Project: Creditworthiness

# Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

## Key Decisions:

Answer these questions

* What decisions needs to be made?

We are loan officers at a small bank that have come into a large number of new customers. We need to decide whether these customers are creditworthy or not.

* What data is needed to inform those decisions?

To make this decision, we will be looking at the customers that we currently have and what factors made them creditworthy. We will be looking at many different variables given in their past applications.

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

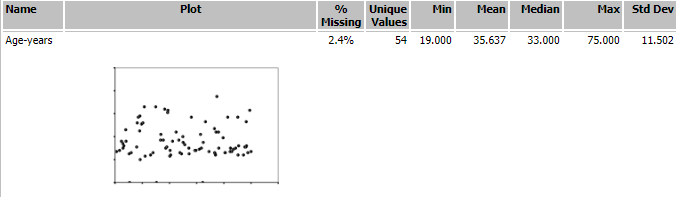
We will be using a binary model as the decisions we are making are whether or not a customer is creditworthy.

# Step 2: Building the Training Set

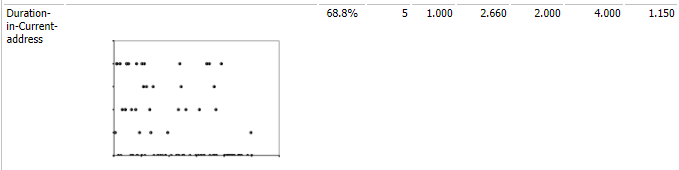
*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn’t* ***need to convert any data fields to the appropriate data types.***

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

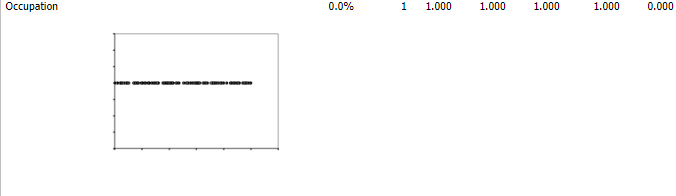
We removed six fields from this data because each field had consisted of mostly one value. These fields were Guarantors, Duration-in-Current-address, Concurrent-Credits, Occupation, No-of-Dependents, and Foreign-Worker. The Telephone field was removed because it was not logical to include this variable. In terms of imputed fields, the Age-years field had a few null values so we imputed those with the median value. The visualizations below will show this information more clearly:



Age-Years is missing 2.4% of data, so we will impute the median according to the project suggestion

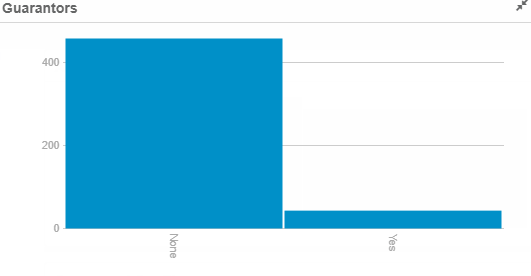


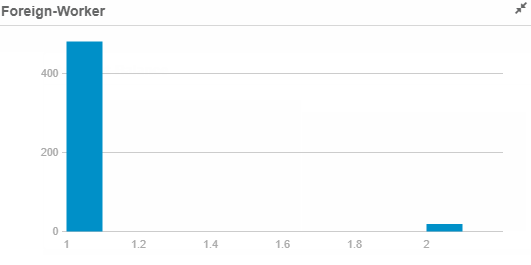
Duration-in-Current-address is missing 68.8% of data, so we will remove this.

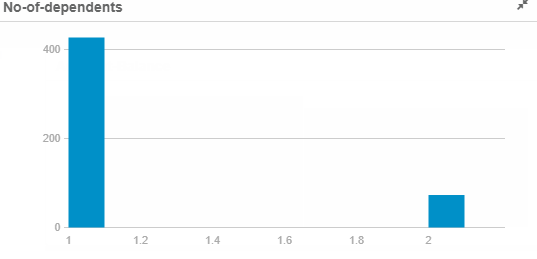




Both Occupation and Concurrent-Credits only had one value, so these are removed for being low-variability







Guarantors, Foreign-Worker, and No-of-dependents were also removed because their values skewed heavily towards one value, making them low-variability as well.

