**Project: Creditworthiness**

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

## **Key Decisions:**

**Answer these questions**

1. **What decisions needs to be made?**

The Decision that needs to be taken here is to classify the customer either as credit worthy or non-credit worthy and based on that we need classify the customer into two categories.

1. **What data is needed to inform those decisions?**

We need to know information that are:

* **‘credit-data-training.xlsx’**: This file contains the data of the customers to whom bank has provided the loan to and based on that data we can make predictive model to analyse the ‘credit-data-training.xlsx’ data set and can categorize the customer into credit worthy or non-credit worthy.
* The variables which will be useful in deciding the credit worthiness of the customer will be: Account Balance, Credit Amount, Payment Status of Previous Credit, Purpose New Car, Value Saving Stocks, Age-Year, Duration of Credit Month.

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

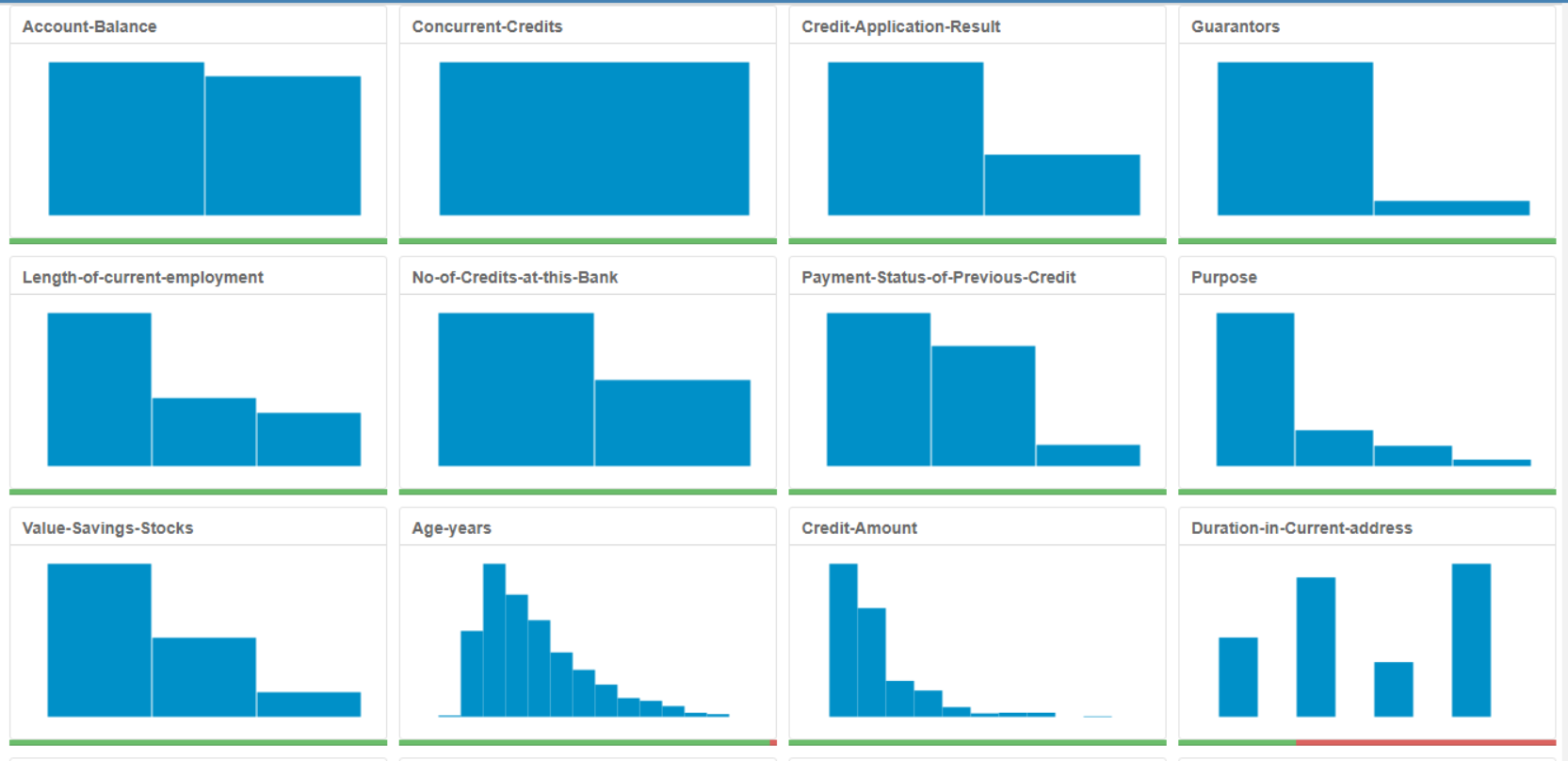
The model that we need to build for this problem is Binary model as we need to decide whether the customer is credit worthy or non-credit worthy.

