Project: Creditworthiness

# Business and Data Understanding

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

In this case, predicting the possibility of the new 500 loan applications to be processed this week. That means determine whether the customers are creditworthy to give a loan?

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

1. Data on all past applications which contains customer credit history, capital, loan purpose, capital, occupation and age.

2. The list of customers that need to be processed in the next few days.

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

Binary model is the best model. Since in this case we are classifying the elements of a given data into two groups (creditworthy or non-creditworthy).

# Building the Training Set

*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn’t* ***need to convert any data fields to the appropriate data types.***

*Here are some guidelines to help guide your data cleanup:*

* 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”.
* Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
* 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.
* Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

***Note:*** *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

***Note:*** *For students using software other than Alteryx, please format each variable as:*

|  |  |
| --- | --- |
| **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 |

*To achieve consistent results reviewers, expect.*

*Answer this question:*

* **In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.**
* **Age-years** and **Duration-in-Current-address** fields contains 2% and 69% of missing values respectively.
* For the field **Age-years**, missing value is very low, imputed the missing data with

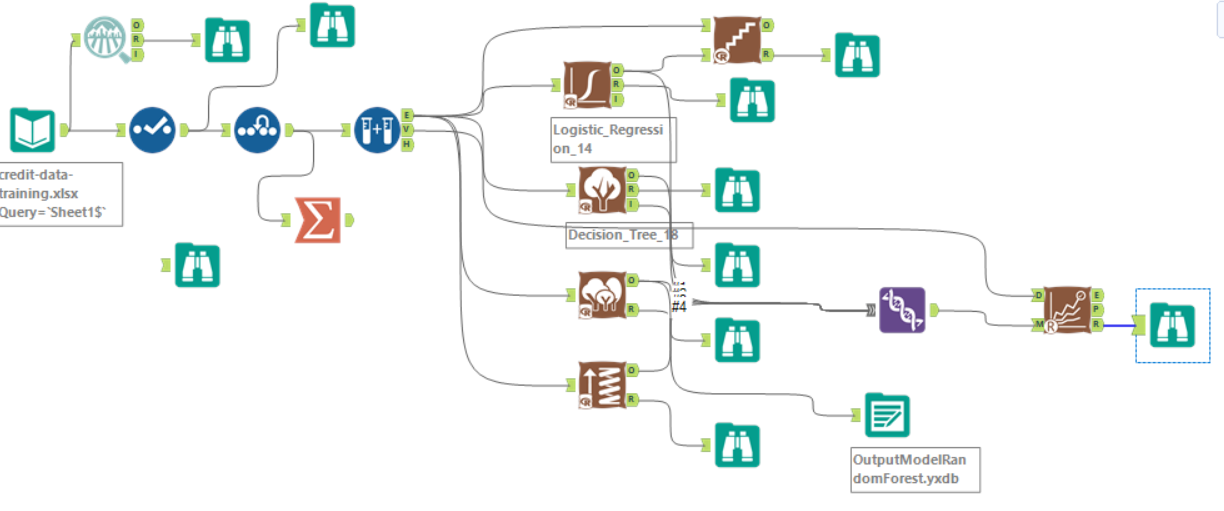
Median of the field Age-years.

* Since there is a large number of missing data, **Duration-in-Current-address**is removed.
* The low variability fields **Concurrent Credit**and **Occupation** only have one value, so we remove them as well.
* The string field **Guarantors** have only 2 number of unique values i.e Yes and None. Yes has 43 values and None has 457 values. This seem low variability field, so we remove this as well
* The numeric fields **Foreign Worker**, **NO. of dependents** and **Telephone** have a very small number of unique values. They seem to be categorical fields with low variability, so they are removed as well.