# Step 3: Train your Classification Models

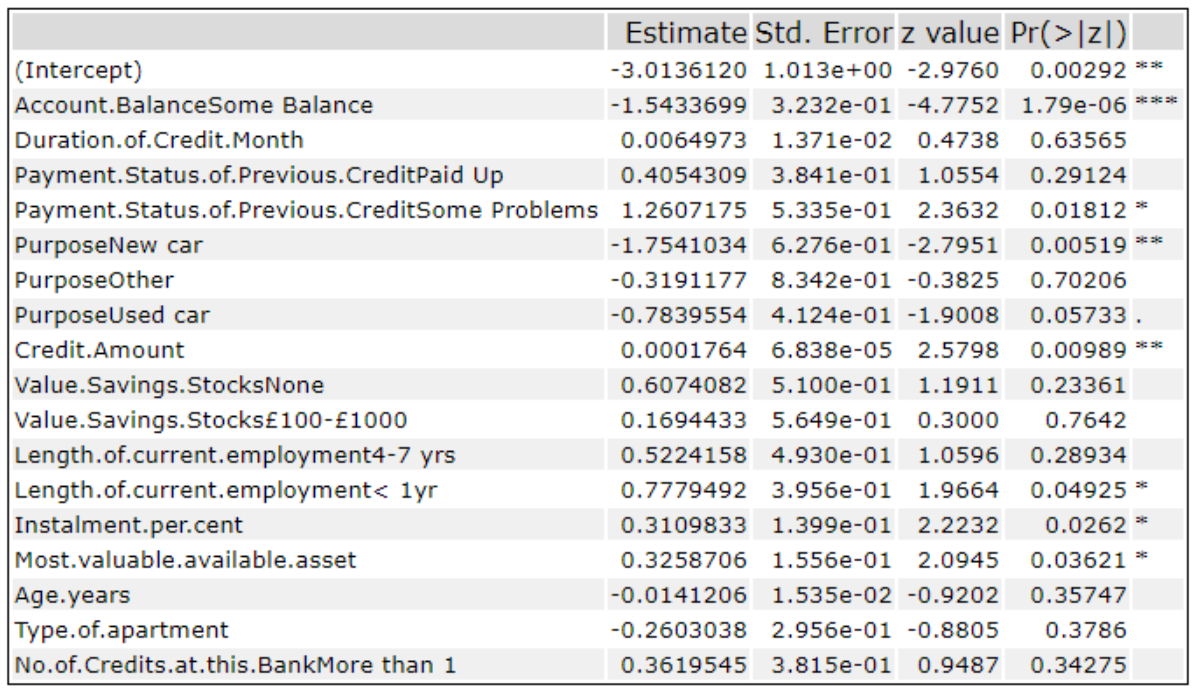
*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.*

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

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

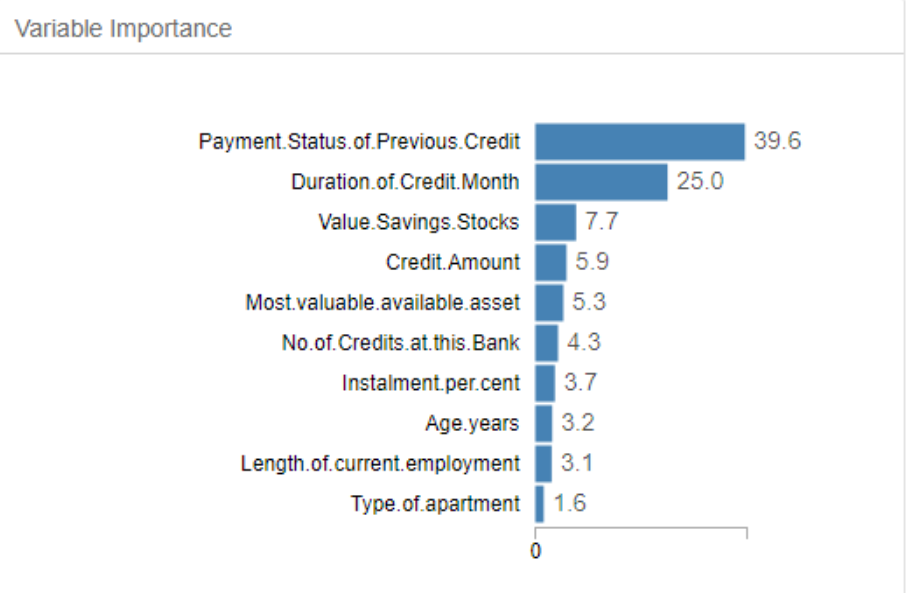
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