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

1. **In your clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.**

There are several fields which are removed and one field which is imputed and they are:

* The Histogram of the variable **Guarantors, Foreign-Worker** and **No of dependents** have shown that majority of data is heavily skewed towards one type of data.These variables are removed due to heavily skewed towards one data.
* **Concurrent Credits** and **Occupation** both of them have entirely uniform data and there are no other variations in the data that’s why both of them are removed due to low variability.
* **Duration in Current address** has 69% of the missing data. That’s why this field is removed.
* **Telephone,** this field does not have any predictive ability to credit application results, so this field is also removed.
* **Age-Year** this field has 2% of missing value. The missing data of this variable has been imputed using the median, 33 of the entire data field.

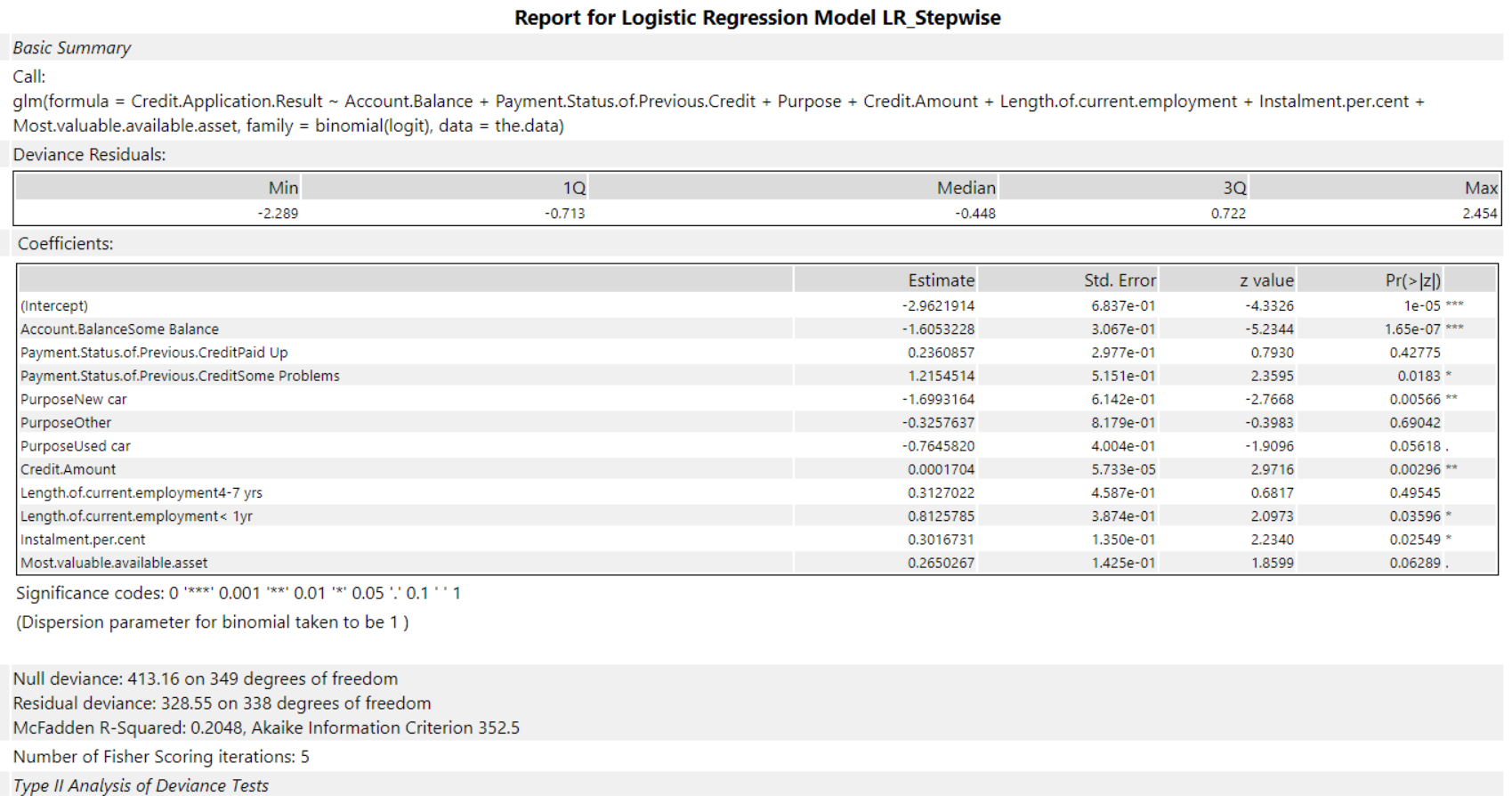
****

****

# **Step 3: Train your Classification Models**

1. **Logistic Regression (Stepwise):**

From below Chart it can be observer that the most important predictor variables for the logistic regression(stepwise) model are Account Balance Some Balance, Purpose New car and Credit Amount. The p-value of all these variables can be observed from below chart.



