

## 1. Business and Data Understanding

As the business analyst of a small bank we should improve the process of checking if a customer is creditworthy or not. For approx. 500 loan applications in one week this is not possible to do by hand. With the right model we can build a solid, stable model for predict if a customer should get a credit.

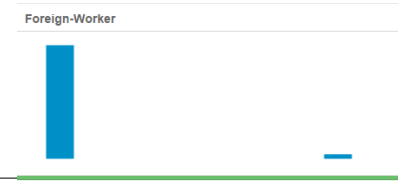
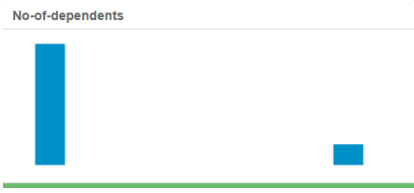
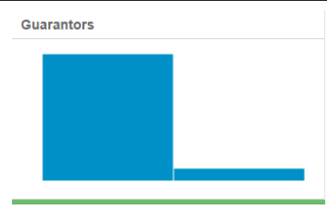
We need data like: The age of the customer, has he already a credit, Is he employed or not (how long is he employed), for what needs the customer the credit, had he problems with paying back a credit, has he an account at our bank (and if so is the account in balance), what is the amount of credit, has he some savings and finally how long will the credit run.

It's a binary problem: Will he get a credit or not.

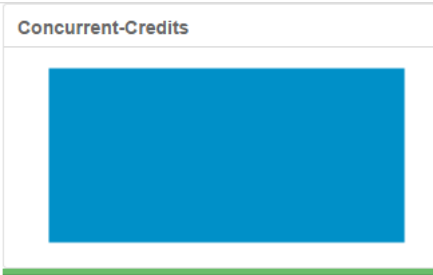
## 2. Building the Training Set

First of all we had to clean the data.

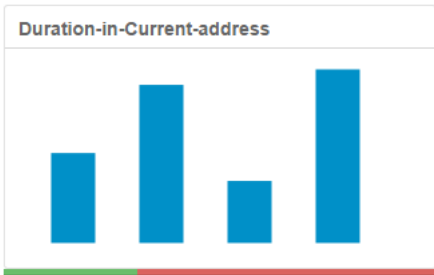
### 2.1 Low Variability

Foreign-Worker	
No-of-dependents	
Guarantors	

## 2.2 No Variability

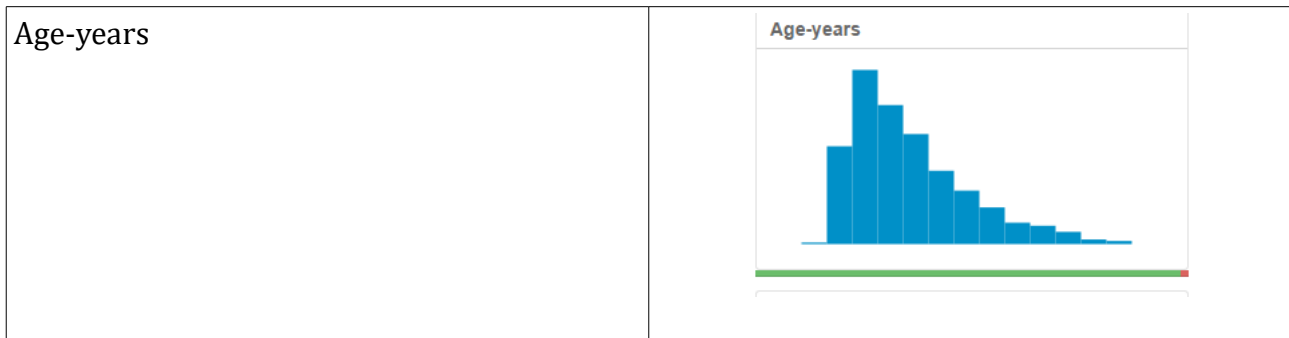
Concurrent-Credits																							
Occupation	<table><thead><tr><th>k</th><th>Occupation</th></tr></thead><tbody><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>1</td></tr></tbody></table>	k	Occupation	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
k	Occupation																						
1	1																						
1	1																						
1	2																						
1	1																						
1	1																						
1	1																						
1	1																						
1	1																						
1	1																						
1	1																						

## 2.3 Missing Data

Duration-in-Current-address	 <p>As we can see nearly 70% of the data is missing.</p>
-----------------------------	--

## 2.4 The Age of the customer

We need the age of the customer for predicting if the customer is getting the credit or not. But there are some data missing (approx. 2% of the data is missing).



