Project: Creditworthiness

# Step 1: Business and Data Understanding

## Key Decisions:

Answer these questions

**What decisions needs to be made?**

The decisions that need to be made are:

We need to know whether or not to approve a loan application.

We also need to know which model is the best for deciding this.

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

We need historic data on loan applications and the outcome of each loan application. This data will need to include enough information to create variable to be used to train the model.

We also need the same data for the new applications.

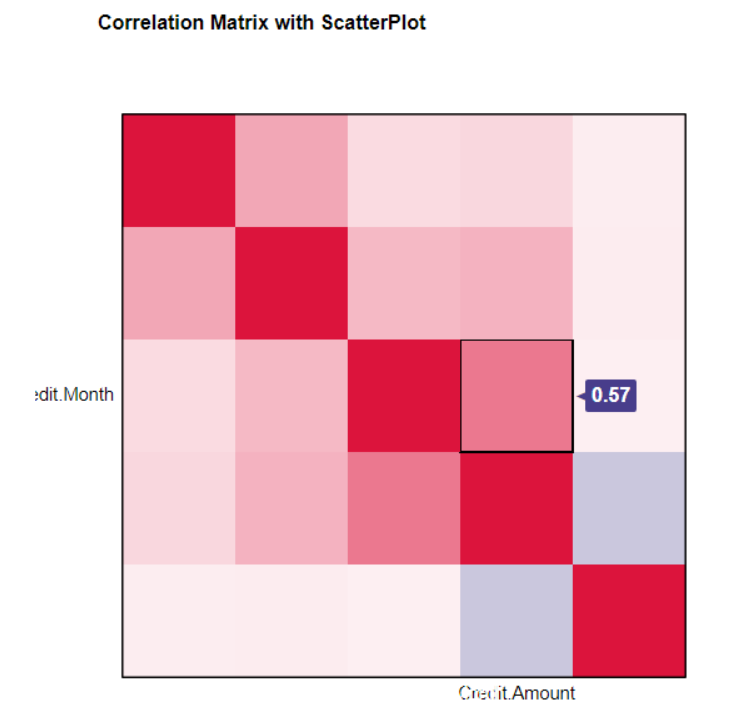
**What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

We need to use a Binary model as the main question to answer is yes or no on whether or not to approve a loan.

# 

# Step 2: Building the Training Set

**For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.**



Based on association analysis, no two fields are highly correlated, I.e. >0.7.

**Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed**

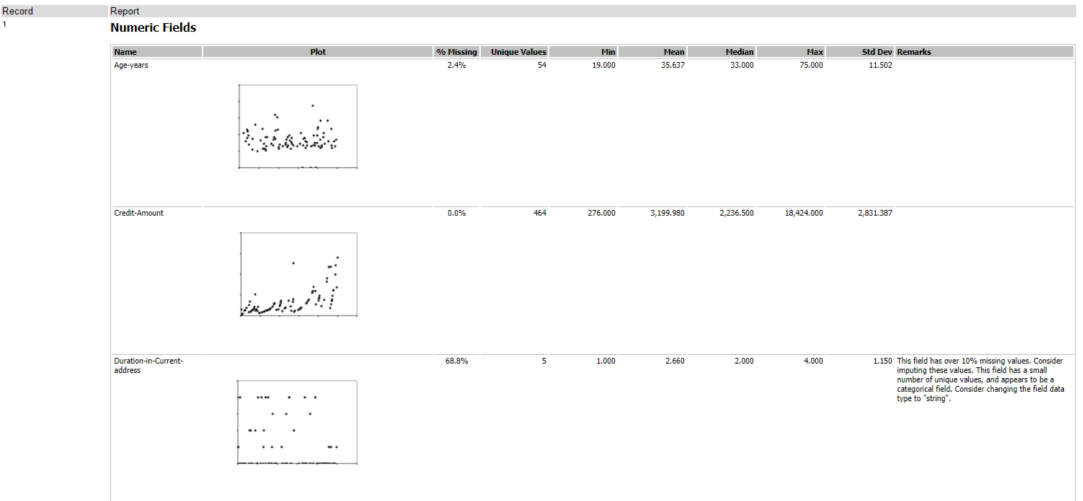


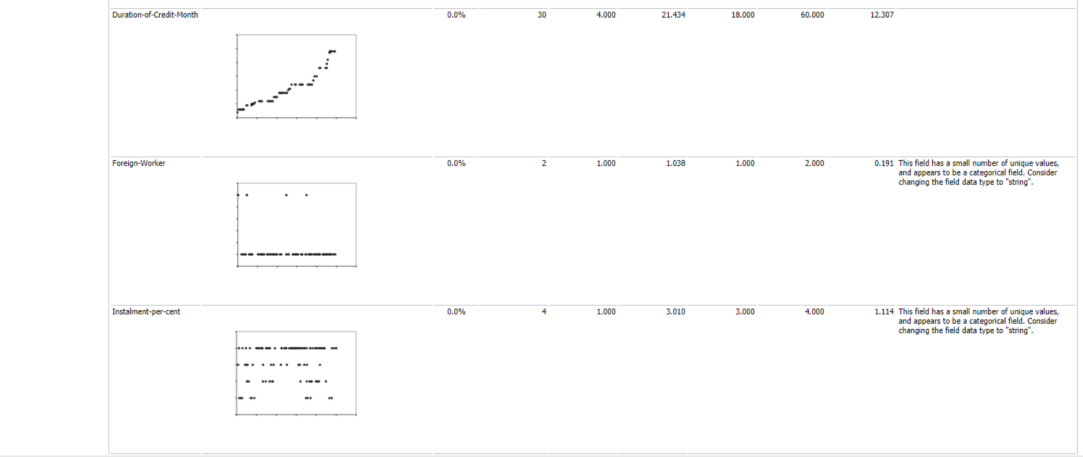


Duration.in.Current.address has 69% missing values so this field will be removed.

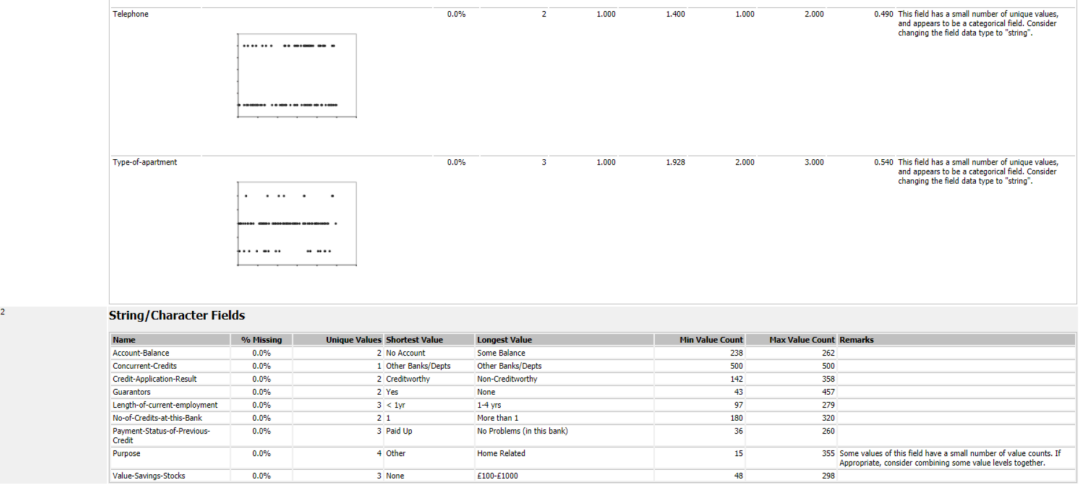
Age-years has 2% missing values so will not be removed but rather imputed with the median Age-years replacing the 2% missing values.

**Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.**









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

The following data fields have low variability and will be removed:

Foreign worker, no of dependents, occupation, concurrent credits and guarantors.

Telephone has also been removed based on its irrelevance to the target.

Age-years has been imputed with the median Age-years due to the small number of missing values.