**Logistic Regression**



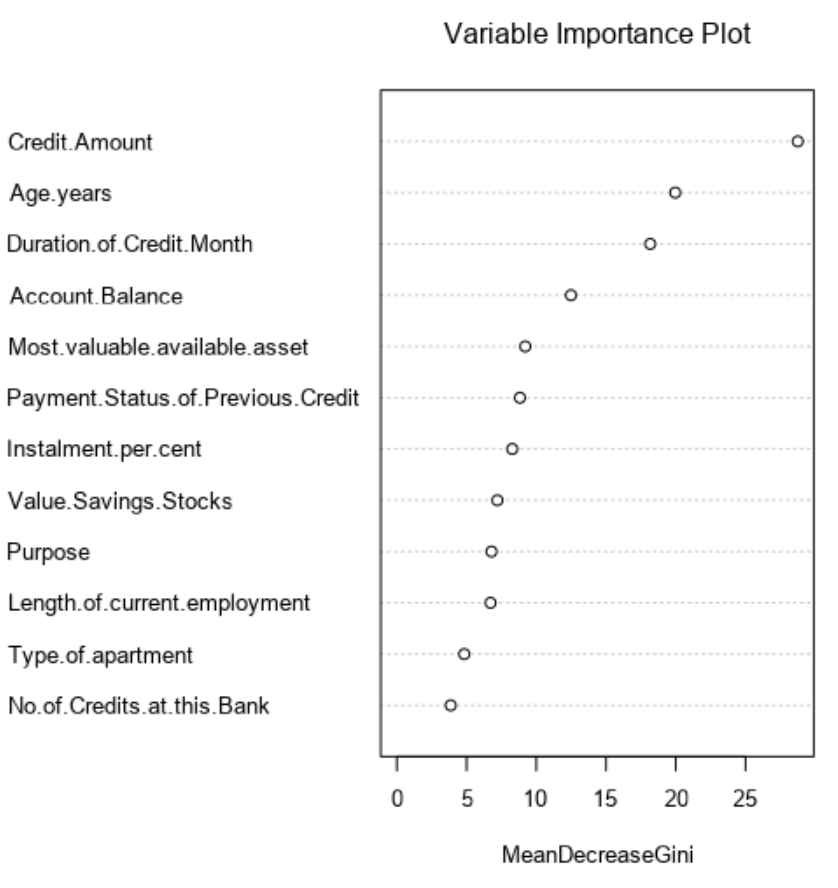
The most significant variable for the logistic regression model is the Account Balance. Having Some Balance had the most statistically relevant p-value (1.79e-06), signified by the three asterisks next to the first row in the coefficient chart. Purpose (New Car), and Credit (Amount) were also significant with p-values of .00519 and .00989 respectively.

**Decision Tree**



The most important variables for the Decision Tree model are the Payment Status of Previous Credit and the Duration of Credit Month.

