Project: Creditworthiness

Step 1: Business and Data Understanding

- What decisions need to be made?

The decisions that need to be made are, process the new loan applications and classify the clients by evaluating the creditworthiness of these new loan applicants.

- What data is needed to inform those decisions?

The data needed to inform those decisions is the data related to all past applications and the list of new customers that need to be processed this week.

From the first file, we have these pieces of information about previous applications and their credit application result:

- Credit-Application-Result
- Account-Balance
- Duration-of-Credit-Month
- Payment-Status-of-Previous-Credit
- Purpose
- Credit-Amount
- Value-Savings-Stocks
- Length-of-current-employment
- Instalment-per-cent
- Guarantors
- Duration-in-Current-address
- Most-valuable-available-asset
- Age-years
- Concurrent-Credits
- Type-of-apartment
- No-of-Credits-at-this-Bank
- Occupation
- No-of-dependents
- Telephone
- Foreign-Worker

We will make at first an analysis of what are the variables that are the most important in our prediction, so we probably won't keep the full dataset.

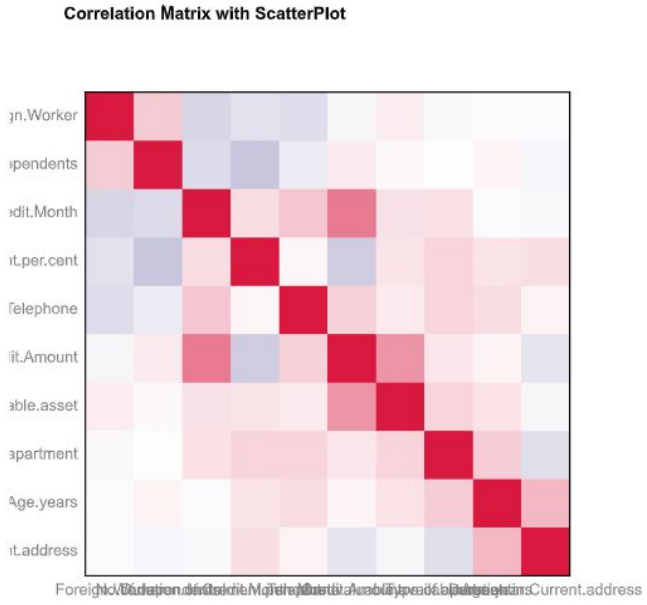
The second file has the same structure except for the Credit-Application-Result that will be predicted with the chosen model.

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

We will use a binary model to classify whether the clients qualify for a loan or not.

Step 2: Building the Training Set

After submitting the dataset to the Association Analysis. The correlation map doesn't show any highly correlated fields.



The field Duration-in-Current-Address has about 69% missing values thus this field should be removed. Age also has 2% missing values, those values can be removed. Besides that, other fields don't have any missing value.

The fields Concurrent-Credit, Guarantors, Occupation, Foreign Worker and No of dependents are fields with low-variability. Those fields will be removed from the dataset.

The field Telephone will also be removed as it is not relevant to the study.

Step 3: Train your Classification Models

First I created an Estimation and Validation samples (70%,30% respectively). I then moved into creating the 4 models we will be testing for this project.

1. **Logistic Regression:**

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

  We can see from the report that the most significant variables are: Account Balance, Purpose and Credit Amount are the top 3 most significant variables with a p-value of less than 0.05.

Record | Report

1 **Report for Logistic Regression Model LR_Credit**

2 *Basic Summary*

3 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)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

6 Coefficients:

7

	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)

8 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

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

The Stepwise-Regression model has an overall accuracy of 76%, 87% for the creditworthy accuracy and 48% for the non-creditworthy. We can, therefore, say that the model is biased toward creditworthy.