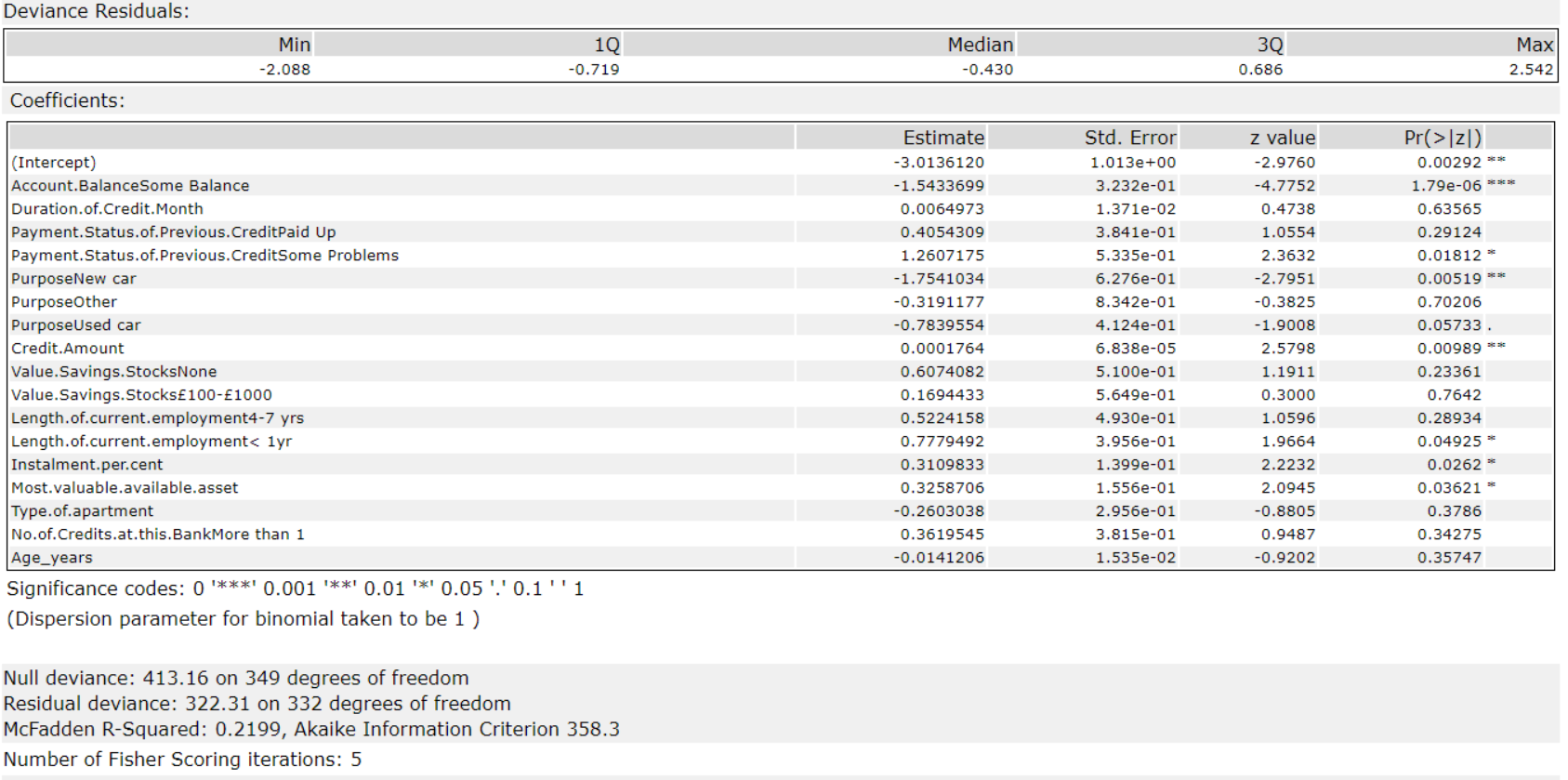
In total I have removed the following seven variables:

Duration in current address, Foreign worker, no of dependents, occupation, concurrent credit, guarantors and telephone.

# 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.**

**Logistic Regression**



Using logistic regression the following variables were found to be statistically significant.