**Forest Model**

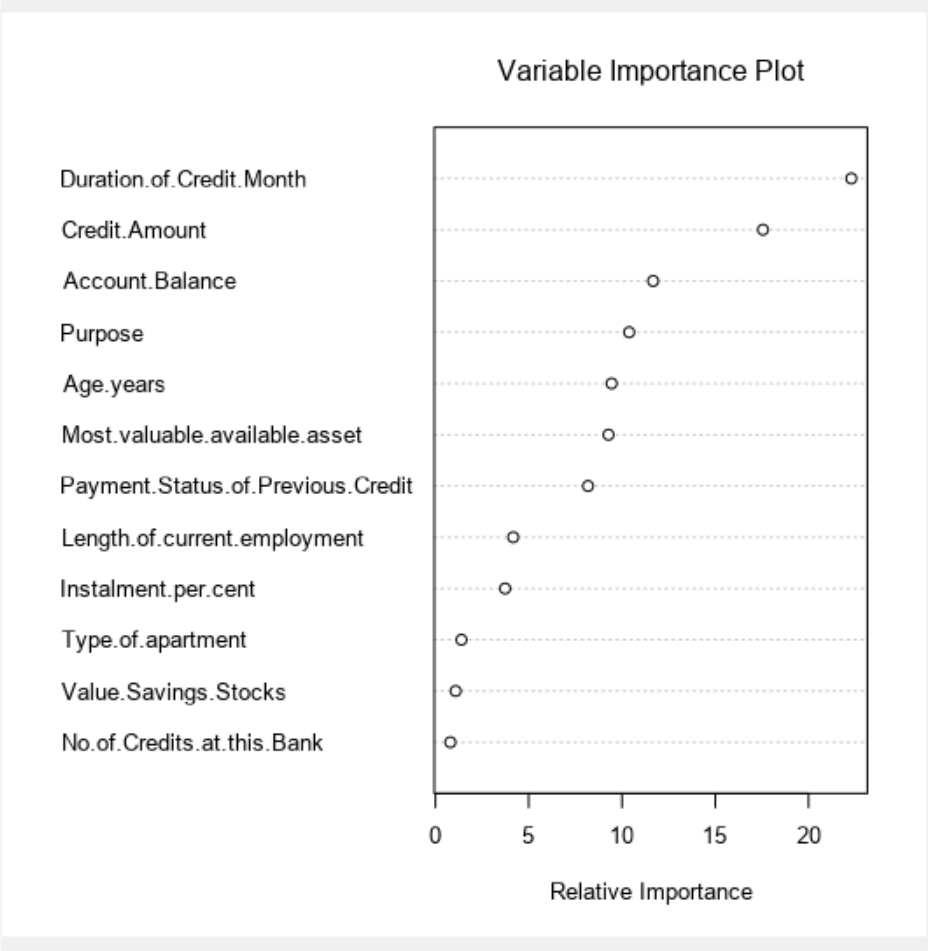


For the Forest Model, the two most important variables are Credit Amount and Age.years. Duration.of.Credit.Month is also relatively close to Age.years in terms of importance.