So we could either use the average age of the customer (36 years) but if we look at our data it make more sense to use the median age (33 years). Otherwise the age sum will be to high.

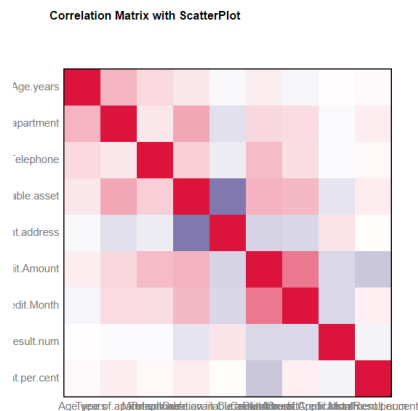
## 2.5. Correlation Analysis

**Pearson Correlation Analysis**

*Focused Analysis on Field Credit.Application.Result.num*

	Association Measure	p-value
Duration.of.Credit.Month	-0.2025036	5.0151e-06 ***
Credit.Amount	-0.2019458	5.3311e-06 ***
Most.valuable.available.asset	-0.1413324	1.5334e-03 **
Duration.in.Current.address	0.1145110	1.0390e-02 *
Instalment.per.cent	-0.0621068	1.6556e-01
Telephone	-0.0289707	5.1807e-01
Type.of.apartment	-0.0265155	5.5417e-01
Age.years	0.0081157	8.5635e-01

*Full Correlation Matrix*



By using the correlation matrix we can see if there is a high correlation and this could led to problems in our models. The telephone number has a small p-value so I made the decision to remove the telephone number because this is something without a value for our credit approval process. The type of apartment has also not such a big p-value but we can always sell his apartment to get back the money.

## 3. Train your Classification Models

### 3.1 Logistic Model

The three most significant variable are: Account\_Balance, Credit\_Amount, PurposeNew car

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
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 *
Account_BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
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 .
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 **
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 *

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

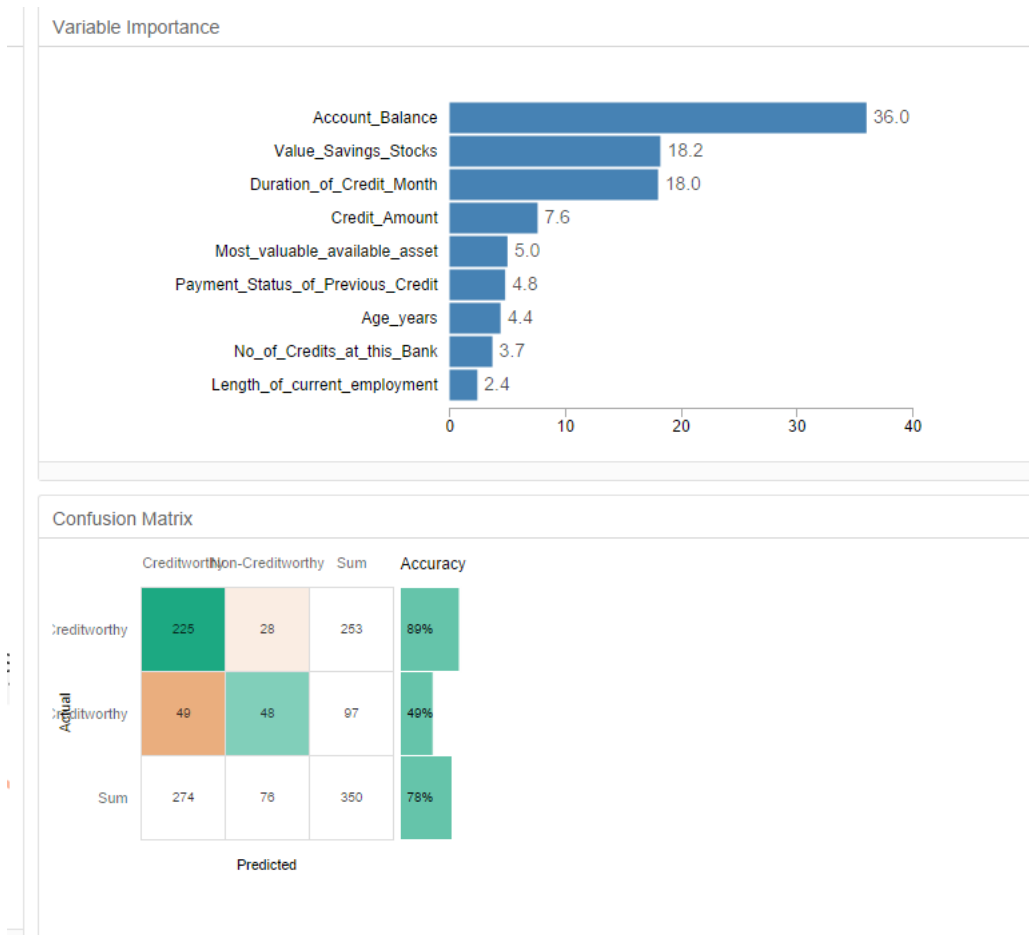
[Dispersion parameter for binomial taken to be 1]

