Project: Creditworthiness Prediction

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

# **Business problem**

A reputable bank had an influx of new people applying for loans. In a given week, there were 500 loan applications to process. The manager of the bank sees this new influx as a great opportunity and wants to figure out how to process all of these loan applications within one week.

With Data on all past applications and the list of customers that need to be processed in the next few days, the manager needs application of analytical techniques to speed up the process. The main concern is to predict who creditworthy with good accuracy.

# **Business and Data Understanding**

## **Key Decisions:**

Decision needs to be made on which customers are credit worthy.

**Data Needed**

Data on past applications and new applications are needed to inform the decision.

The data should include but not limited to Customer ID, Credit application results, Account balance, Duration of credit month, Payment status of previous credit, Purpose, Credit amount, Value savings stocks, Installation percent, Most-valuable available asset, Age-years, Type of apartment, Number of credits at the bank, working status (Foreign worker or not) and any other data that will help assess whether a customer is credit worthy or not.

The data on new application will include all that is mentioned above with the exception of credit application results.

**Data Availability.**

The bank has enough data to build a classification model.

**Data Science Technique needed to solve the problem.**

A binary classification model is needed to help make the decision.

Binary classification is needed because the target variable contains two classes being credit worthy and non-credit worthy.

# **Step 2: Data preparation and feature selection**

The raw data contain columns as displayed in the table below:

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

* In the cleanup process, Occupation, Guarantor, No-of-dependents, Concurrent credits, Foreign-Worker, were removed because they have low variance.

Occupation and Concurrent-credits have one unique value and one unique category respectively.

Guarantors and No-of-dependents had one category dominating resulting in low variance.

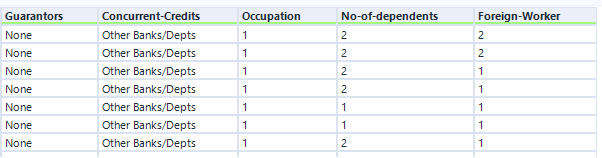
Duration in-current-address was removed because it has a lot (68.8%) of missing values.

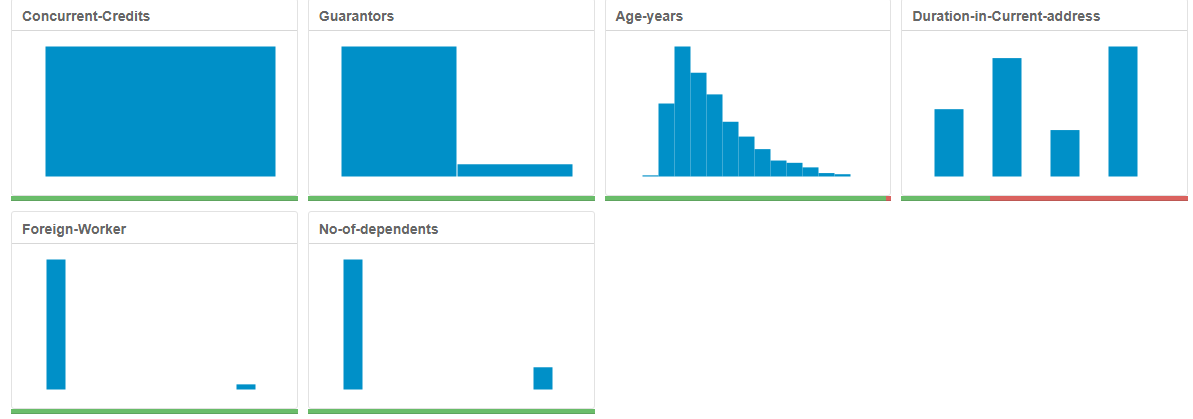
Telephone was also removed because it does not make logical sense to be included.

Age- years has few missing values (2.4% missing values), and were imputed using the median.

.

These are shown in the pictures below.





