**Creditworthiness of Loan Applicants.**

**1 - Business and Data Understanding.**

**What business decisions need to be made?**

Due to a financial scandal that hit a competitive bank last week, there has been a sudden influx of new people applying for loans at the bank. All of a sudden there are nearly 500 loan applications to process this week as opposed to typical 200 loan applications per week which are approved by hand.

This new influx is a great opportunity and the bank wants to figure out how to systematically evaluate the creditworthiness of these new loan applicants.

**What data is needed to inform these decisions?**

The data needed will come from “credit-data-training.xlsx”. The data has already been cleaned, however it will still need to be checked for missing data and later used to train four different models. The models will be compared to find the most suitable.

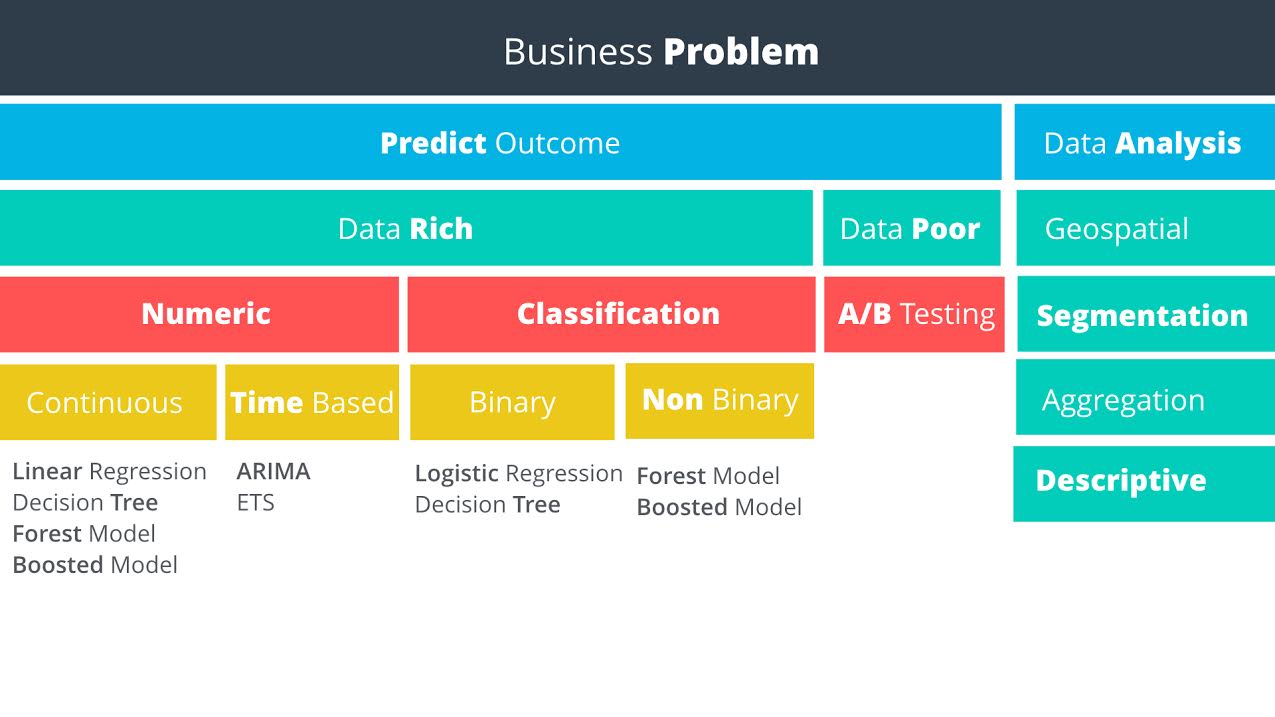
The chosen model will be used to predict the loan applicants worthy of a loan from “customers-to-score.xlsx”.

The columns used from “credit-data-training.xlsx” are:

|  |
| --- |
| **Columns Names** |
| Credit-Application-Result |
| Account-Balance |
| Duration-of-Credit-Month |
| Payment-Status-of-Previous-Credit |
| Purpose |
| Credit-Amount |
| Value-Savings-Stocks |
| Length-of-current-employment |
| Most-valuable-available-asset |
| No-of-Credits-at-this-Bank |
| Type-of-Apartment |
| Instalment-per-cent |
| Age-years |

**What kind of model do we need to use to help make these decisions?**

Using the methodology map below to aid my decision making:



I can see that the business problem requires me to:

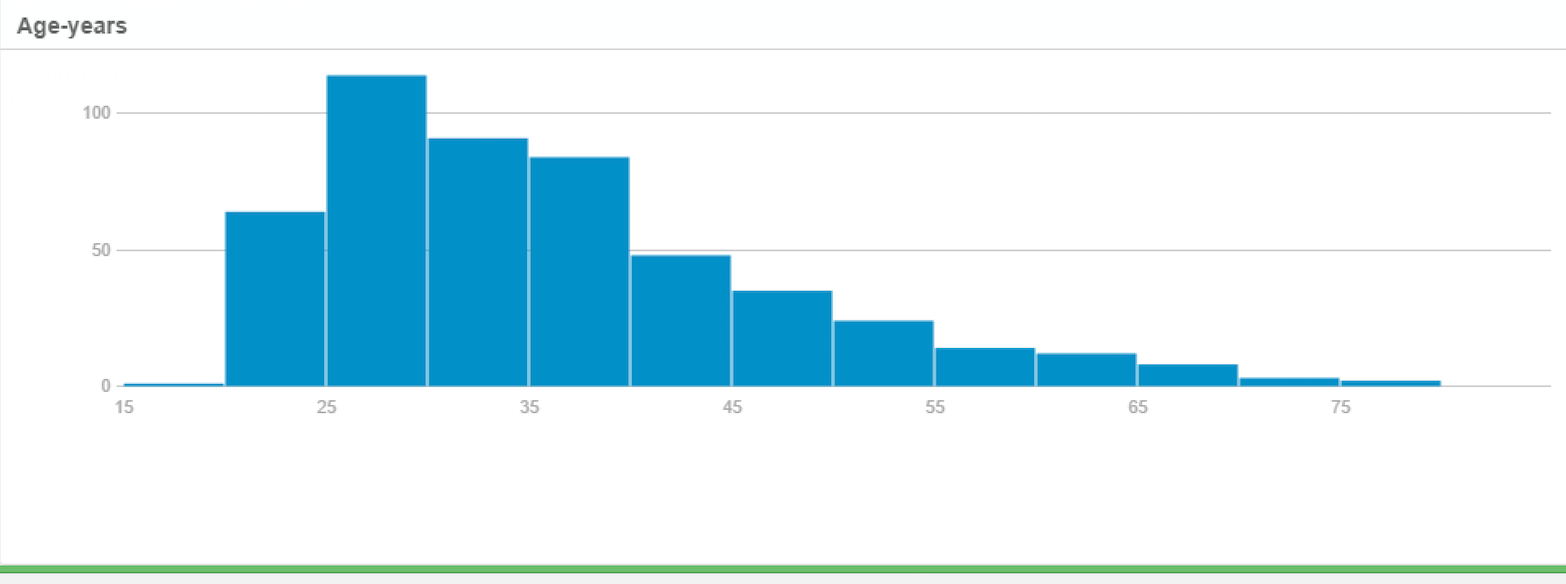
1. Predict an Outcome
2. Use Rich Data
3. Classify the available data
4. Obtain a binary outcome ie, give the applicant a loan or not.

The model used will likely be a binary classification model.

**2 - Building the training set.**

The data used to train the model will come from “credit-data-training.xlsx”. Predictor variables will need to be chosen based on their relationship with the target variable which is whether the applicant will be creditworthy.

We will be looking at variability of data, an example of which is the Age-years variable which is demonstrated in the below chart.