Below is the model comparison report for Logistic Regression (Stepwise) which shows that this model has an accuracy of 76%. Using Confusion Matrix,

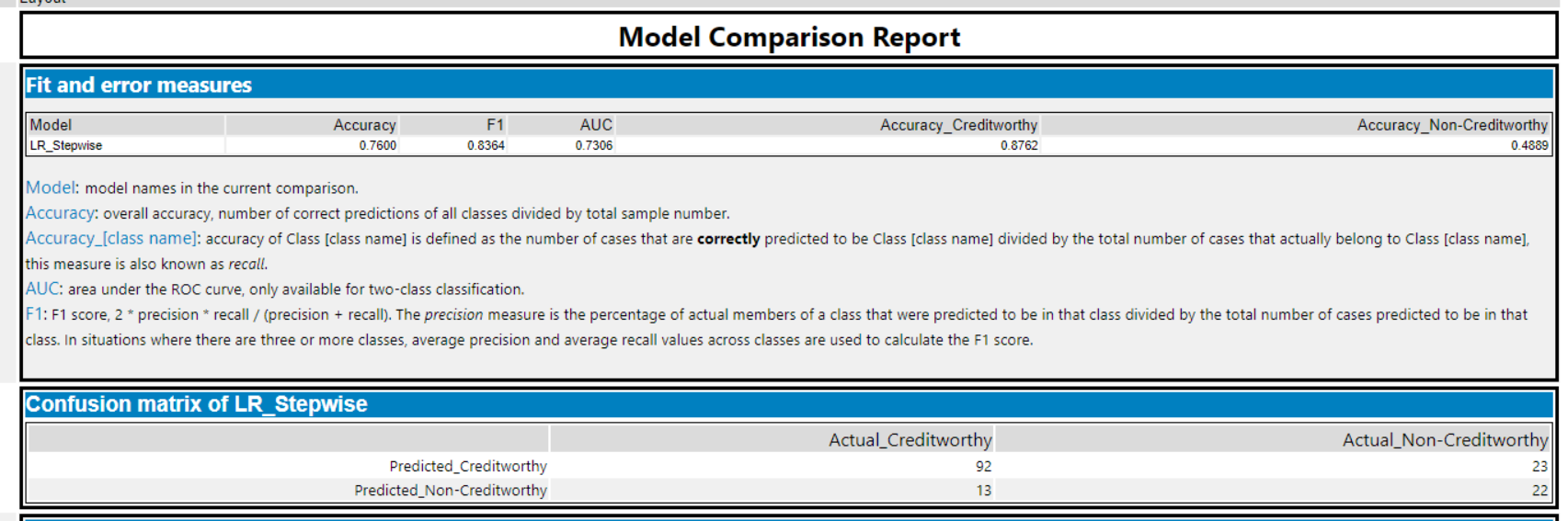
**Accuracy for creditworthy** = (actual creditworthy) / (predicted creditworthy)