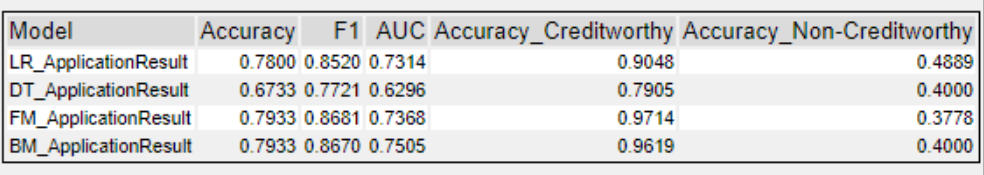
**Boosted Model**

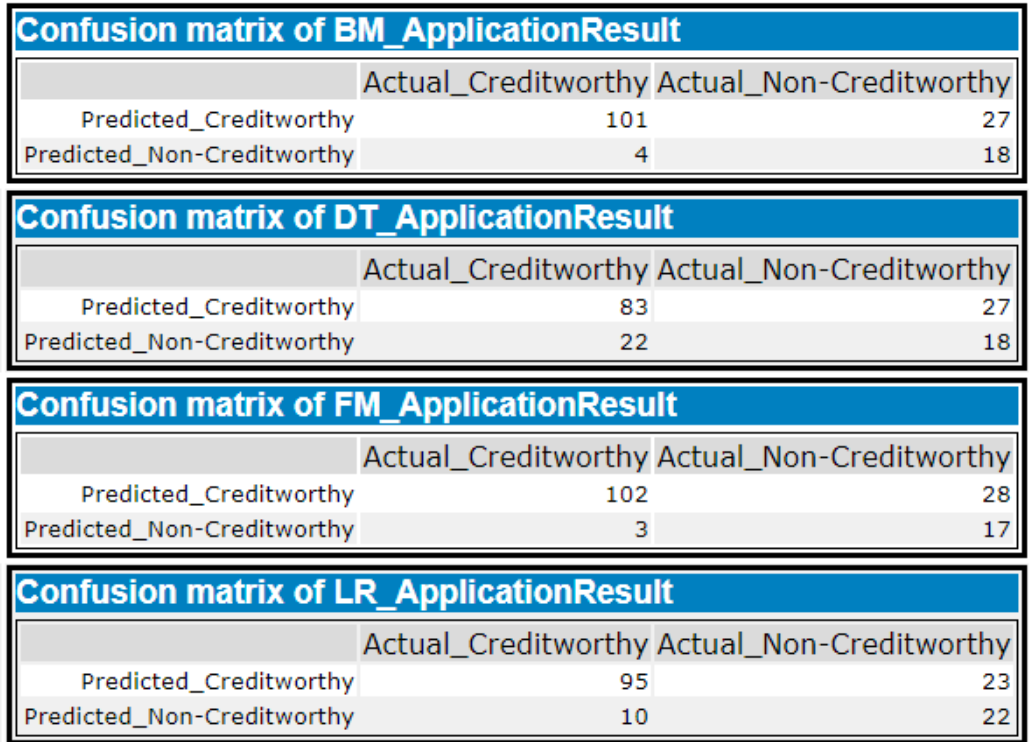
The most important variables for the Boosted Model are Duration.of.Credit.Month

and Credit.Amount.



* 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?