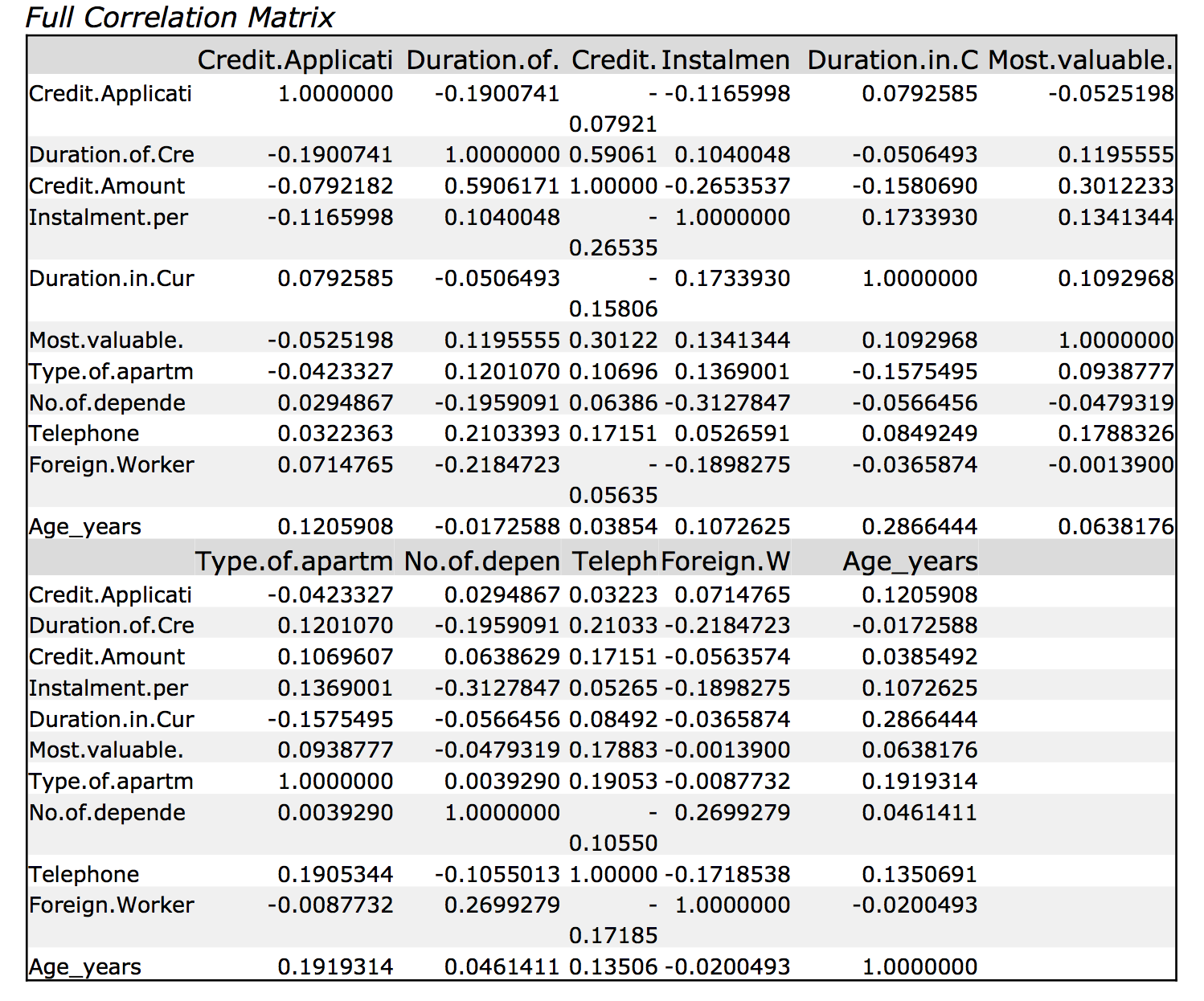
Correlation of the data is another factor used to help find suitable predictor variables. The full correlation matrix is summarised further in the document.

**For numerical data fields, are there any fields that are highly correlated?**

Below is a summary of the data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Field Category** | **Min** | **Max** | **Median** | **Std. Dev.** | **Percent Missing** | **Unique Values** | **Mean** | **Shortest\_Value** | **Longest\_Value** | **MinValueCount** | **MaxValueCount** |
| Age-years | Numeric | 19 | 75 | 33 | 11.50152219 | 2.4 | 54 | 35.63729508 |  |  |  |  |
| Credit-Amount | Numeric | 276 | 18424 | 2236.5 | 2831.386861 | 0 | 464 | 3199.98 |  |  |  |  |
| Duration-in-Current-address | Numeric | 1 | 4 | 2 | 1.150017082 | 68.8 | 5 | 2.66025641 |  |  |  |  |
| Duration-of-Credit-Month | Numeric | 4 | 60 | 18 | 12.30742009 | 0 | 30 | 21.434 |  |  |  |  |
| Foreign-Worker | Numeric | 1 | 2 | 1 | 0.191387718 | 0 | 2 | 1.038 |  |  |  |  |
| Instalment-per-cent | Numeric | 1 | 4 | 3 | 1.113723826 | 0 | 4 | 3.01 |  |  |  |  |
| Most-valuable-available-asset | Numeric | 1 | 4 | 3 | 1.064267509 | 0 | 4 | 2.36 |  |  |  |  |
| No-of-dependents | Numeric | 1 | 2 | 1 | 0.353459853 | 0 | 2 | 1.146 |  |  |  |  |
| Occupation | Numeric | 1 | 1 | 1 | 0 | 0 | 1 | 1 |  |  |  |  |
| Telephone | Numeric | 1 | 2 | 1 | 0.490388583 | 0 | 2 | 1.4 |  |  |  |  |
| Type-of-apartment | Numeric | 1 | 3 | 2 | 0.539813669 | 0 | 3 | 1.928 |  |  |  |  |
| Account-Balance | String |  |  |  |  | 0 | 2 |  | No Account | Some Balance | 238 | 262 |
| Concurrent-Credits | String |  |  |  |  | 0 | 1 |  | Other Banks/Depts | Other Banks/Depts | 500 | 500 |
| Credit-Application-Result | String |  |  |  |  | 0 | 2 |  | Creditworthy | Non-Creditworthy | 142 | 358 |
| Guarantors | String |  |  |  |  | 0 | 2 |  | Yes | None | 43 | 457 |
| Length-of-current-employment | String |  |  |  |  | 0 | 3 |  | < 1yr | 1-4 yrs | 97 | 279 |
| No-of-Credits-at-this-Bank | String |  |  |  |  | 0 | 2 |  | 1 | More than 1 | 180 | 320 |
| Payment-Status-of-Previous-Credit | String |  |  |  |  | 0 | 3 |  | Paid Up | No Problems (in this bank) | 36 | 260 |
| Purpose | String |  |  |  |  | 0 | 4 |  | Other | Home Related | 15 | 355 |
| Value-Savings-Stocks | String |  |  |  |  | 0 | 3 |  | None | £100-£1000 | 48 | 298 |

Below is the correlation matrix of all the variables with Credit\_Applicant\_Result as the target variable.