Confusion matrix of Log_Stepwise		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

The overall percent accuracy is: 76 %

Regarding the bias topic: The model accuracy for predicting the creditworthy of: 80 % and the accuracy for predicting non-creditworthy is: 63%. As we can see there is a difference of 17% between the creditworthy and the non-creditworthy. This means the Logistic model has bias towards correctly predicting creditworthy because we're having here the higher value (80%).

## 3.2 Decision Tree

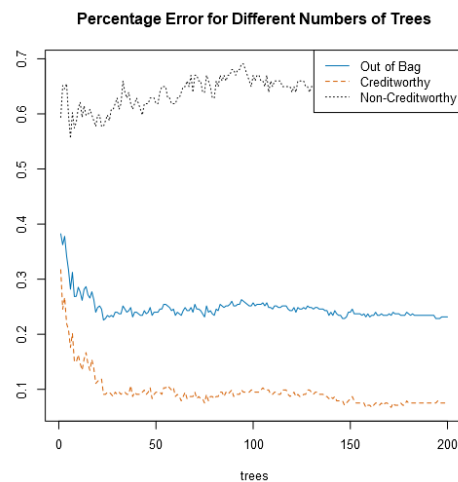
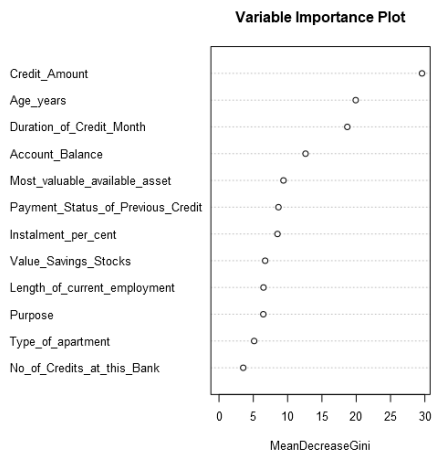


As we can see for the Decision tree: the Account\_Balance, Value Saving Stocks and the Duration of Credit Month are the most important variables.

The overall percent accuracy is: 78 %

Regarding the bias topic: The model accuracy for predicting the creditworthy of: 79 % and the accuracy for predicting non-creditworthy is: 60%. As we can see there is a difference of 19% between the creditworthy and the non-creditworthy. This means the Decision tree model has bias towards correctly predicting creditworthy because we're having here the higher value (79%).