*You should have four sets of questions answered. (500 word limit)*

# Step 4: Writeup

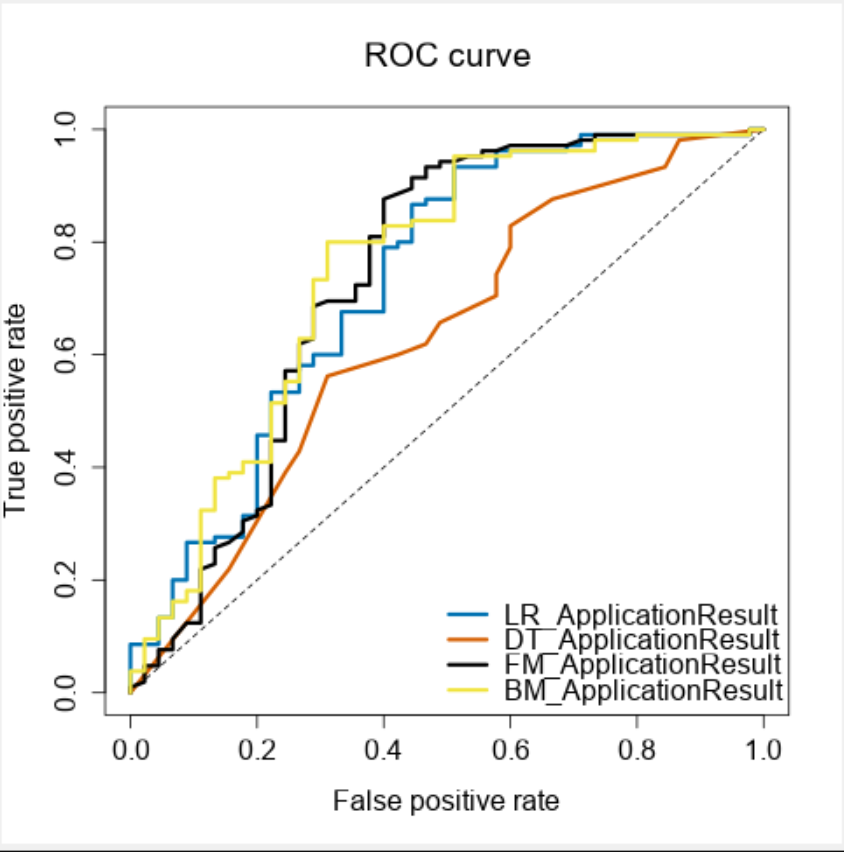
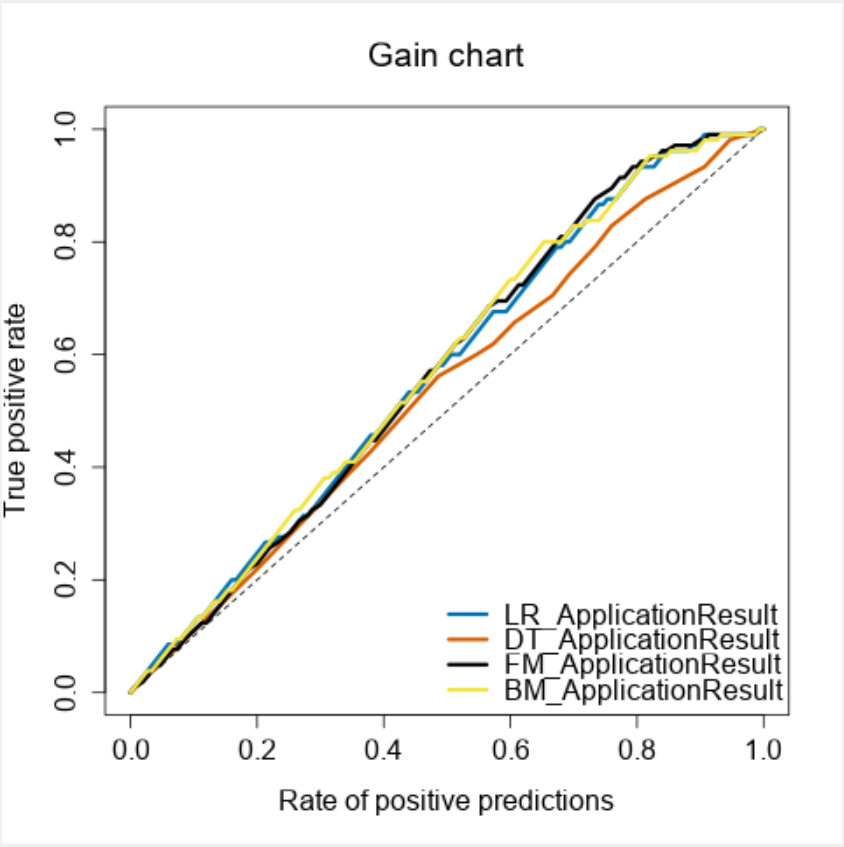
*Decide on the best model and score your new customers. For reviewing consistency, if Score\_Creditworthy is greater than Score\_NonCreditworthy, the person should be labeled as “Creditworthy”*

*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

Looking at the Model Comparison Report, we can see the Accuracy of Creditworthy applicants ranging from 79% to 97%. Although the Forest Model and the Boosted Models both share the top overall accuracy numbers at 79.33%, we chose to use the Forest Model because the Accuracy of Creditworthy applicants is slightly higher (97.14% vs 96.19%).



The black line denoting the Forest Model reaches the highest true positive rate first among all of the models in the gain chart, whereas it is second against the logistic regression for the ROC curve.

**Note:** Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

* How many individuals are creditworthy?

There are 408 individuals that are creditworthy and 92 that are not creditworthy.