Looking through the correlation matrix and using 0.7 as the benchmark for high correlation, there seems to be nothing of high correlation with the numerical data fields.

**Are there any missing data fields?**

Looking at the summary, there are missing fields in Age-years and Duration-in-apartment.

There are too many missing fields in Duration-in-apartment and therefore this field will be excluded from the analysis.

Age-in-years is missing only 2.4% of its data and in turn I will substitute missing data here with the average age of the dataset.

**Are there any fields with low variability?**

Below are columns that potentially show low variability due to the majority of its data being one sided:

|  |
| --- |
| Foreign-worker |
| Guarantors |
| Concurrent-Credits |
| Telephone |
| Occupation |
| No-of-dependents |

Data with low variability will be excluded from the model.

The below is my final dataset for modelling.

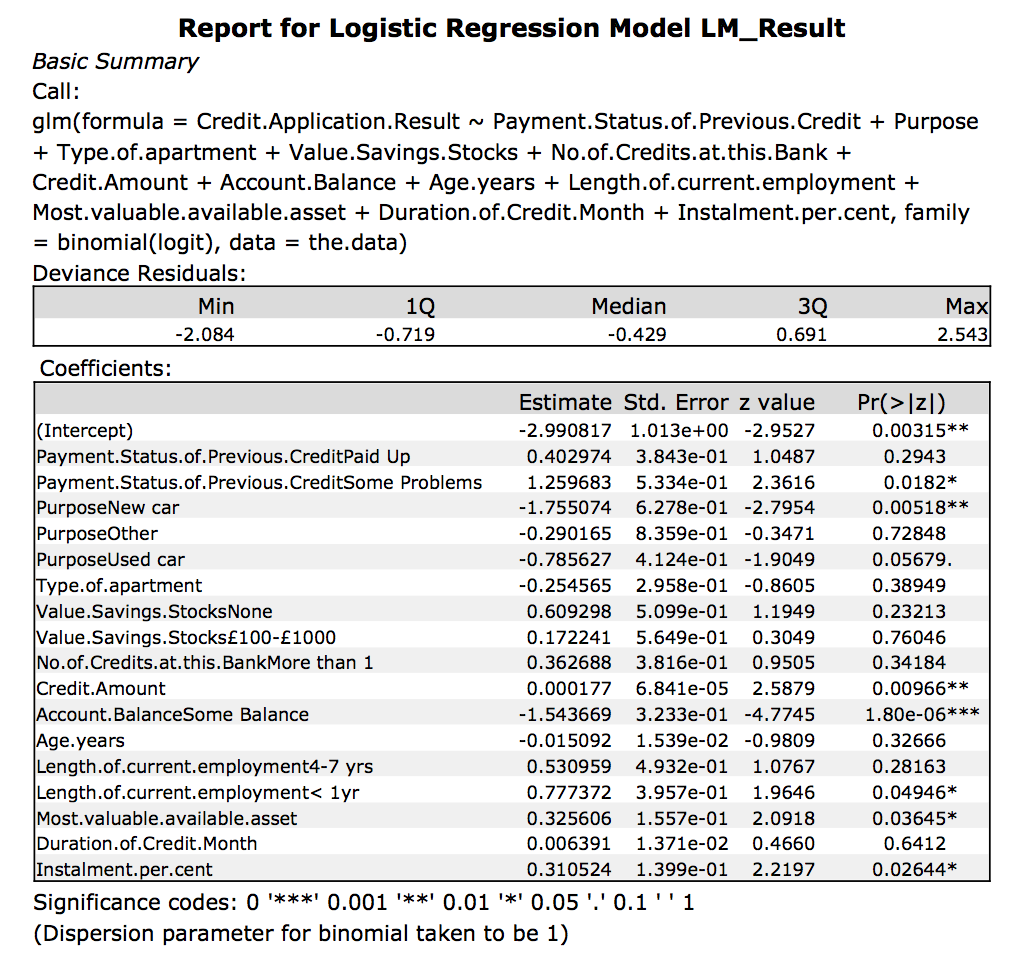
|  |
| --- |
| **Columns Names** |
| Credit-Application-Result |
| Account-Balance |
| Duration-of-Credit-Month |
| Payment-Status-of-Previous-Credit |
| Purpose |
| Credit-Amount |
| Value-Savings-Stocks |
| Length-of-current-employment |
| Most-valuable-available-asset |
| No-of-Credits-at-this-Bank |
| Type-of-Apartment |
| Instalment-per-cent |
| Age-years |

A sanity check of the 13 final columns and gives an **average age of 35.637 or 36 yrs** (rounded up to the nearest year)

**Creating the model.**

In order to create the models, a 70/30 split was done to create an estimation and validation dataset.