= 92/ (92+23) = 0.8 = 80%

**Accuracy for non-creditworthy** = (actual non-creditworthy) / (predicted non-creditworthy)

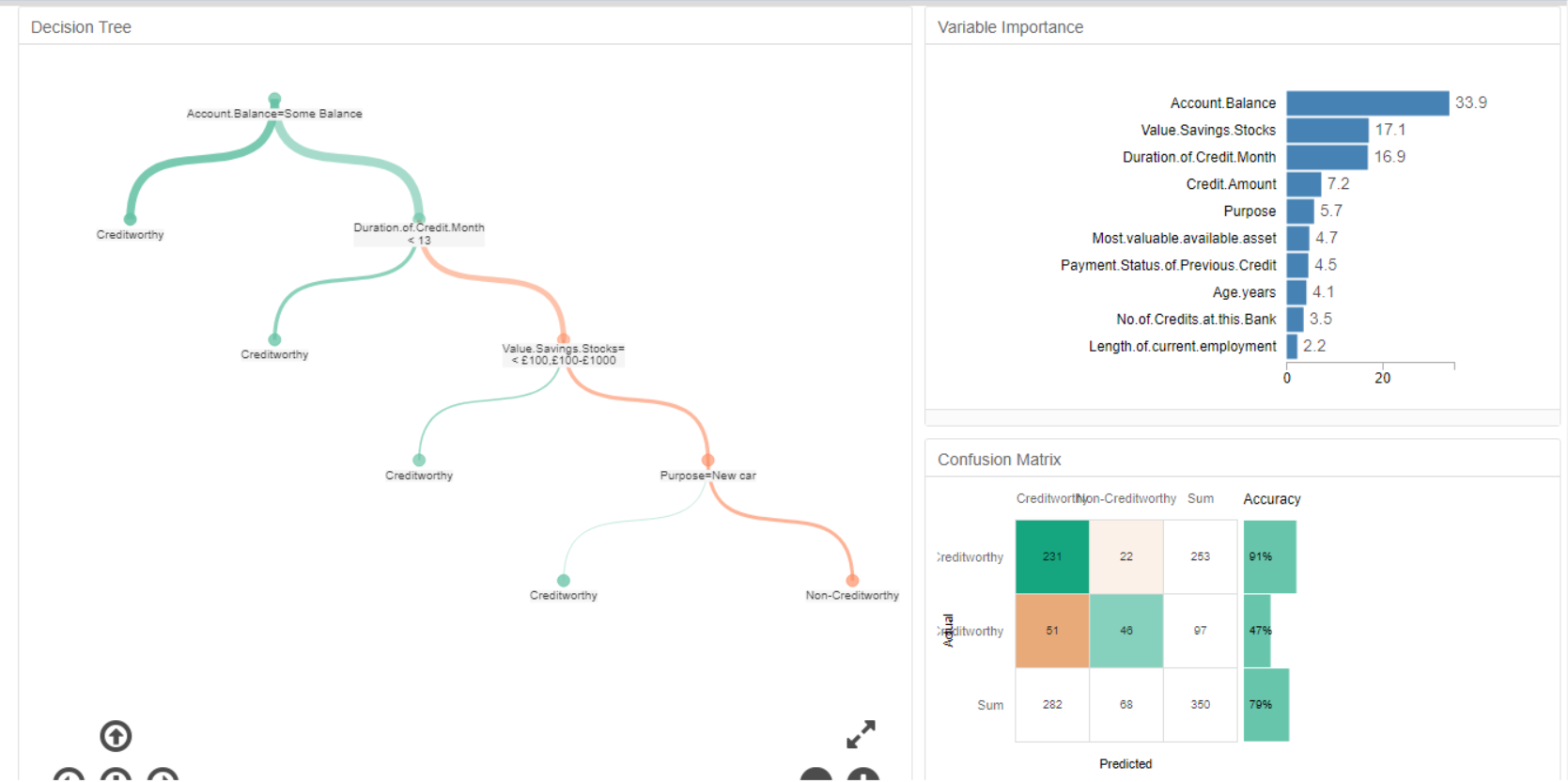
= 22/ (13+22) = 0.6286 = 62.86%

The models seem to be slightly biased towards predicting customer as non-creditworthy.



1. **Decision Tree**

The significant predictor variable for Decision Tree model can be observed from the variable important graph and these are: Account Balance, Value Savings Stocks and Duration of Credit Month.



Below is the model comparison report for Decision Tree which shows that this model has an accuracy of 74.67%. Using Confusion Matrix,

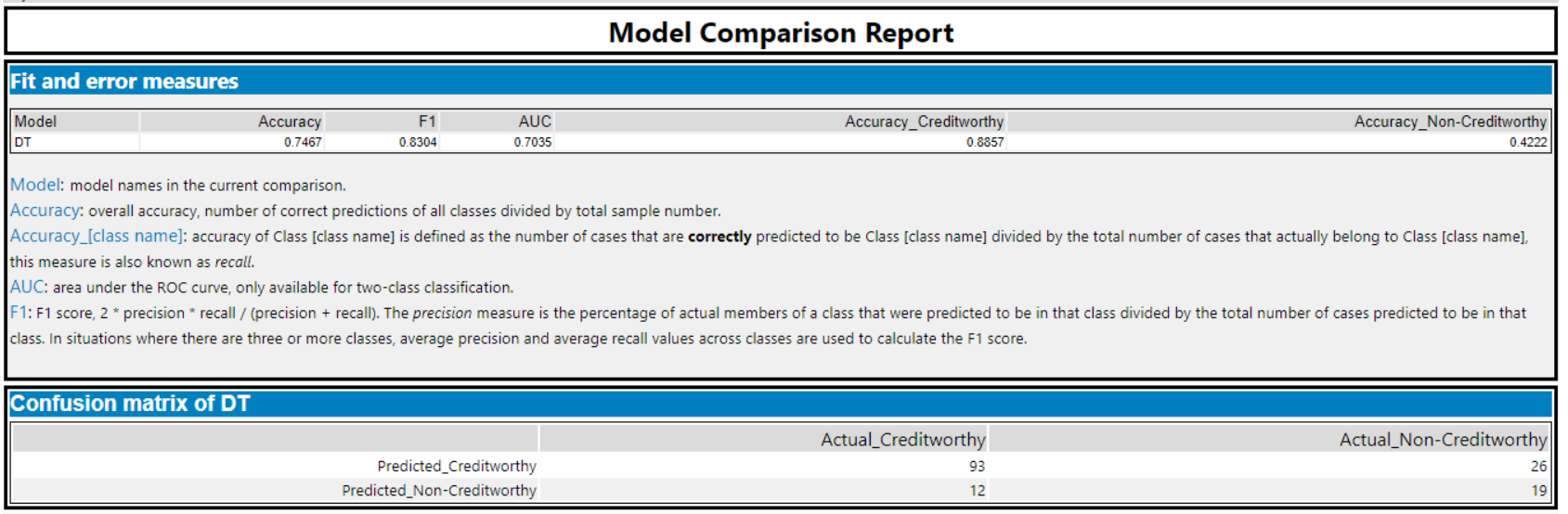
**Accuracy for creditworthy** = (actual creditworthy) / (predicted creditworthy)

