**Project: Creditworthiness**

# **Step 1: Business and Data Understanding**

* What decisions need to be made?

Since due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process per week. We need to develop a model which will systematically evaluate the creditworthiness of these new loan applicants. We need to determine the best model that will do the evaluation by having a high accuracy rate. Indeed, we do not want the model to make mistakes and classify some borrowers as creditworthy, but they are not. Thus, it will cost money because they will default on their loans. If the numbers of customers that are not creditworthy are a lot, the bank will have a lot of customers who will default on their loans.

* What data is needed to inform those decisions?

To inform those decisions, we will need to collect data about customers that we have already classified as creditworthy and not creditworthy. We will also need the predictors that determine whether they were creditworthy or not creditworthy. These predictors are:

* Account Balance
* Duration of credit month
* Payment status of previous credit
* Purpose of Credit
* Value of Savings and Stocks
* Employment
* Age
* Most Valuable assets
* Residency
* Numbers of credits at the bank
* Numbers of Dependents
* Concurrent Credits.
* What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

We need to use a binary model to make these decisions because the result will be creditworthy or not creditworthy which are two results.

# **Step 2: Building the Training Set**

* In your cleanup process, which fields did you remove or impute?

In the data clean-up process, there were two variables that had missing values which were:

* Duration in Current Address
* Age.

I impute the missing values with the median 33 for the age and removed the duration in the current address because there were too many missing values.

* Guarantors
* Duration Current Credit
* Occupation
* Numbers of Dependents
* Telephone
* Foreign Workers

# **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.
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any biases seen in the model’s predictions?

**Logistic Regression:**   
  
The predictor’s variables that are significant and most important are:

* Account Balance, p-value = 0.0063
* Duration of Credit – Month, p-value = 0.01157
* Payment Status of Previous Credit, p-value = 0.01217
* Length of Current Employment less than 1 year, p-value = 0.01783
* Numbers of Credit at the Bank, p-value = 0.04126
* Numbers of Dependents, p-value = 0.01365.

The overall percent accuracy is 0.7442. There is bias in the model because there 8 customers who were not creditworthy, but the model classifies them as creditworthy.

|  |
| --- |
| **Confusion matrix of Stepwise** |
| |  | Actual\_Creditworthy | Actual\_Non-Creditworthy | | --- | --- | --- | | Predicted\_Creditworthy | 30 | 8 | | Predicted\_Non-Creditworthy | 3 | 2 | |

**For Decision Tree:** The variables used in the tree construction are:

* Age
* Duration of Credit.

The overall percent accuracy is 0.7209. There is bias in the model because there 10 customers who were not creditworthy, but the model classifies them as creditworthy.

|  |
| --- |
| **Confusion matrix of Decision\_Tree\_** |
| |  | Actual\_Creditworthy | Actual\_Non-Creditworthy | | --- | --- | --- | | Predicted\_Creditworthy | 31 | 10 | | Predicted\_Non-Creditworthy | 2 | 0 | |

**Forest Model:** According to the variable importance plot, the predictor’s variables that are significant and most important are:

* Credit Amount (Very Important)
* Duration of Credit
* Age
* Payment Status of Previous Credit
* Installment
* Account Balance
* Value Savings and Stocks (Important)
* Length of Current Employment
* Purpose
* Numbers of dependents. (Important)

The overall percent accuracy is 0.7674. There is bias in the model because there 9 customers who were not creditworthy, but the model classifies them as creditworthy.

|  |
| --- |
| **Confusion matrix of FM\_Creditworthy** |
| |  | Actual\_Creditworthy | Actual\_Non-Creditworthy | | --- | --- | --- | | Predicted\_Creditworthy | 32 | 9 | | Predicted\_Non-Creditworthy | 1 | 1 | |
|  |

**Boost Model:** According to the variable importance plot, the predictor’s variables that are significant and most important are:

* Duration of Credit (Very Important)
* Age
* Account Balance
* Credit Amount
* Installment
* Value of Savings and Stocks
* Payment of Previous Credit (Important)
* Length of Current Employment.

The overall percent accuracy is 0.7674. There is bias in the model because there 10 customers who were not creditworthy, but the model classifies them as creditworthy.

# **Step 4: Writeup**

I chose the Forest Model because of the overall percent accuracy compared to other models. When I looked at the ROC Graph, the Forest Model is better than the Boost Model even though it has slightly more bias than the Boost Model. The Forest Model and Boost Model have the same overall percent accuracy.

* How many individuals are creditworthy?

To determine the number of customers who are creditworthy, I used the output tool at the end of the workflow on the Forest Model so that I can score the model. I was able to score the model and I found the percent of creditworthy for each customer. I decide to use a filter tool so that I can determine which customers are creditworthy. I decide that customers who have a percent that is greater or equal to 0.70 are creditworthy.

According to the Forest Model, there are 335 customers who are creditworthy and 164 customers who are not creditworthy.