Account.Balance

Payment.Status.of.Previous.Credit

Purpose

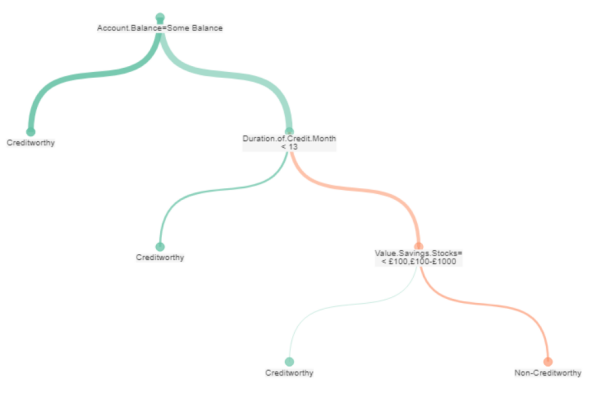
Credit.Amount

Length.of.current.employment

Installment.per.cent

Most.valuable.available.asset

**Decision Tree**



Using a decision tree the following variables were found to be important:

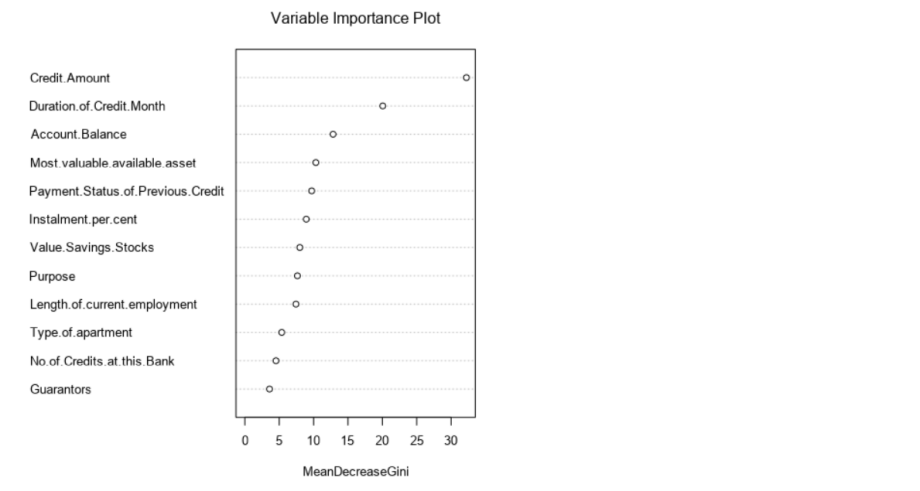
Account.Balance

Duration.of.Credit.Month

Value.Savings.Stocks

Added Credit.Amount due to importance in Decision Forest and Boosted models.

**Decision Forest**



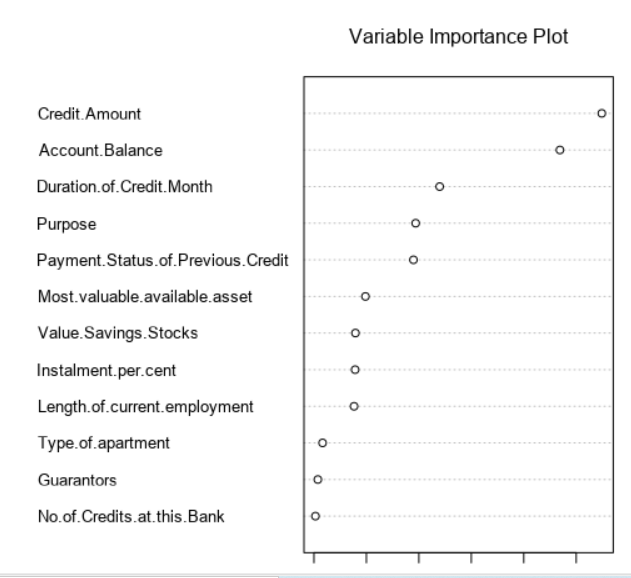
Using a decision forest model, all variables were found to be important to the model. However, the top three were:

Credit.Amount

Duration.of.Credit.Month

Account.Balance

**Boosted Model**



Using a boosted model, all variables contributed however the following three were the most important:

Credit.Amount

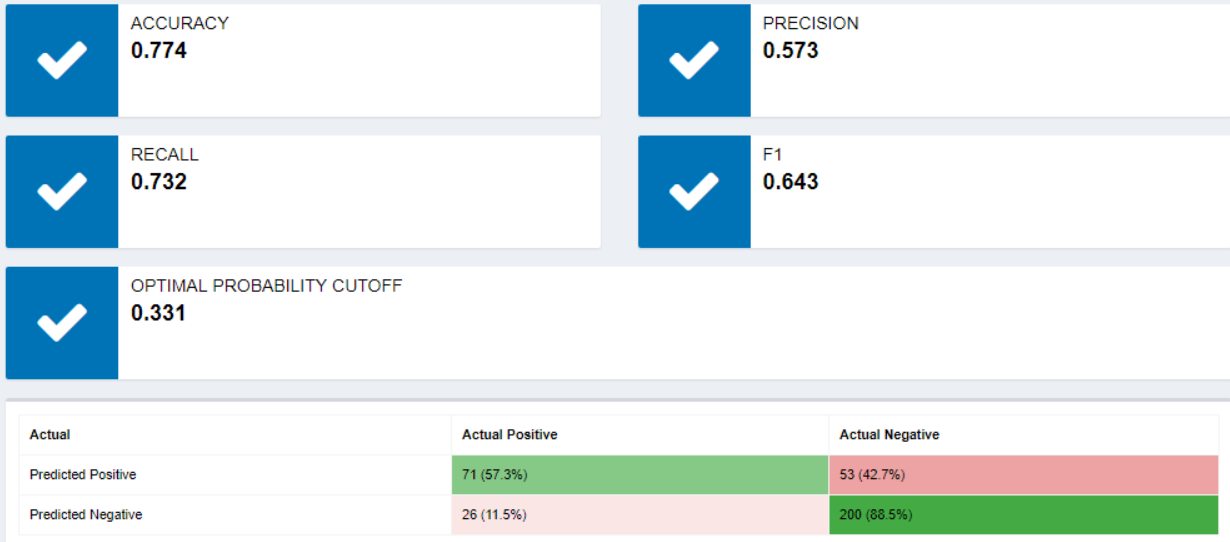
Account.Balance

Duration.of.Credit.Month

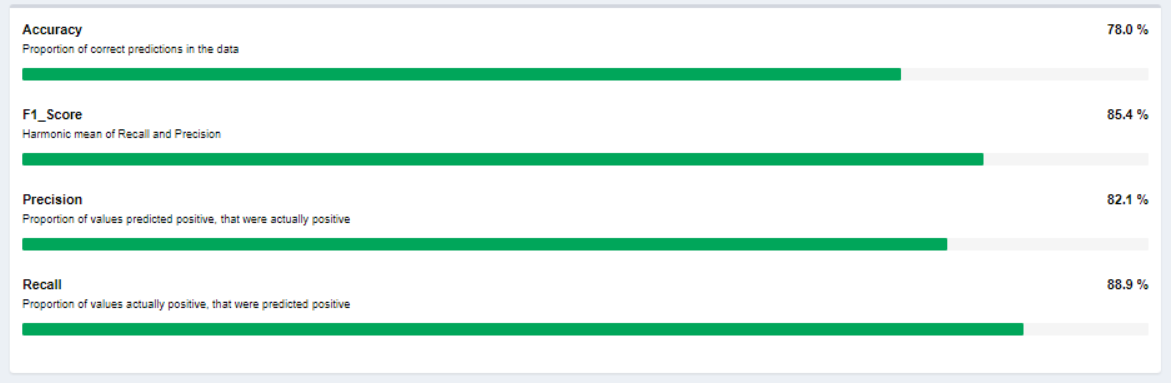
**Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?**

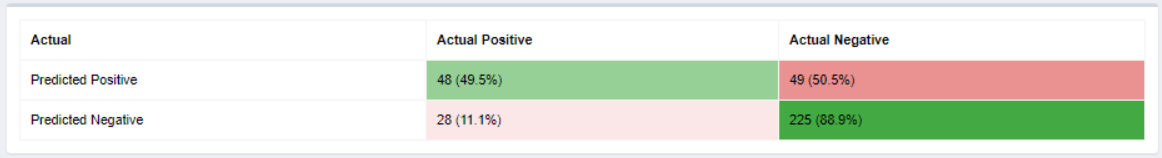
Using the statistically significant variables, a model was created of each of the four types below and tested against the 30% validation set.