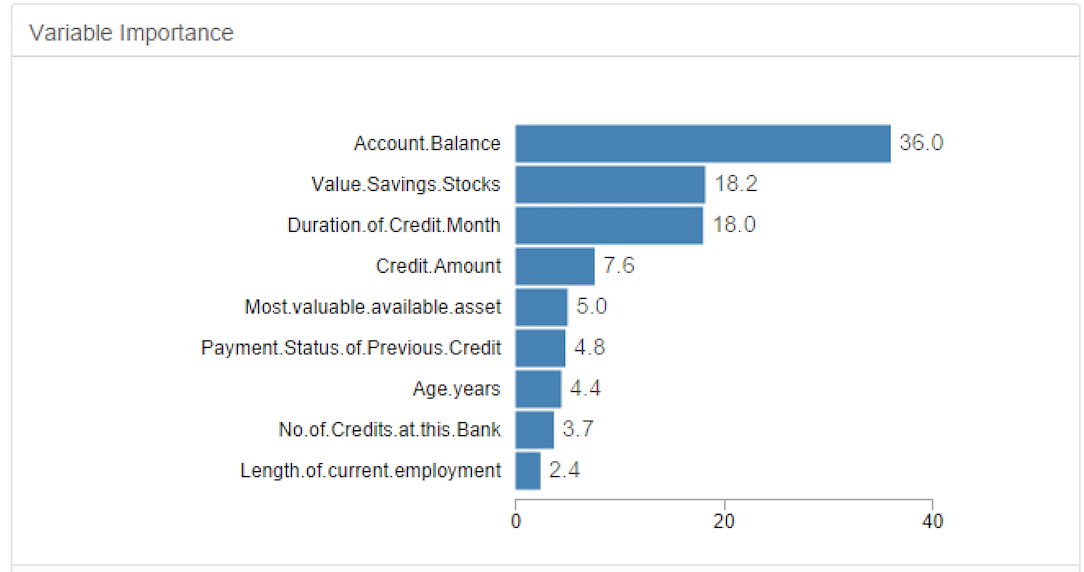
The models were run and each of the model summaries are below.



For the Logistic Model, the most significant predictor variables are:

|  |  |
| --- | --- |
| **Variable Name** | **p-value** |
| Payment.Status.of.Previous.CreditSome.Problems | 0.0182 |
| PurposeNew car | 0.00518 |
|  | 0.00966 |
| Account.BalanmceSome balance | 1.80e-06 |
| Length.of.current.em,plyment<1yr | 0.04946 |
| Most.valuable.available.asset | 0.03645 |
| Instalment.per.cent | 0.02644 |