Model Comparison Report

Fit and error measures

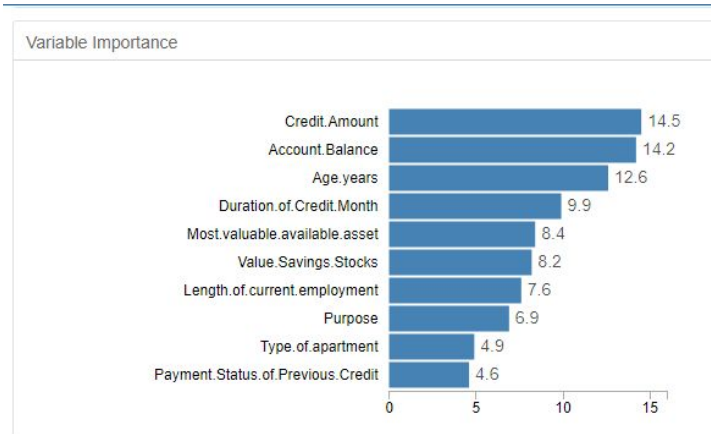
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of LR_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	92	23
Predicted_NonCreditworthy	13	22

2. Decision Tree:

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  We can see from the variable importance analysis that the most important variables are Credit.Amount, Account.Balance & Age.years.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

  The Overall accuracy for the Decision Tree model is 70%, for the creditworthy accuracy, it is at 79%, whereas the accuracy for non-creditworthy is 60%. Therefore we can say that here too, the model is biased toward creditworthy.

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7067	0.7982	0.6764	0.8286	0.4222

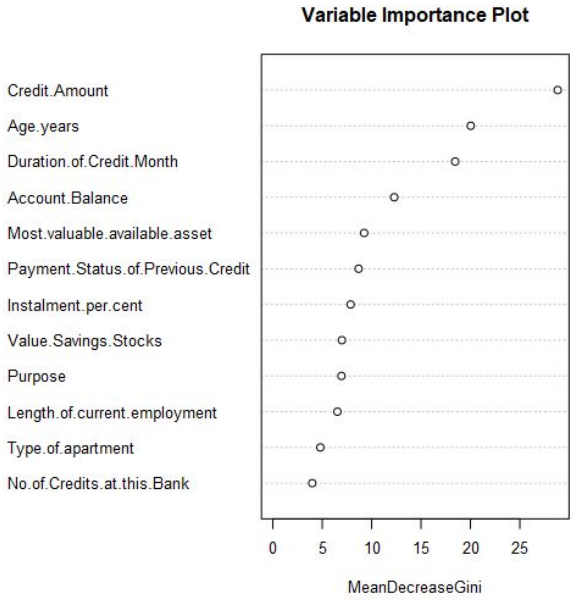
Confusion matrix of DT_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	87	26
Predicted_NonCreditworthy	18	19

3. Random Forrest:

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

  From the variable importance plot, we can see that the Credit.Amout, Age.years, and Duration.of.Credit.Month are the most important variables.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

  The Overall accuracy for the Random Forest model is 81%, for the creditworthy accuracy, it is at 97%, whereas the accuracy for non-creditworthy is 44%. Therefore we can say that here too, the model is highly biased toward creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_NonCreditworthy
RF_Credit	0.8133	0.8793	0.7299	0.9714	0.4444

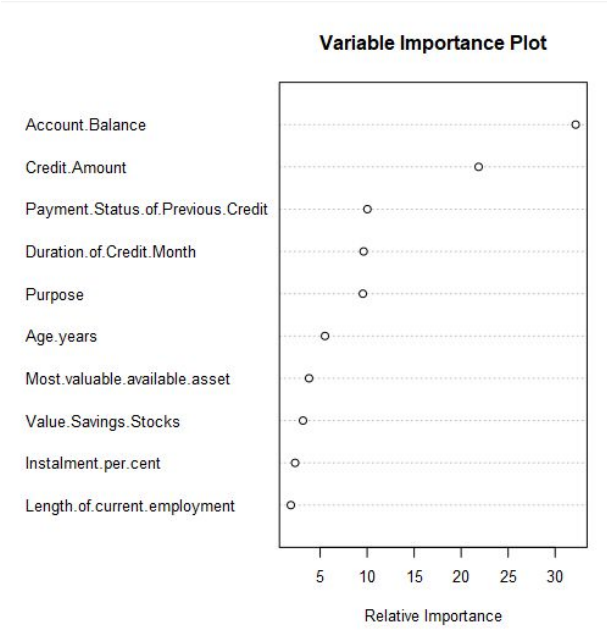
Confusion matrix of RF_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	102	25
Predicted_NonCreditworthy	3	20