*Logistic Regression:*

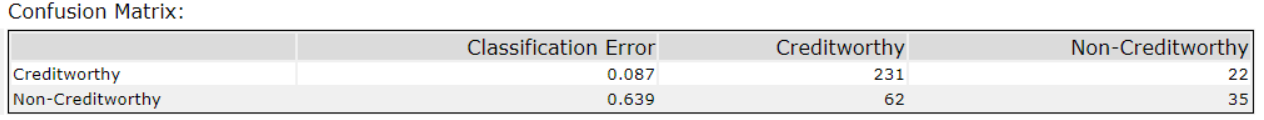


*Decision Tree:*

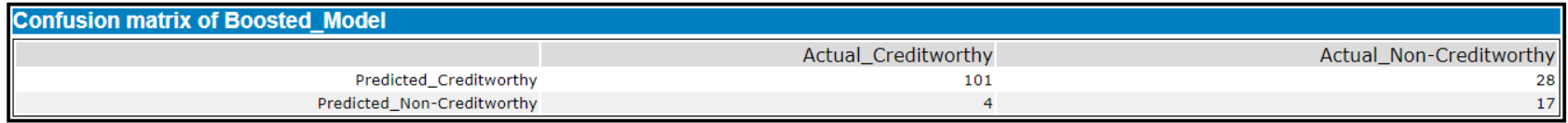




*Decision Forest:*



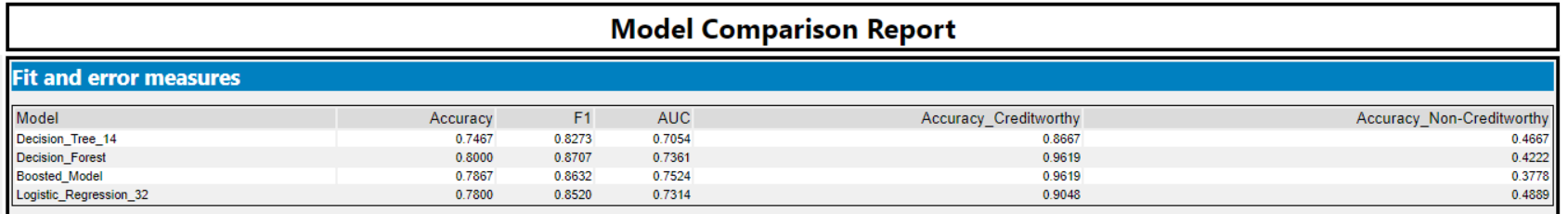
*Boosted Model:*

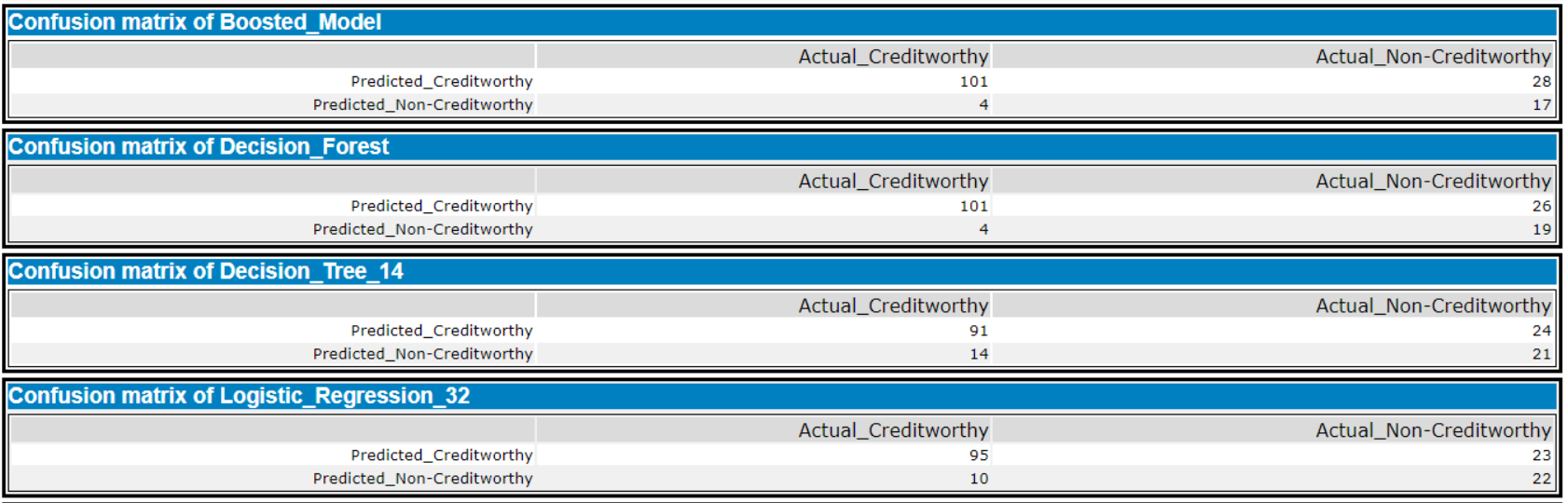


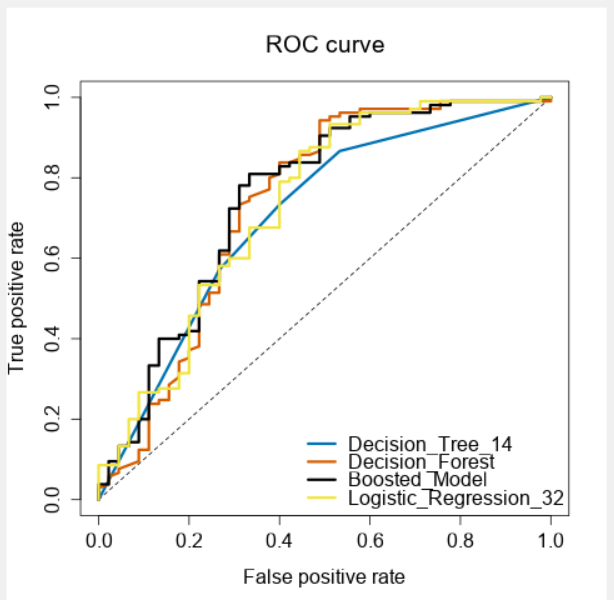
All models are bias toward Actual\_Negative showing much higher accuracy when predicting Negative values while the accuracy when predicting Positive values is much lower.

# Step 4: Writeup

*Decide on the best model and score your new customers. For reviewing consistency, if Score\_Creditworthy is greater than Score\_NonCreditworthy, the person should be labeled as “Creditworthy”*







*Answer these questions:*

**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:**

The model I have chosen to use is the forest model. This model has the highest overall accuracy, the third highest AUC from the ROC curve, the second highest accuracy when predicting non-creditworthy and the highest accuracy when predicting creditworthy applicants.

**Overall Accuracy against your Validation set**

Overall the decision forest model was the most accurate with an overall accuracy of 80.00%.

The decision tree, boosted and logistic regression models were 74.67%, 78.67% and 78% respectively.

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

All models showed a higher accuracy predicting creditworthy with accuracies higher on Actual\_Creditworthy applicants. The logistic regression model had the highest accuracy when predicting negative applicants with an accuracy of 48.89%.

**ROC graph**

The ROC graphs shows that the boosted model is the best model when using ROC as a metric. The AUC of the boosted model is 75.24% which is higher than the other models.

**Bias in the Confusion Matrices**

All models show a bias toward predicting creditworthy applicants which is a bit of a worry as this means several actual non-creditworthy applicants are being classified as creditworthy. The boosted model is the most biased while the other models are all similarly biased.

**How many individuals are creditworthy?**

I have decided on a Forest model as the best model I retrained the Forest model on the entire dataset and ran the new data through.

Based on the new model with the new data, **408** individuals are creditworthy.

Alteryx Workflow