= 93/ (93+26) = 0.7815 = 78.15%

**Accuracy for non-creditworthy** = (actual non-creditworthy) / (predicted non-creditworthy)

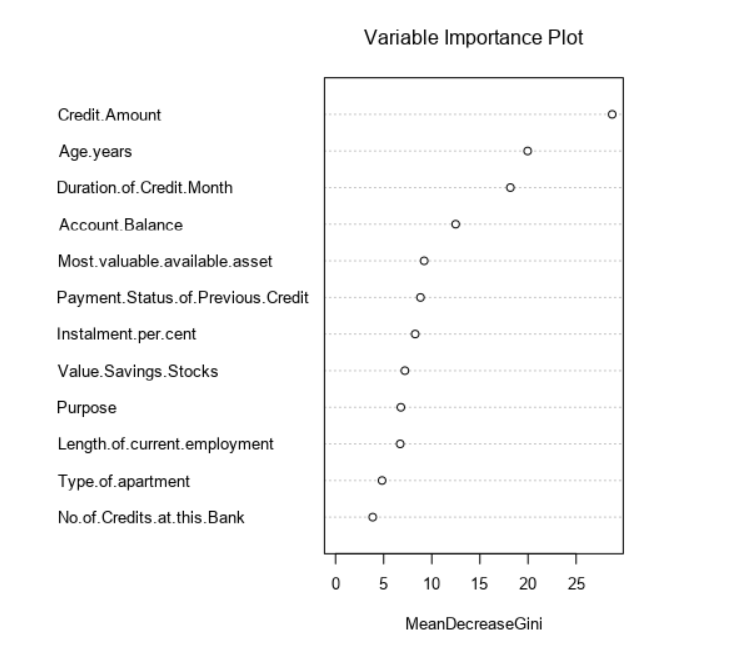
= 19/ (12+19) = 0.6129 = 61.29%

The models seem to be slightly biased towards predicting customer as non-creditworthy.



1. **Forest Model:**

From the Variable importance plot of the forest model, we can infer that the most important predictor variables for this model are Credit Amount, Age Years and Duration of Credit Month.



Below is the model comparison report for Forest Model which shows that this model has an accuracy of 79.33%. Using Confusion Matrix,

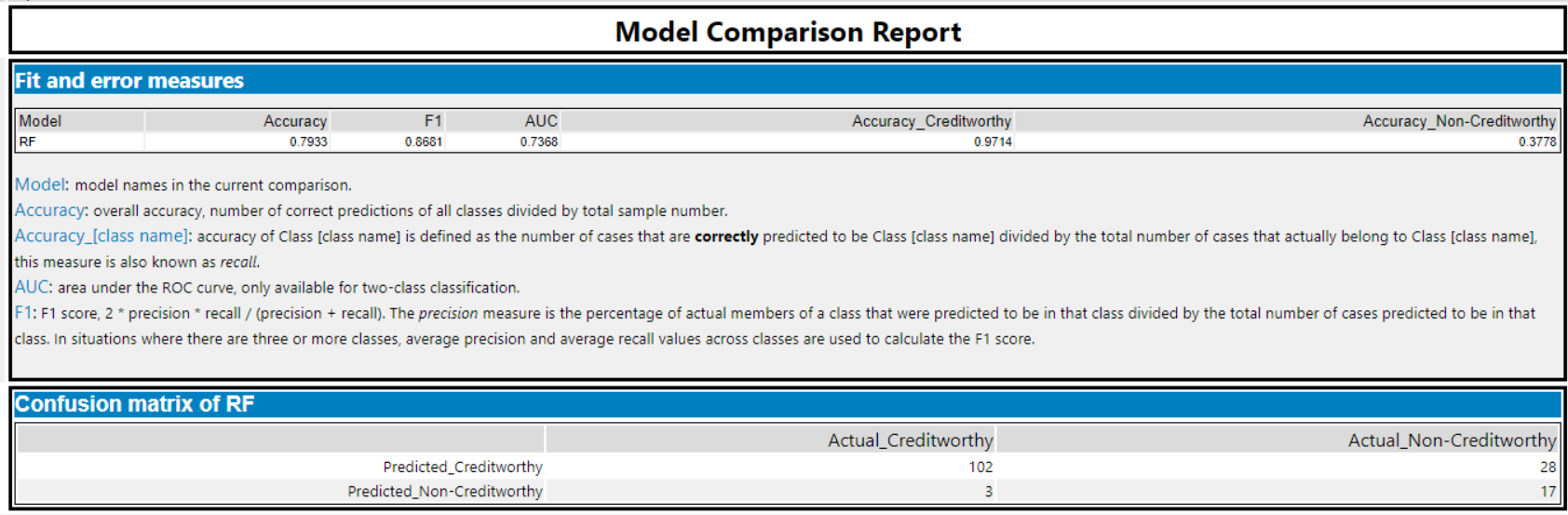
**Accuracy for creditworthy** = (actual creditworthy) / (predicted creditworthy)