# Train your Classification Models

*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.*

*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

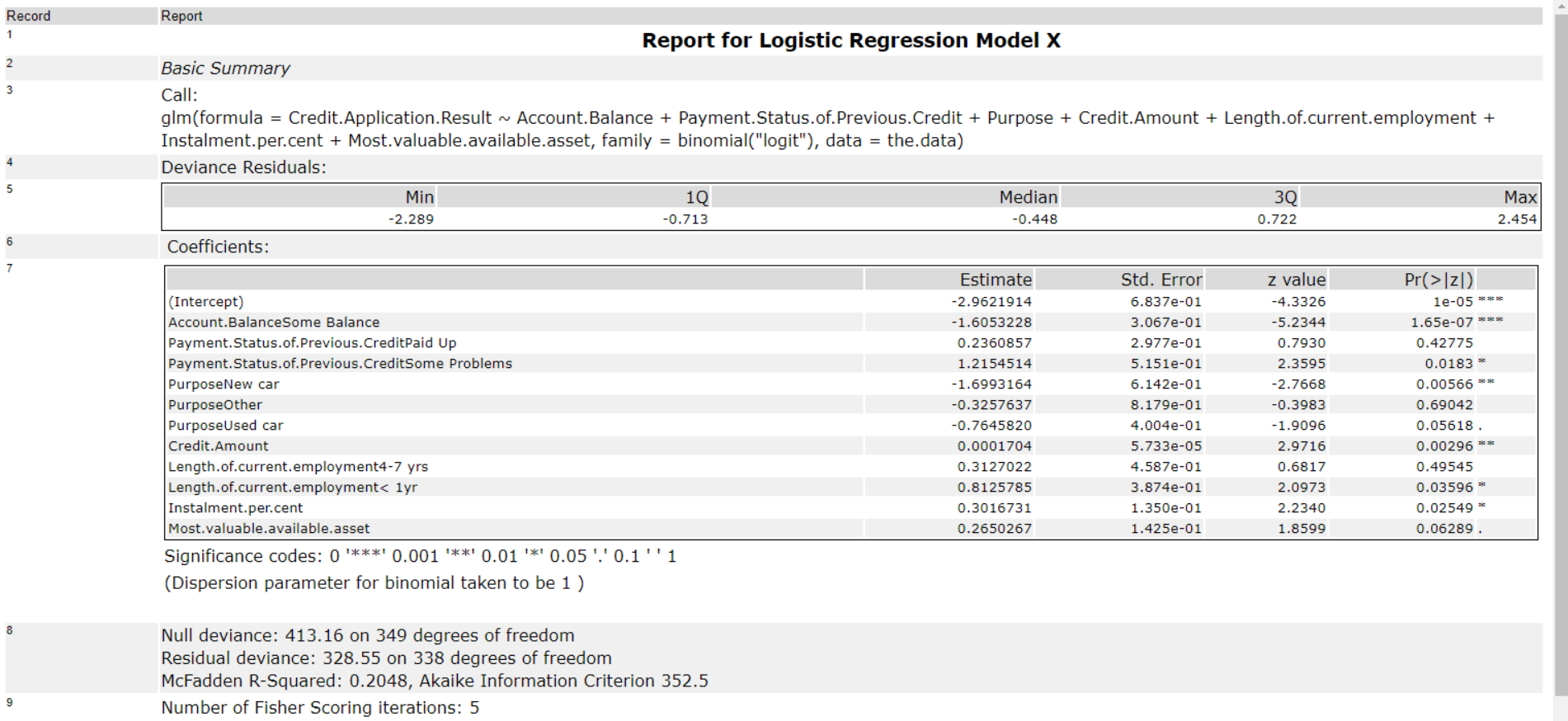


*Answer these questions for* ***each model*** *you created:*

* **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**

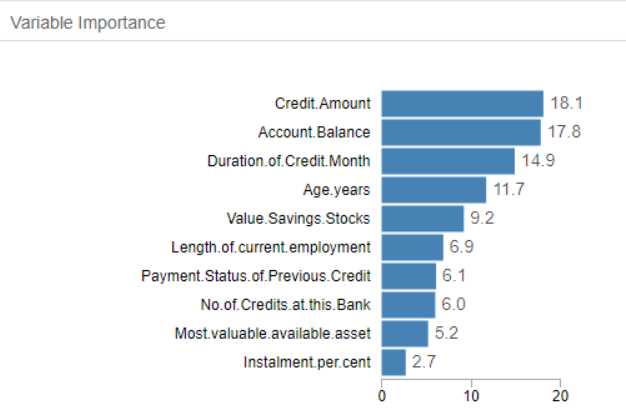
1. **Logistic Regression Stepwise:**

Based on the report, the significant predictive variables are **Account Balance**, **Payment