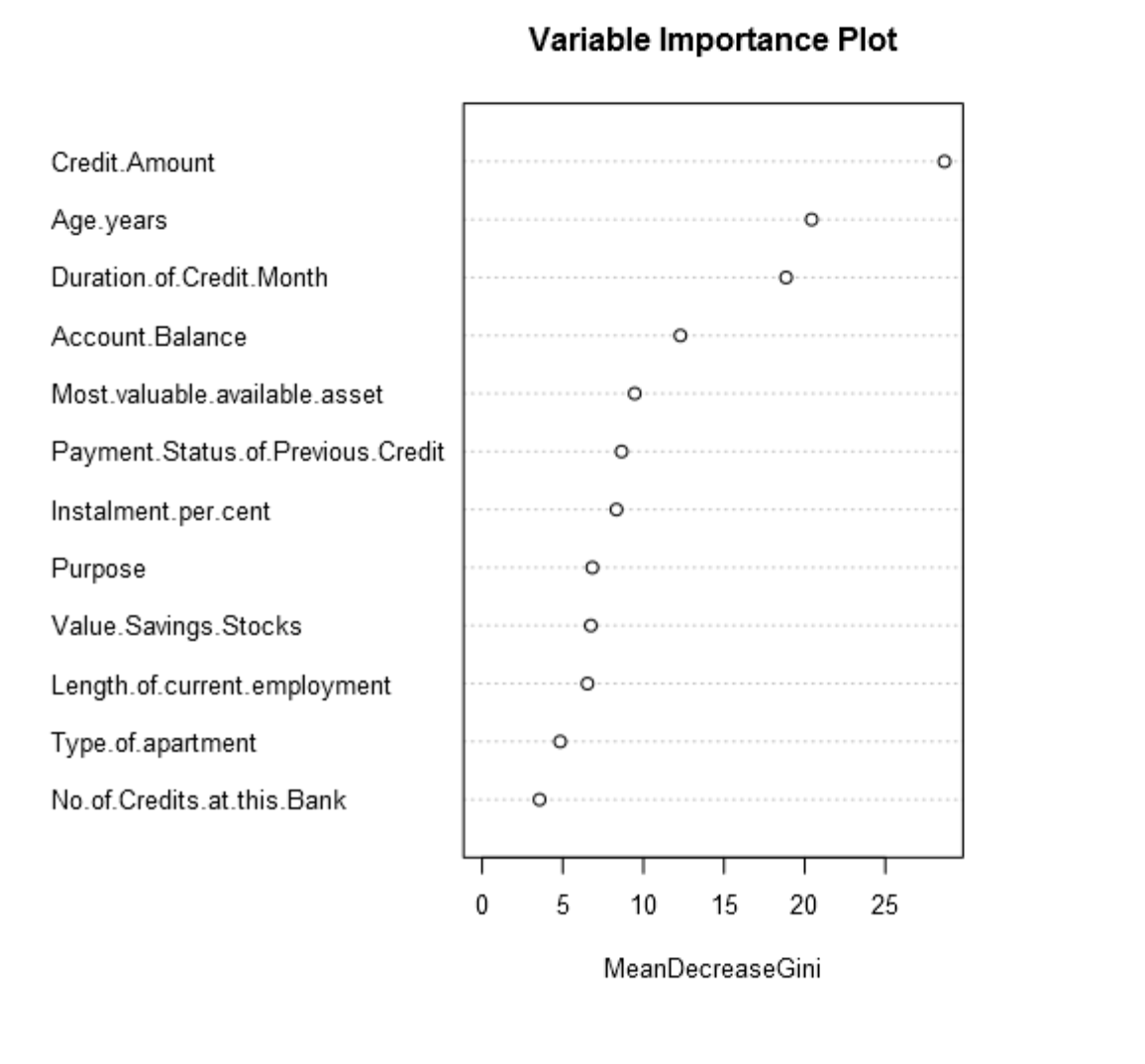
Variable Importance Chart for the decision tree model.



The most important variables in the decision tree model are:

|  |
| --- |
| Account Balance |
| Value.Savings.Stocks |
| Duration.of.Credit.Month |

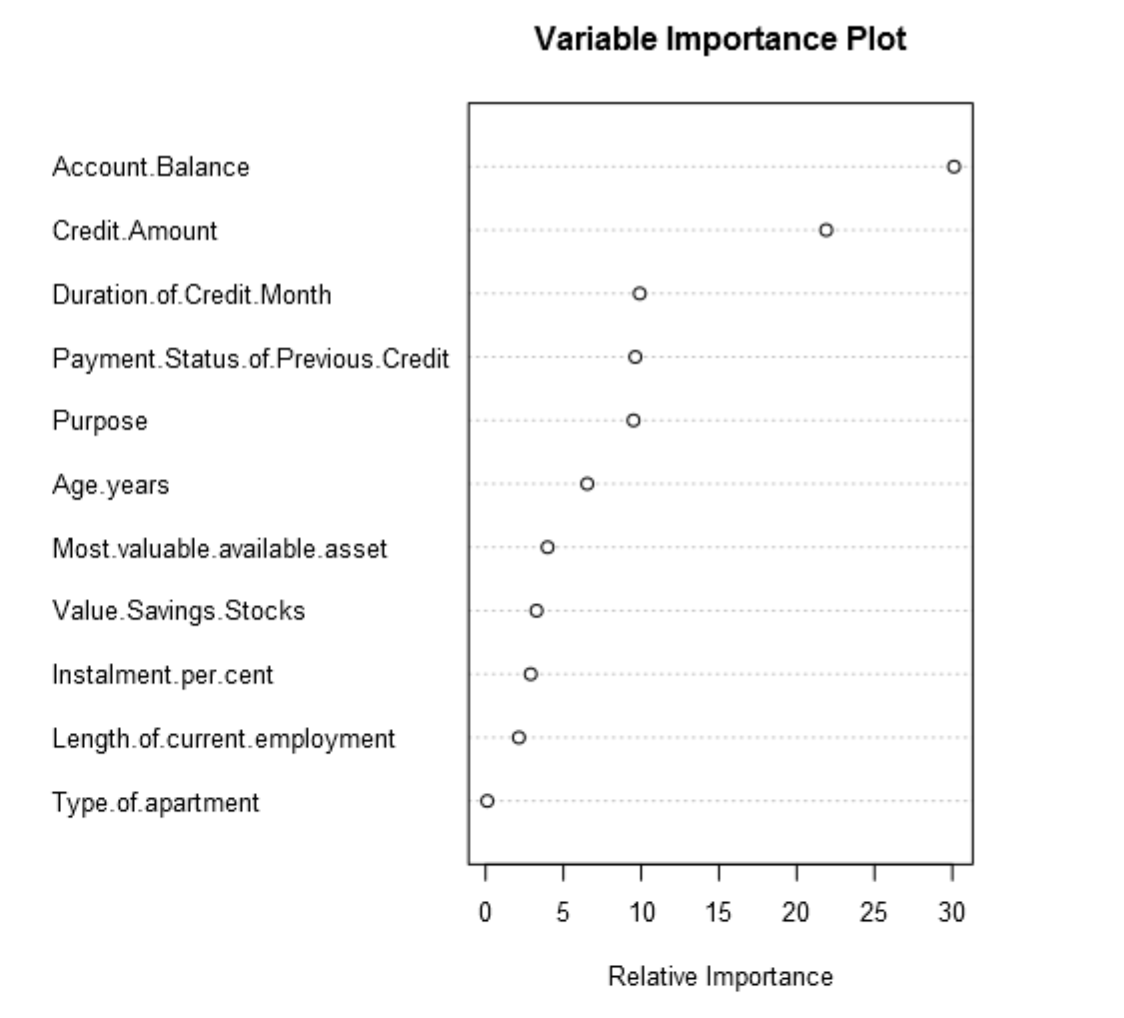
Below is the variable importance chart for the random forest model.