4. Boosted Model:

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

  From the variable importance plot, we can see that the Account Balance and Credit.Amount are the most important variables.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

  The Overall accuracy for the Boosted Model is 78%, for the creditworthy accuracy, it is at 96%, whereas the accuracy for non-creditworthy is 37%. Therefore we can say that here too, the model is highly biased toward creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit	0.7867	0.8632	0.7520	0.9619	0.3778

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	101	28
Predicted_NonCreditworthy	4	17

Step 4: Writeup

The chosen model for this problem is Random Forest model, because it had the highest accuracy compared to the other 3 models.
The bias between the creditworthy and non-creditworthy prediction is also almost the same in the 4 models, all of them were biased toward creditworthy. Since the bias is in the 4 models, I decided to only focus on the overall accuracy. I think the bias might be caused by the dataset itself.
When comparing the confusion matrix, Random Forest model seem also to be the best model for this problem.
Finally, Random Forest model reaches the true positive rate at the fastest rate as shown in the ROC curve.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_NonCreditworthy
LR_Credit	0.7600	0.8364	0.7306	0.8762	0.4889
RF_Credit	0.8133	0.8793	0.7299	0.9714	0.4444
DT_Credit	0.7067	0.7982	0.6764	0.8286	0.4222
BM_Credit	0.7867	0.8632	0.7520	0.9619	0.3778

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	101	28
Predicted_NonCreditworthy	4	17

Confusion matrix of DT_Credit

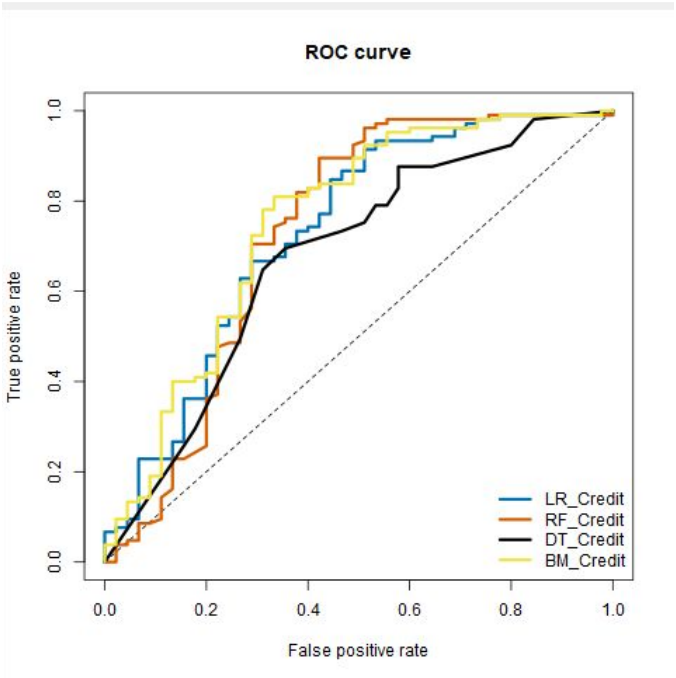
	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	87	26
Predicted_NonCreditworthy	18	19

Confusion matrix of LR_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	92	23
Predicted_NonCreditworthy	13	22

Confusion matrix of RF_Credit

	Actual_Creditworthy	Actual_NonCreditworthy
Predicted_Creditworthy	102	25
Predicted_NonCreditworthy	3	20



- How many individuals are creditworthy? The number of individuals that are creditworthy is 408.

.