Status of previous credit, Credit Amount and Length of current employment.**



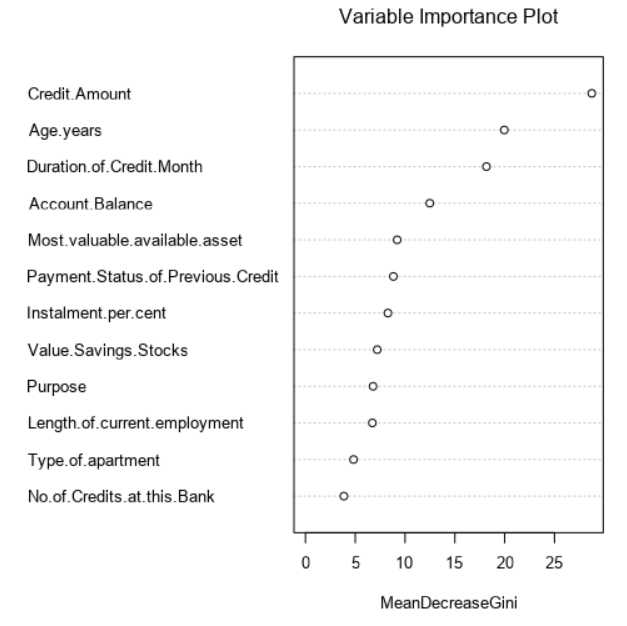
1. **Decision trees:**

Based on variable importance plot, the top 3 significant variables are **Account Balance, Duration of Credit Month** and **Credit Amount.**



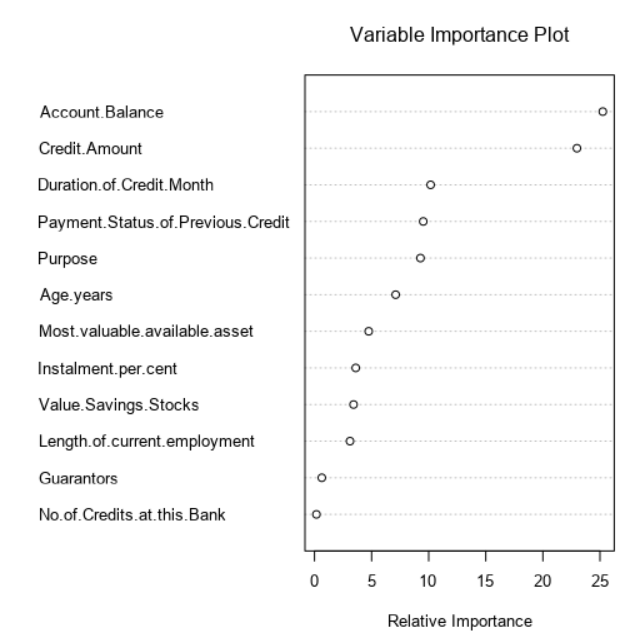
1. **Forest Model:**

Based on variable importance plot, the top 3 significant variables are **Credit Amount, Age.years and Duration of credit Month.**