= 102/ (102+28) = 0.7846 = 78.46%

**Accuracy for non-creditworthy** = (actual non-creditworthy) / (predicted non-creditworthy)

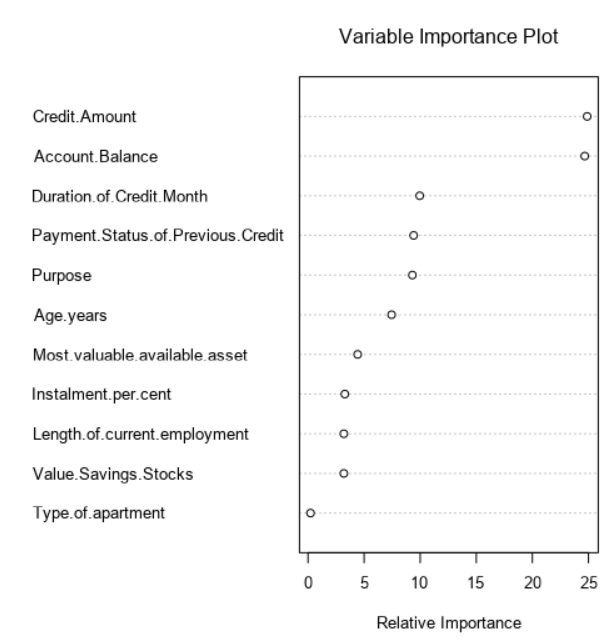
= 17/ (3+17) = 0.85 = 85%

Since accuracies for creditworthy and non-creditworthy are comparable 78.46% and 85% respectively, this model isn’t biased.



1. **Boosted Model:**

From the Variable importance plot of the boosted model, we can infer that the most important predictor variables for this model are Credit Amount, Account Balance and Duration of Credit Month.



Below is the model comparison report for Boosted Model which shows that this model has an accuracy of 78.67%. Using Confusion Matrix,

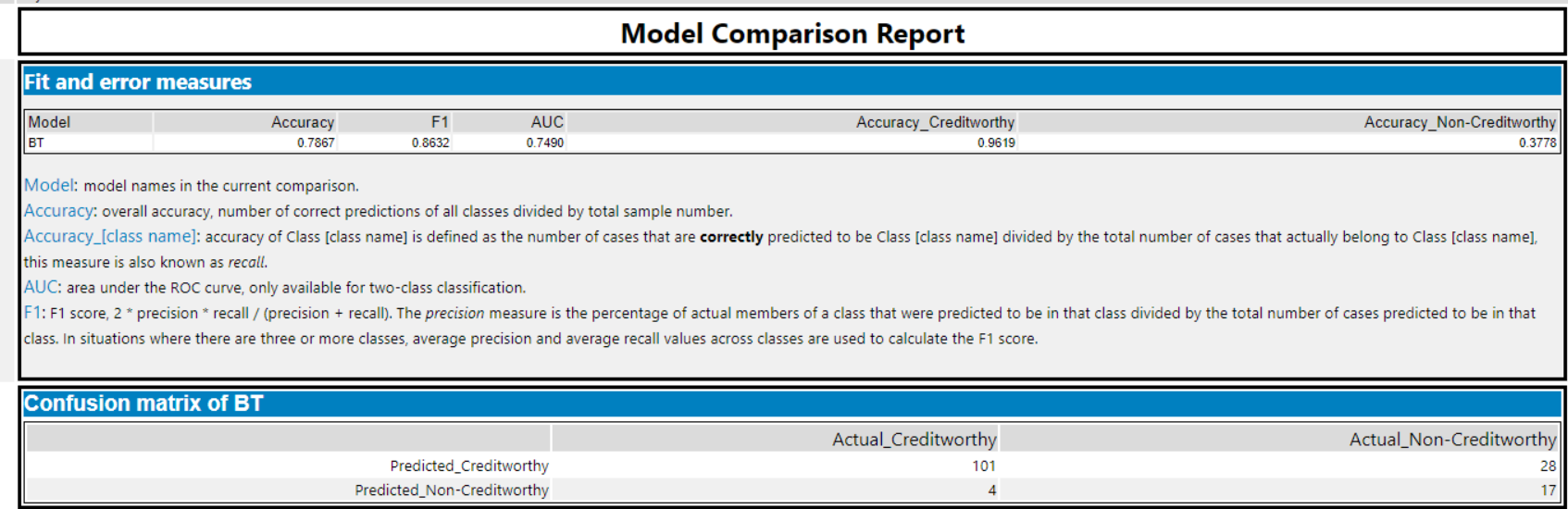
**Accuracy for creditworthy** = (actual creditworthy) / (predicted creditworthy)

= 101/ (101+28) = 0.7829 = 78.29%

**Accuracy for non-creditworthy** = (actual non-creditworthy) / (predicted non-creditworthy)

= 17/ (4+17) = 0.6129 = 80.95%

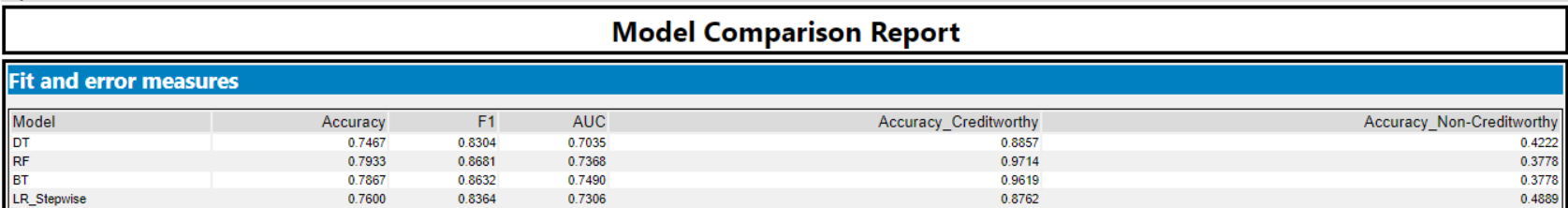
Since accuracies for creditworthy and non-creditworthy are comparable 78.29% and 80.95% respectively, this model isn’t biased.