The chart indicates the most important predictor variables for the random forest model are:

|  |
| --- |
| Credit.Amount |
| Age.years |
| Duration.of.Credit.Month |

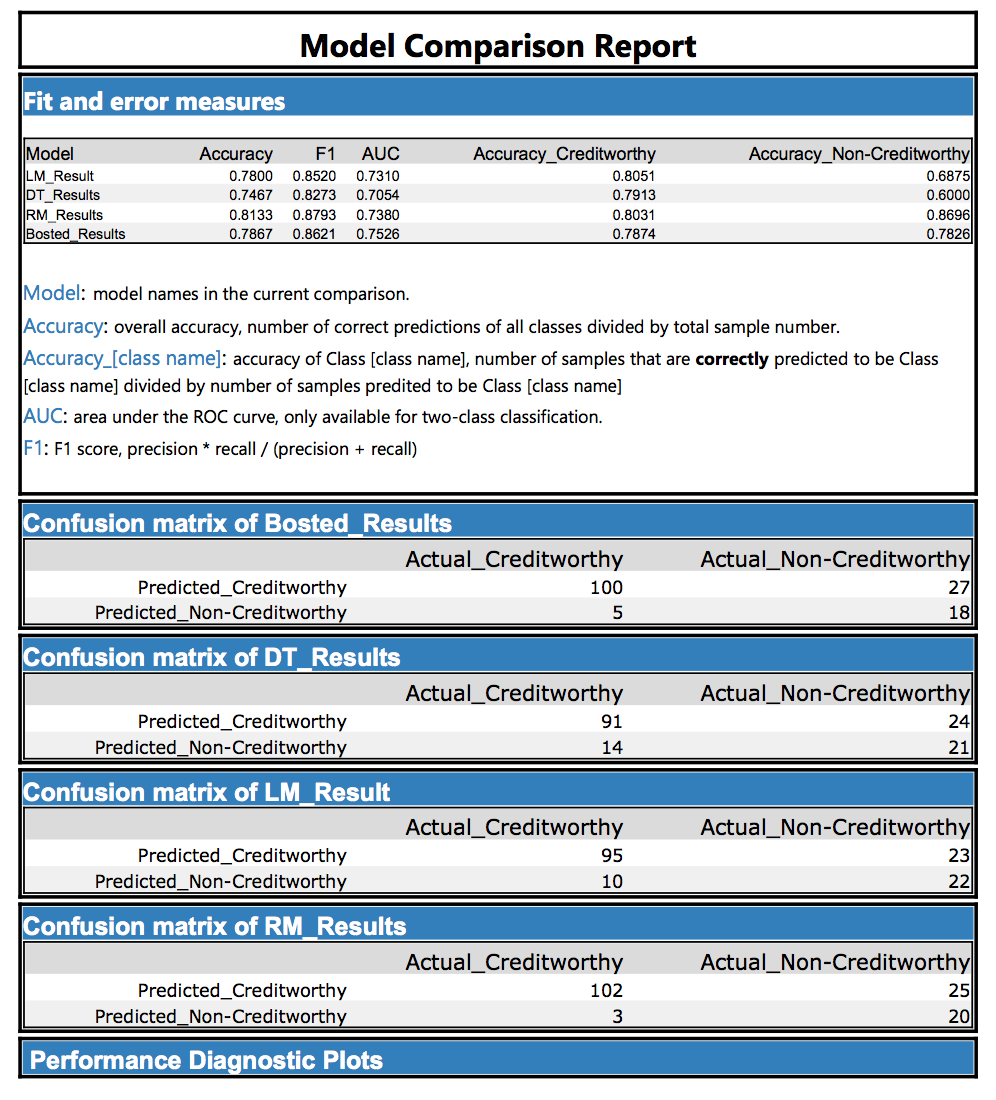
Below is the variable importance plot of the boosted model.