# Step 3: Building Classification Models

70% of the data was used to for training the data whilst 30% was used for validation*.*

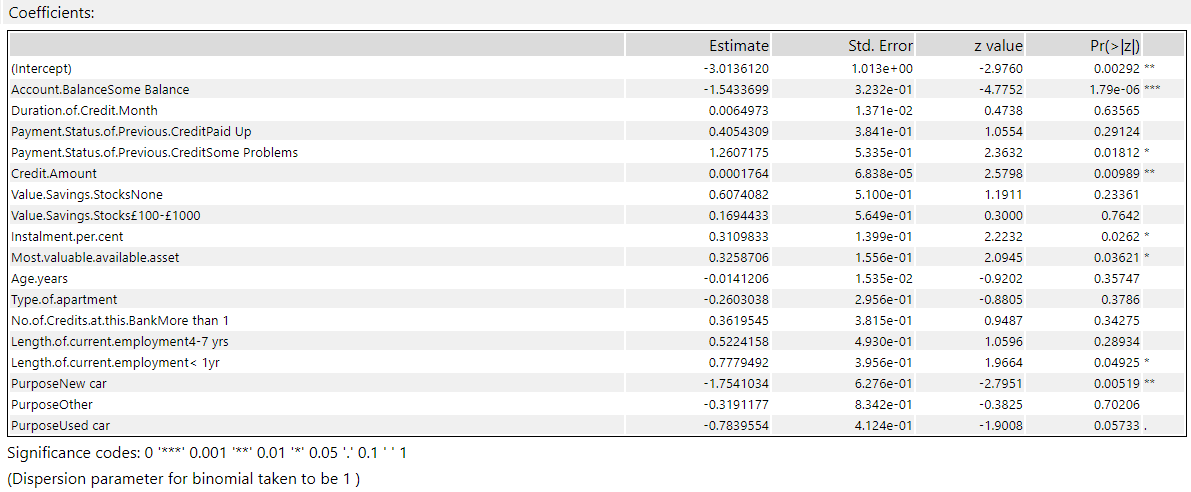
Random Seed was set to 1.

Logistic Regression, Decision Tree, Forest Model, Boosted Model were built. The models will later be compared and the best will be chosen

For each model, the most important (significant) variables were listed to provides information on which variables will needed the most in case the model is selected for production.

Graph are provided to provide more insight.

**Logistic regression**



Considering the chart above, variables that are significant in the Logistic regression at 0.05 significant level includes:

1.Account balance

2.Credit Amount

3.Purpose

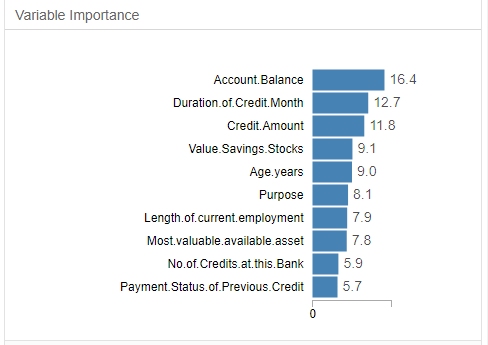
4. Payment Status

5.Instalment percent

6.Most valuable asset

7.Length of current employment

**Decision Tree Model**



Most important variables in the Decision Tree model include:

1.Account Balance.

2.Duration of credit month

3.Credit Amount

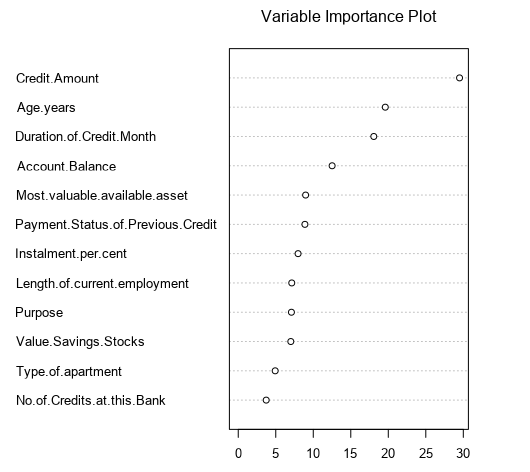
4.Age years

5.Value Savings Stocks

6.Purpose

7.Length of current employment

**Random Forest Model**



Most important variables in the Random Forest model include:

1.Credit Amount

2.Age years

3.Duration of credit month

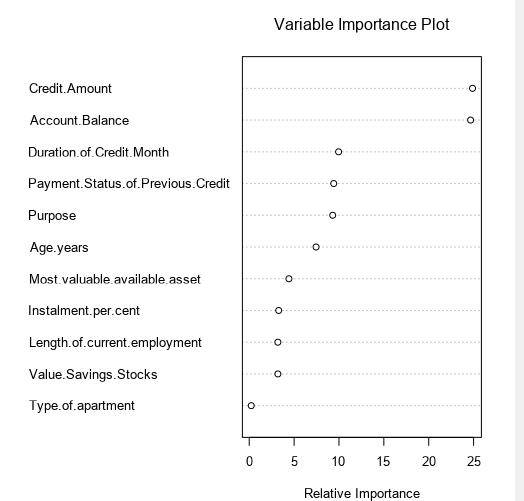
4.Account Balance

5.Payment Status of Previous Credit

6.Most valuable available Asset

7.Instalment percent

**Boosted Model**



Most important variables in the Boosted model include:

1.Credit Amount

4.Account Balance

3.Duration of credit month

4.Purpose

5.Payment Status of Previous Credit

6.Age years

7.Most valuable available asset

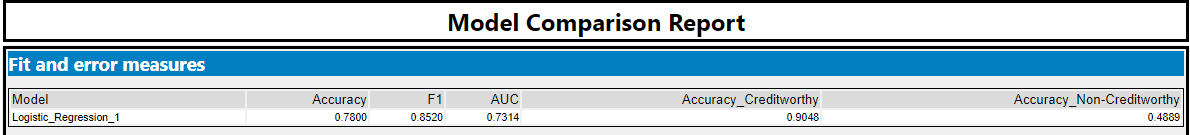
**Step 4: Model Validation**

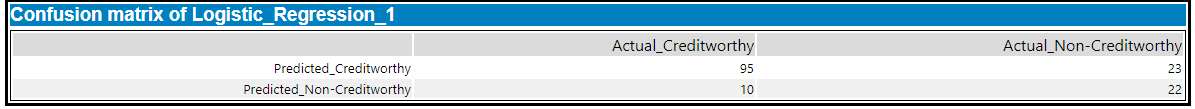
The models were validated against the validation set. The overall percent accuracy, confusion matrix, F1 score and ROC graph were obtained as metrics for validating the models.

The outputs are shown below.

**Overall percent accuracy and Confusion Matrix**

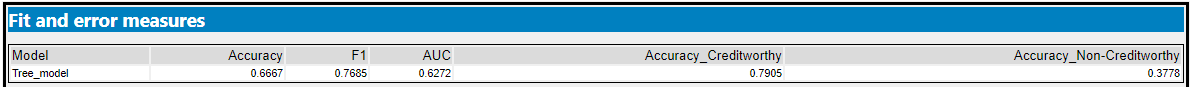
**Logistic Regression**

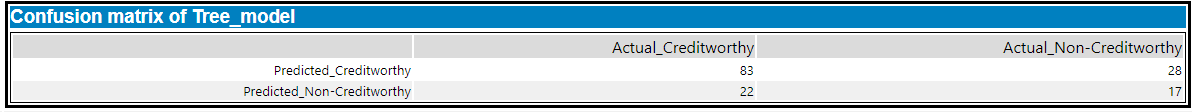
****

****

The overall accuracy of the logistic regression is 0.7800. Considering the confusion matrix, the model is biased toward Creditworthy. It predicts cases as creditworthy than Non-Creditworthy at most times.