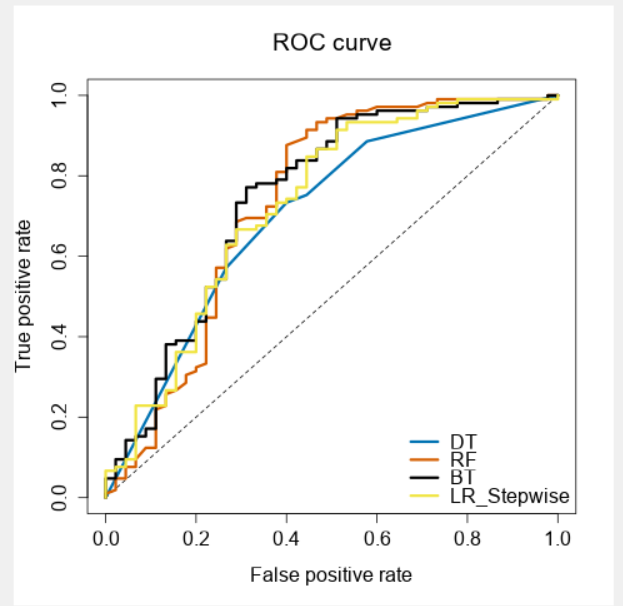
# **Step 4: Writeup**

1. **Which model did you choose to use? Please justify your decision using all of the following techniques. Please only use these techniques to justify your decision:**
   * **Overall Accuracy against your Validation set**
   * **Accuracies within “Creditworthy” and “Non-Creditworthy” segments**
   * **ROC graph**
   * **Bias in the Confusion Matrices**

Forest Model has been chosen since it has the highest accuracy of 79.33%among all four classification models. Also, the accuracies for creditworthy and non-creditworthy are among the highest of all.



Forest reaches the top true positive rate the quickest and overall, the highest the most.



Using the confusion matrix of the Forest Model:

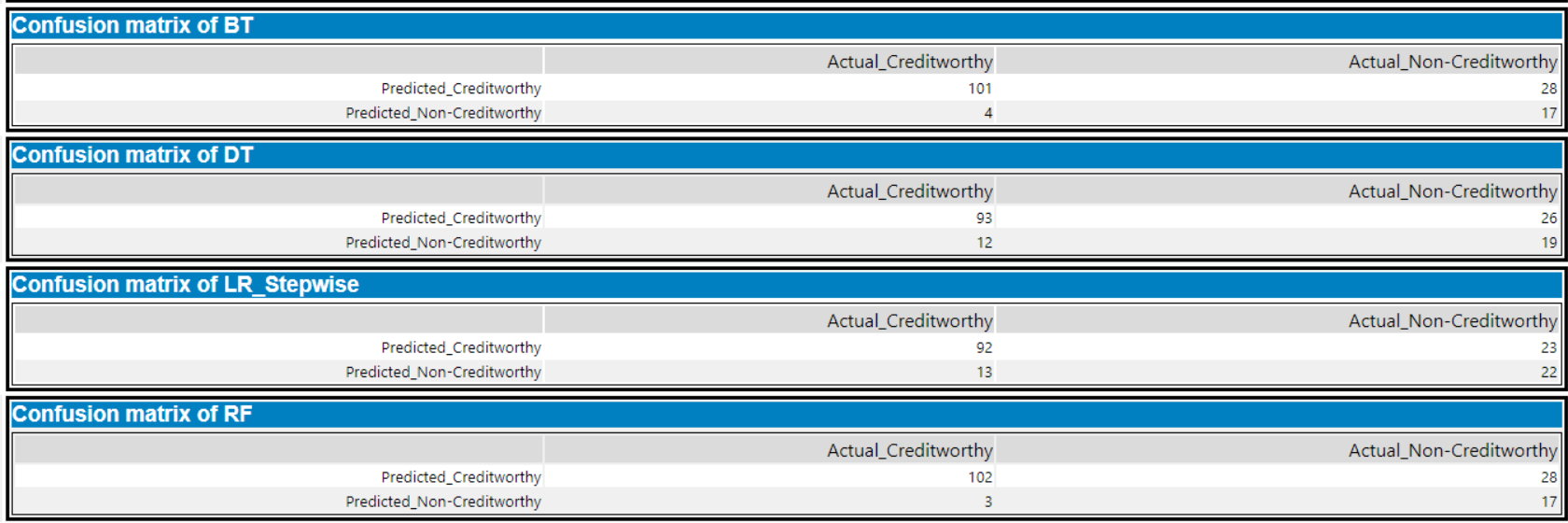
**Accuracy for creditworthy** = (actual creditworthy) / (predicted creditworthy)