The most important variables for the boosted model are:

|  |
| --- |
| Credit.Amout |
| Account.Balance |

Below is the accuracy and confusion matrix to each model having been validated through the validation dataset.



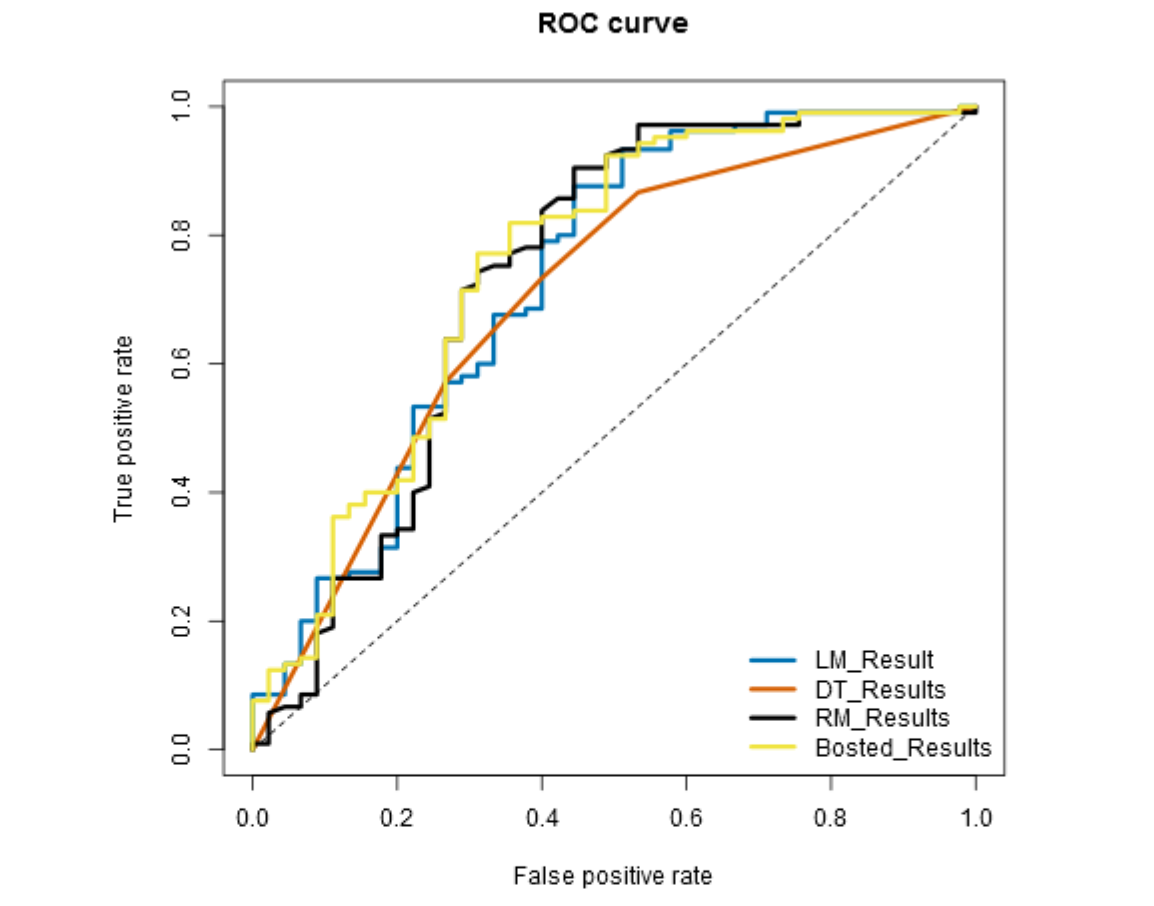
The model with the highest accuracy score is the Random Forest Model at 0.8133.

The models appear to predict Creditworthy more accurately than Non-Creditworthy. It also looks like there are more applicants that are creditworthy and not.

**Final Model.**

The final model used for prediction will be the Random Forest model due to its highest overall accuracy at 0.8133. It has a high accuracy, 0.803, score for predicting Creditworthy applicants and also the highest accuracy score, 0.8696, for predicting non-creditworthy applicants

Below is the ROC chart for the models.



The ROC plots shows the Random Forest model to be the second best with an AUC of 0.7380.

Applying the model to the new dataset, customers-to-score.xls and taking any applicant that has a greater Creditworthy accuracy score than non-creditworthy to mean the applicant should be granted a loan, the final count of **individuals whom are creditworthy are 411**.