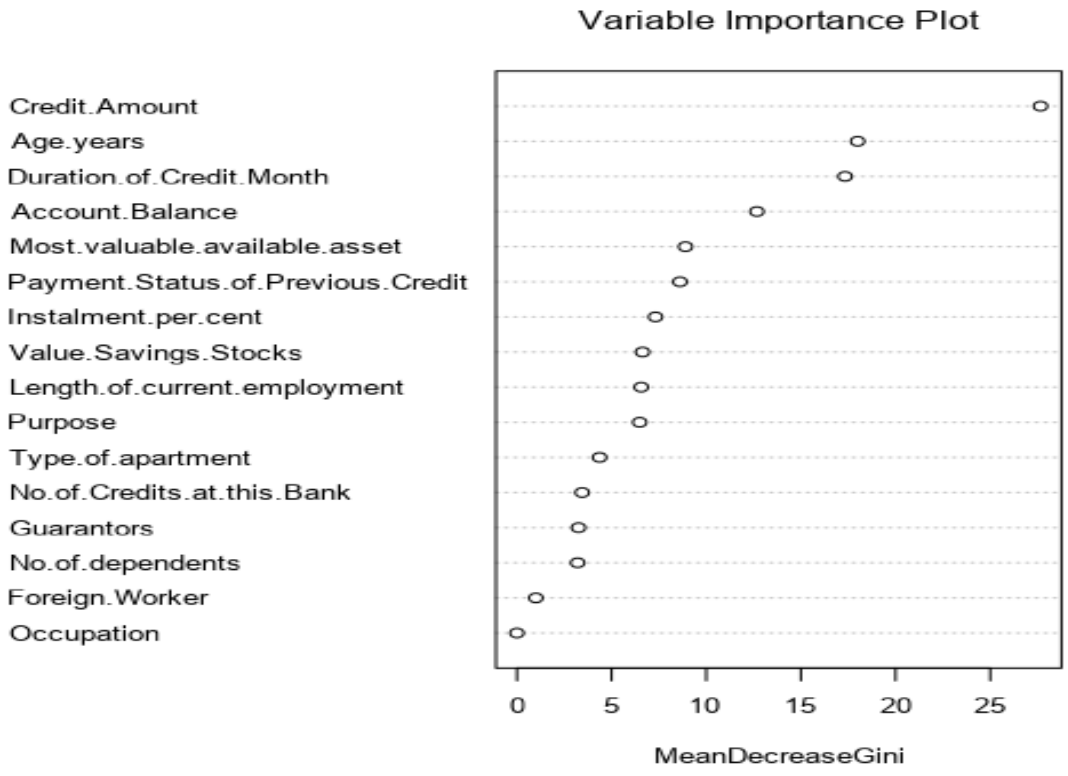
The overall accuracy for this model is : 0.67 and this is the confusion matrix of the decision tree model.

Confusion matrix of DT

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

Random Forest model:

this is the variable importance plot for the Random Forest model, the three most important variables are “credit.Amount”, “Age.Years” and “Duration.of.Credit.Month”

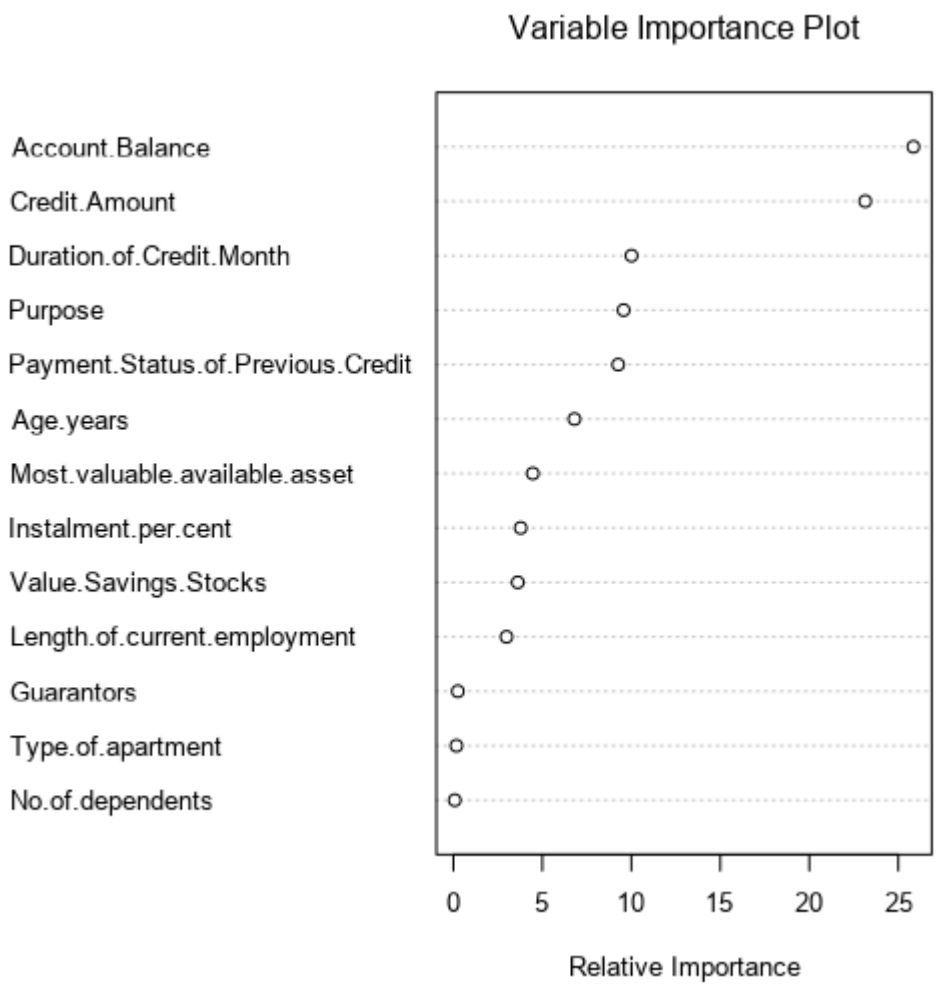


The overall accuracy of this model is: 0.79, and thi is the confusion matrix of the Random Forest model

Confusion matrix of RF

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Boosted model:
This is the Varible Importance Plot for the Boosted model: the three most important variables are “credit.Amount”, “Age.Years” and “Duration.of.Credit.Month”



The overall accuracy of this model is: 0.78. This is the confusion matrix of the Boosted model.

Confusion matrix of BO

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

Step 4: Writeup

Answer these questions:

I have chosen the Random Forest model because it's overall accuracy is the highest with a value of 0.79. This model also has the biggest Area Under the Curve: 0.77. The confusion matrix shows that the Random Forest model classifies 101 customers as “Creditworthy” over 105 Creditworthy customers and it classifies 18 customers as “non-Creditworthy” over the total of 45 non-Creditworthy customers.
The accuracy of the classification it is way better for the “Creditworthy” class but even in the other 3 model this accuracy was still pretty low.

The numer of individuals that are classified as “Creditworthy by this model is 416.