= 102/ (102+28) = 0.7846 = 78.46%

**Accuracy for non-creditworthy** = (actual non-creditworthy) / (predicted non-creditworthy)

= 17/ (3+17) = 0.85 = 85%

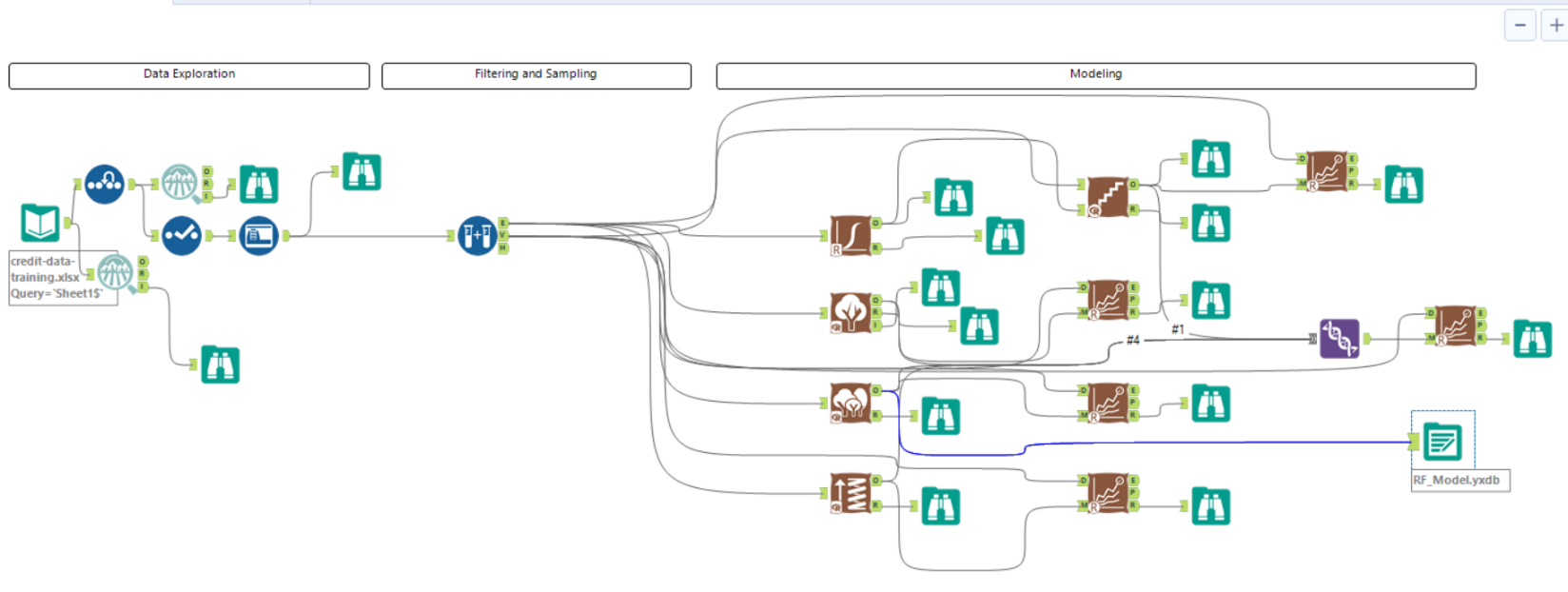
Since accuracies for creditworthy and non-creditworthy are comparable 78.46% and 85% respectively, this model isn’t biased.

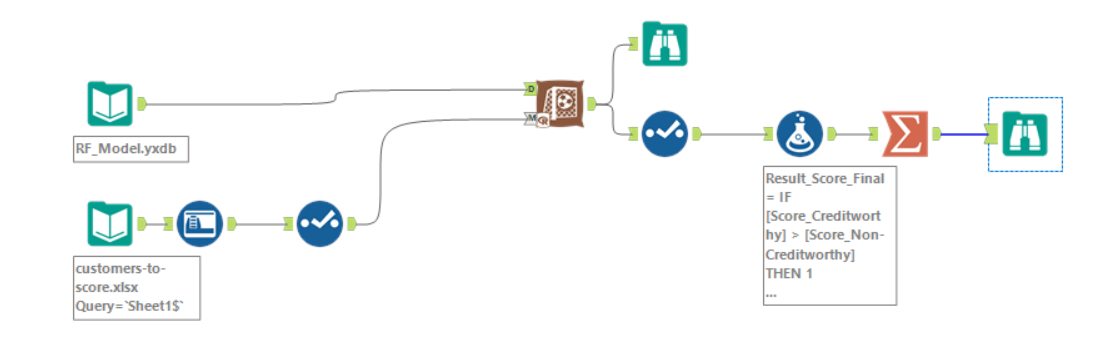


1. **How many individuals are creditworthy?**

There are 410 creditworthy new customers that we could approve for a loan and 92 non-creditworthy customers that should not be approved for a loan.

**My Alteryx Workflow**

****

****