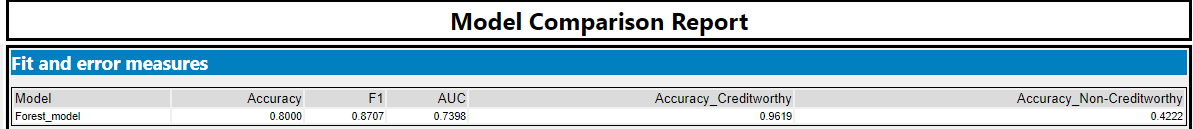
**Decision Tree**

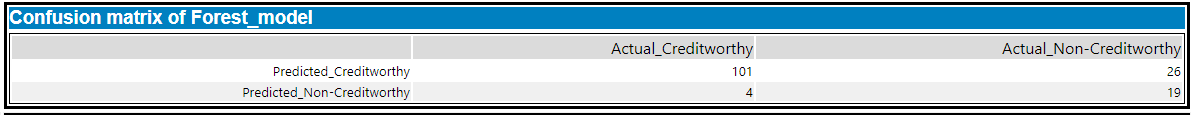
****

****

The overall accuracy of the Decision Tree is 0.6667. Considering the confusion matrix, the model is biased toward Creditworthy. It predicts cases as creditworthy than Non-Creditworthy at most times.

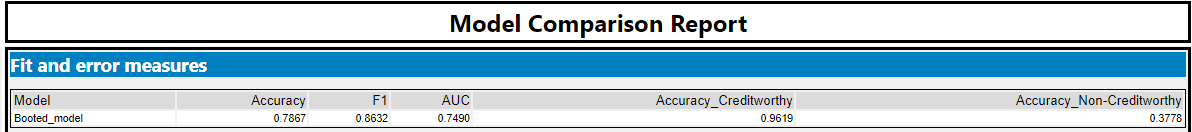
**Random Forest**

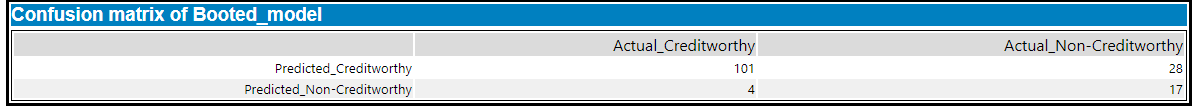
****

****

The overall accuracy of the Random Forest model is 0.8000. Considering the confusion matrix, the model is biased toward Creditworthy. It predicts cases as creditworthy than Non-Creditworthy at most times.

**Boosted Model**

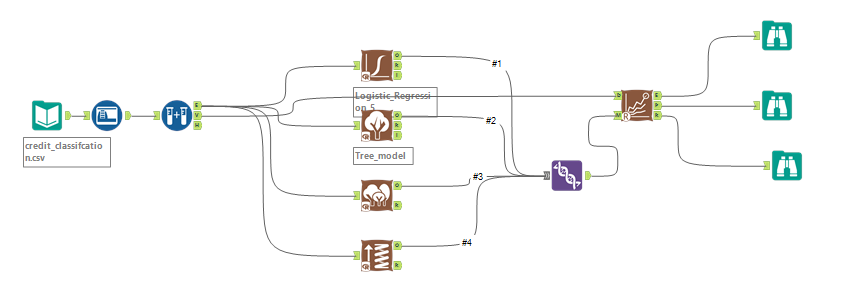
****

****

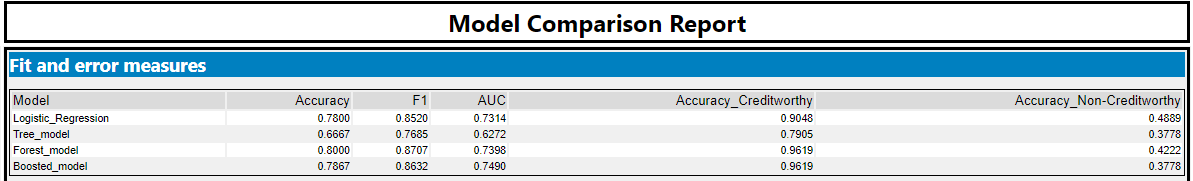
The overall accuracy of the Boosted model is 0.7867. Considering the confusion matrix, the model is biased toward Creditworthy. It predicts cases as creditworthy than Non-creditworthy at most times.