### 3.3 Forrest Model

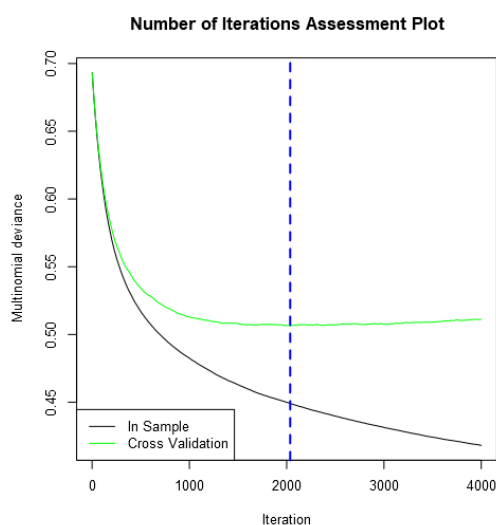
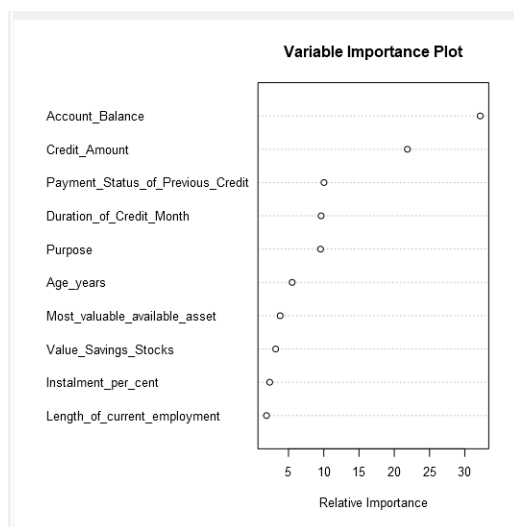


The most important variables are: Credit\_Amount, Age\_years and Duration of credit month.

The overall percent accuracy is: 81 %

Regarding the bias topic: The model accuracy for predicting the creditworthy of: 79 % and the accuracy for predicting non-creditworthy is: 90%. As we can see there is a difference of 11% between the creditworthy and the non-creditworthy. This means the Forest model has bias towards correctly predicting non-creditworthy because we're having here the higher value (90%).

### 3.4 Boosted Model



The most important variables are: Account\_Balance, Credit\_Amount and Payment status of previous credit.

The overall percent accuracy is: 79 %

Regarding the bias topic: The model accuracy for predicting the creditworthy of: 80 % and the accuracy for predicting non-creditworthy is: 63%. As we can see there is a difference of 17% between the creditworthy and the creditworthy. This means the Boosted model has bias towards correctly predicting creditworthy because we're having here the higher value (80%).

#### 4. Write up

Based on section 3 I'm using the Forest mode because of the great accuracy rate of 80% against the validation data.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Regression	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Tree_Regression	0.8967	0.8766	0.7279	0.7623	0.6000
Boosted_Model	0.7867	0.8632	0.7524	0.7629	0.6095
Log_Stepwise	0.7800	0.8364	0.7306	0.8000	0.6288

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

Confusion matrix of Decision_Tree_Regression		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

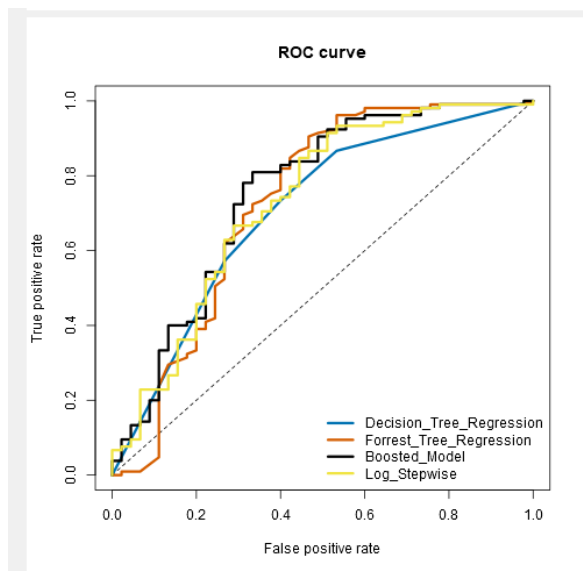
Confusion matrix of Forrest_Tree_Regression		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	27
Predicted_Non-Creditworthy	2	18

Confusion matrix of Log_Stepwise		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

At 'fit and error measures' we can see the accuracy for creditworthy and non-creditworthy for our Forest model we are having: Creditworthy 80% and Non-Creditworthy 90 % if we're using the validation data set.

At the confusion matrix for the Forest model we can see for the true positives we're having the highest value (103) and for the true negative second lowest number (18) of all models. This means the Forest model has a bias towards correctly predicting creditworthy because we're having here the higher value (103).

By using the ROC curve we can also visually see which model is the green line is fairly far up toward the up left, this indicates we've got a fairly good instrument. And if we skip back to the "fit and error measures" we can see that the F1 value for the Forest method is the highest.



We can give a number of 402 customers a credit. For this 402 customers the score\_creditworthy is higher as the score\_non-creditworthy rate.