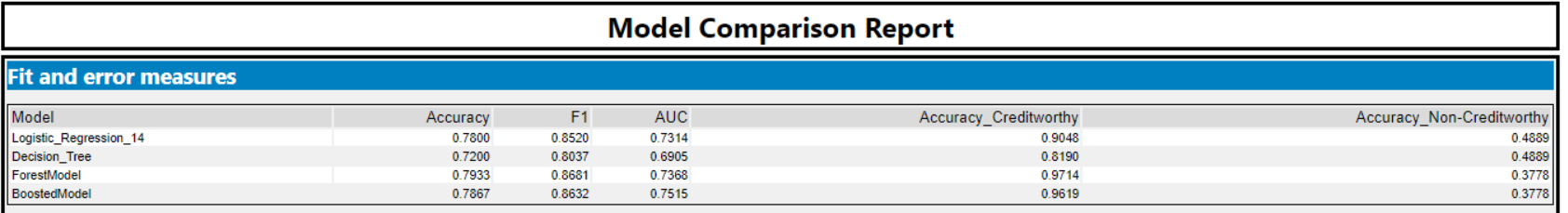
1. **Boosted Model:**

Based on variable importance plot, the top 3 significant variables are **Account Balance, Credit Amount and Duration of credit Month.**



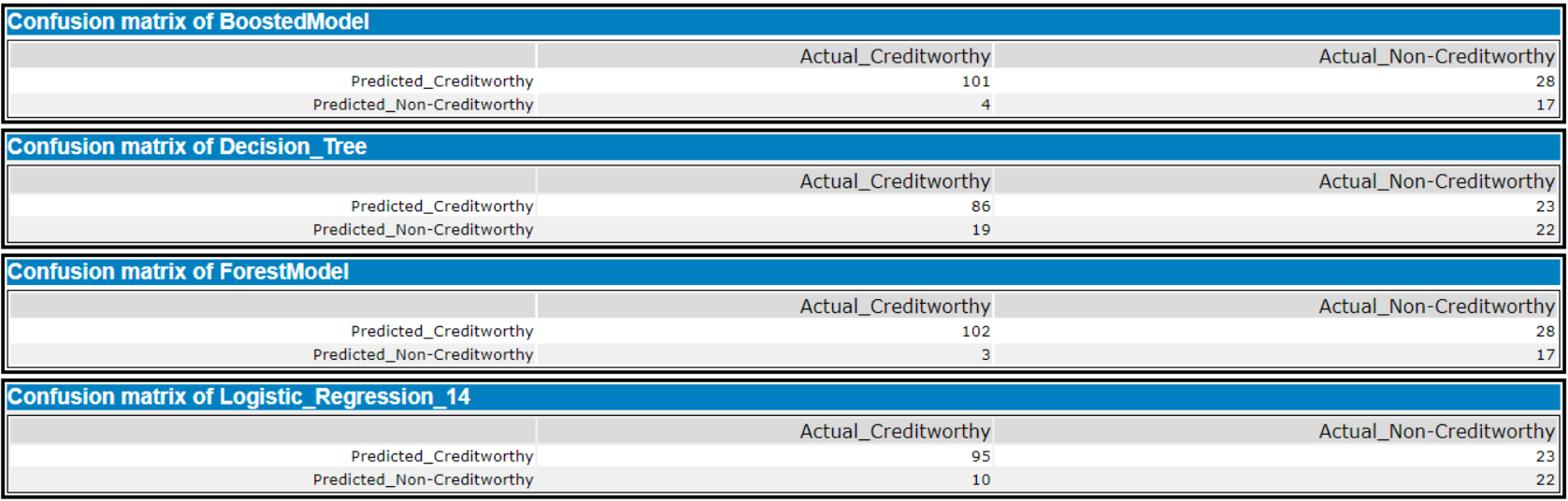
* **Validate** **your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?**

We can validate and compare the accuracy of the models using model comparison tool.



Based on the model comparison, the overall accuracy for Logistic Regression stepwise model is 0.78, the overall accuracy for Decision Tree model is 0.72, the overall accuracy for Random Forest model is 0.79 and the overall accuracy for Boosted model is 0.78.