# Step 5 Model comparison and selection:

**Work flow**

****

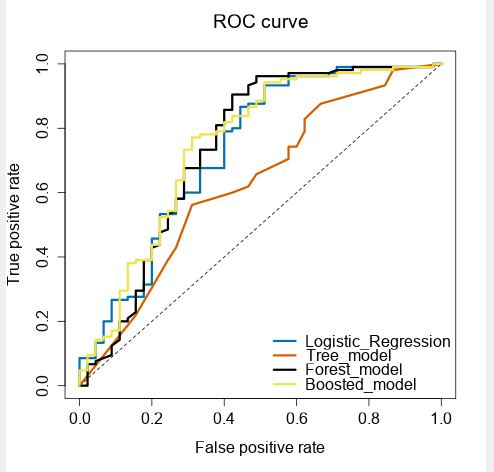
**Overall Accuracy and Accuracies within “Creditworthy” and “Non-Creditworthy” segments**



Random Forest model has the highest overall accuracy of 0.8000, followed by the Boosted model with an overall accuracy of 0.7867. The Decision tree model has the lowest overall accuracy of 0.6667.

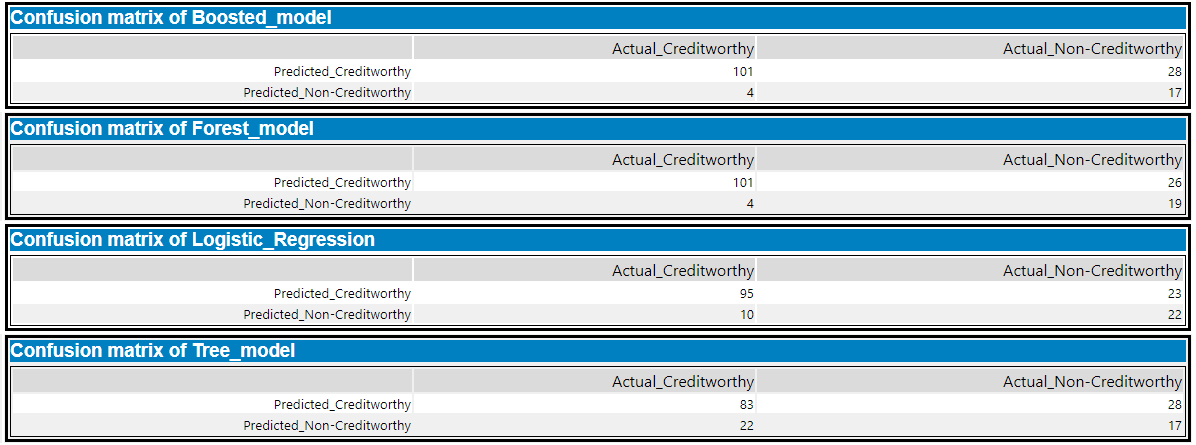
Considering the prediction accuracies for Creditworthy and Non-Creditworthy from the model comparison report, the Forest model and the Boosted model have the highest prediction accuracy of 0.9619 for creditworthy. However, the Forest model has a higher prediction accuracy (0.4222) for Non-Creditworthy than the Boosted model (0.3778) making it a better choice.

Logistic regression has the highest prediction accuracy (0.4889) for Non-Creditworthy but that is not a major concern based on the objectives of management.



The ROC graph suggests that the Boosted model and the Forest model are doing well. However, considering the curves of the two models, the Boosted performs slightly better than Forest model.

**Confusion Matrix**



From the confusion matrix report above, The Logistic regression model and the Tree model do not classify creditworthy correctly as compared to the Forest model and the Boosted model.

The Forest model and the Boosted model rightly classified 101 and misclassified 4 Creditworthy. But the Boosted model classifies (28) Non-Creditworthy better than the Forest model (26).

However, because, my management only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments, Random Forest model is chosen since it has the best prediction accuracy for both Creditworthy and Non-Creditworthy.

Using the selected model**, 412** individuals are predicted as creditworthy.