Below is the confusion matrix for all models.



The overall percent accuracy of the **Logistic model** is 78% which is strong

PPV= true positives \ (true positives + false positives) = 95/ (95+10) =.90  
NPV= true negatives \ (true negatives + false negatives) = 22/ (22+23) = .48  
So after checking confusion matrix there is bias seen in the model's prediction to Creditworthy.

The overall percent accuracy of the **Decision Tree** model is 0.72, which is strong

PPV= true positives \ (true positives + false positives) = 86/ (86+19) =.82  
NPV= true negatives \ (true negatives + false negatives) = 22/ (22+23) = .48  
So after checking confusion matrix there is bias seen in the model's prediction to Creditworthy.

The overall percent accuracy of the **Random Forest** model is 0.79, which is strong

PPV= true positives \ (true positives + false positives) = 102/ (102+3) =.97  
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+28) = .37  
So after checking confusion matrix there is bias seen in the model's prediction to Creditworthy.

The overall percent accuracy of the **Boosted model** is 0.78, which is strong

PPV= true positives \ (true positives + false positives) = 101/ (101+4) =.96  
NPV= true negatives \ (true negatives + false negatives) = 17/ (17+28) = .37  
So after checking confusion matrix there is bias seen in the model's prediction to Creditworthy.

# 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”*

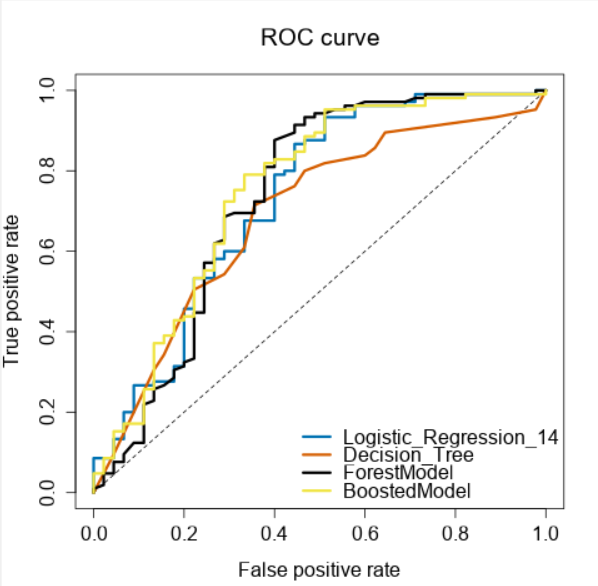
*Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)*

*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:**
  + **Overall Accuracy against your Validation set**
  + **Accuracies within “Creditworthy” and “Non-Creditworthy” segments**
  + **ROC graph**
  + **Bias in the Confusion Matrices**

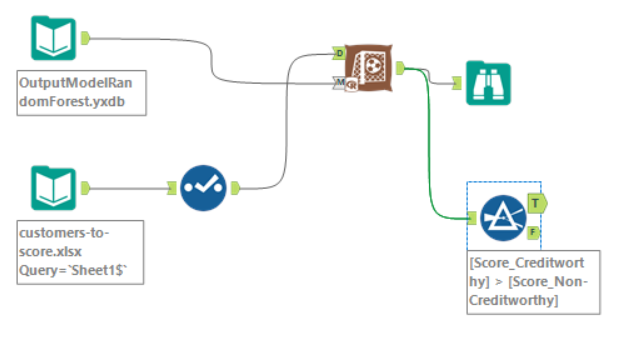
Mixing Overall Accuracy, Accuracies within Creditworthy and Non-Creditworthy values and F1 score. I have chosen **Forest Model** is the best model because this model have higher accuracy.

ROC Graph:



We might say that whose models are biased towards predicting individuals who are creditworthy, as they do not predict individuals who are not creditworthy nearly at the same level as those who are.

* **How many individuals are creditworthy?**



The number of individuals who are creditworthy are 408.