Project Creditworthiness

Dung Nguyen

## Step 1: Business and Data Understanding

### What decisions needs to be made?

Determining whether customers are creditworthy to give a loan to.

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

* Data set of past applications for training models:
  + Outcome: Credit-Application-Result.
  + Potention predictors: Duration\_of\_Credit\_Month, Credit\_Amount, Age\_years, Instalment\_per\_cent, Most\_valuable\_available\_asset, Type\_of\_apartment, and so on.
* New data set (without outcome variable) to apply the chosen model and make decisions.

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

Binary model.

## Step 2: Building the Training Set

After importing, the data set were formatted as suggested.

### Identifying and imputing missing data

The number of missing values in each field:

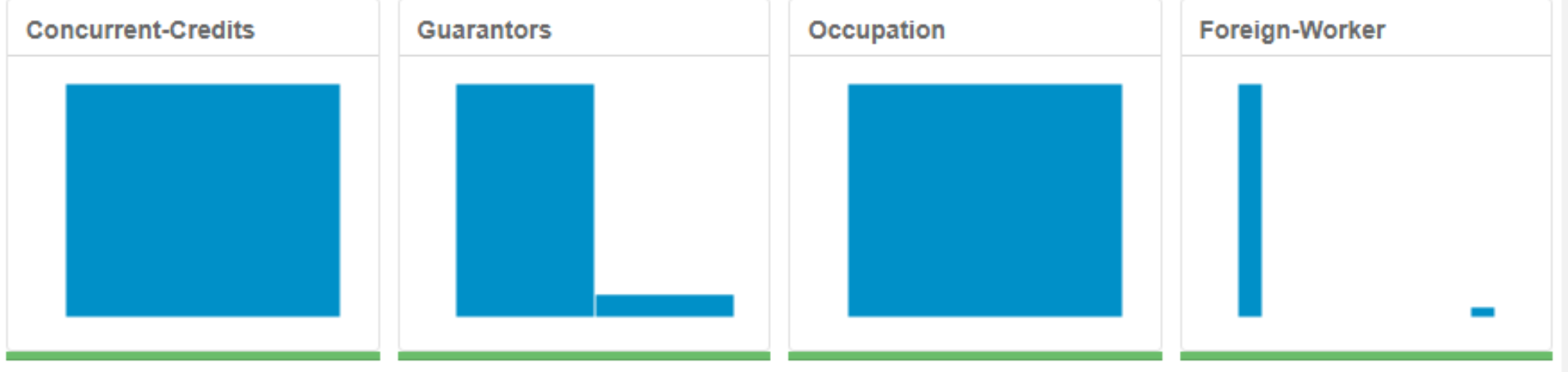
## field no\_of\_missing  
## 1 Credit\_Application\_Result 0  
## 2 Account\_Balance 0  
## 3 Duration\_of\_Credit\_Month 0  
## 4 Payment\_Status\_of\_Previous\_Credit 0  
## 5 Purpose 0  
## 6 Credit\_Amount 0  
## 7 Value\_Savings\_Stocks 0  
## 8 Length\_of\_current\_employment 0  
## 9 Instalment\_per\_cent 0  
## 10 Guarantors 0  
## 11 Duration\_in\_Current\_address 344  
## 12 Most\_valuable\_available\_asset 0  
## 13 Age\_years 12  
## 14 Concurrent\_Credits 0  
## 15 Type\_of\_apartment 0  
## 16 No\_of\_Credits\_at\_this\_Bank 0  
## 17 Occupation 0  
## 18 No\_of\_dependents 0  
## 19 Telephone 0  
## 20 Foreign\_Worker 0

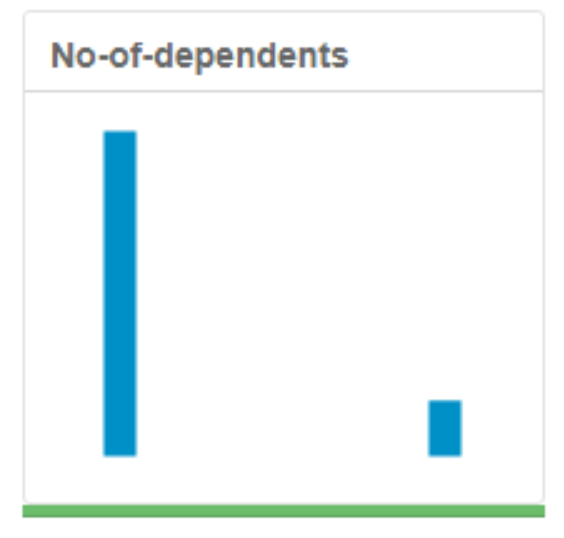
Duration\_in\_Current\_address has 344 missing values so I dropped it. Age\_years has 12 missing values so I imputed it by median.

## [1] 35.574

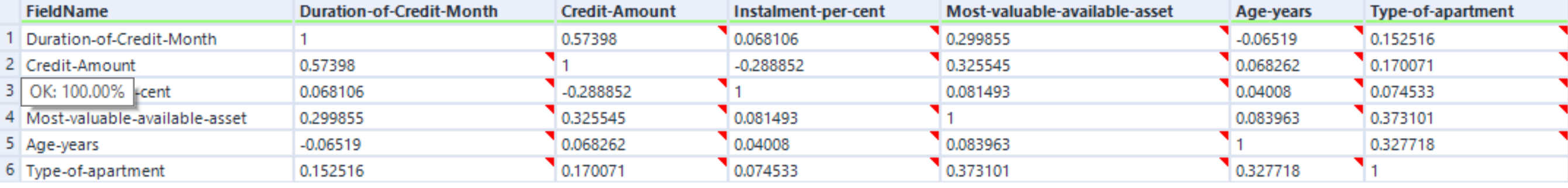
### Identifying low-variability fields and removing them

Histogram of low-variability fields:

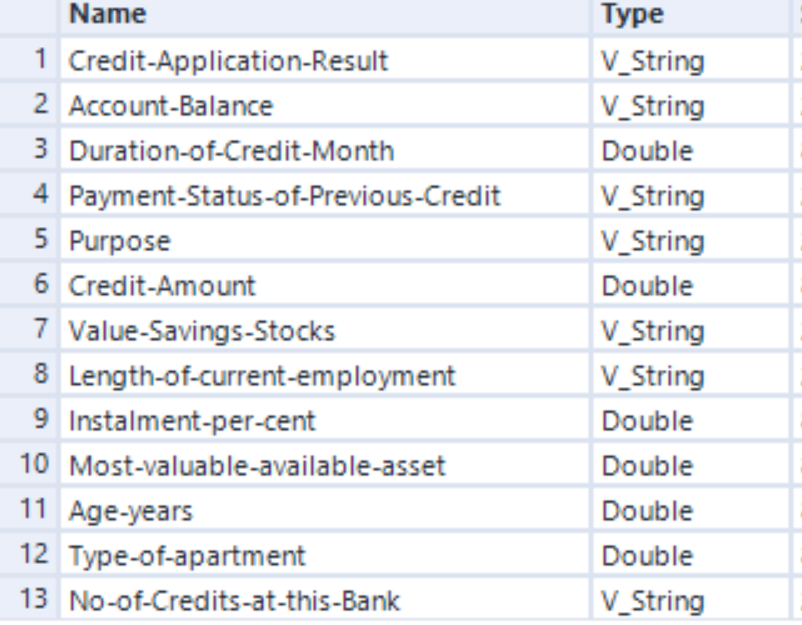




After removing those fields plus Telephone, no high correlation was detected among numeric fields:



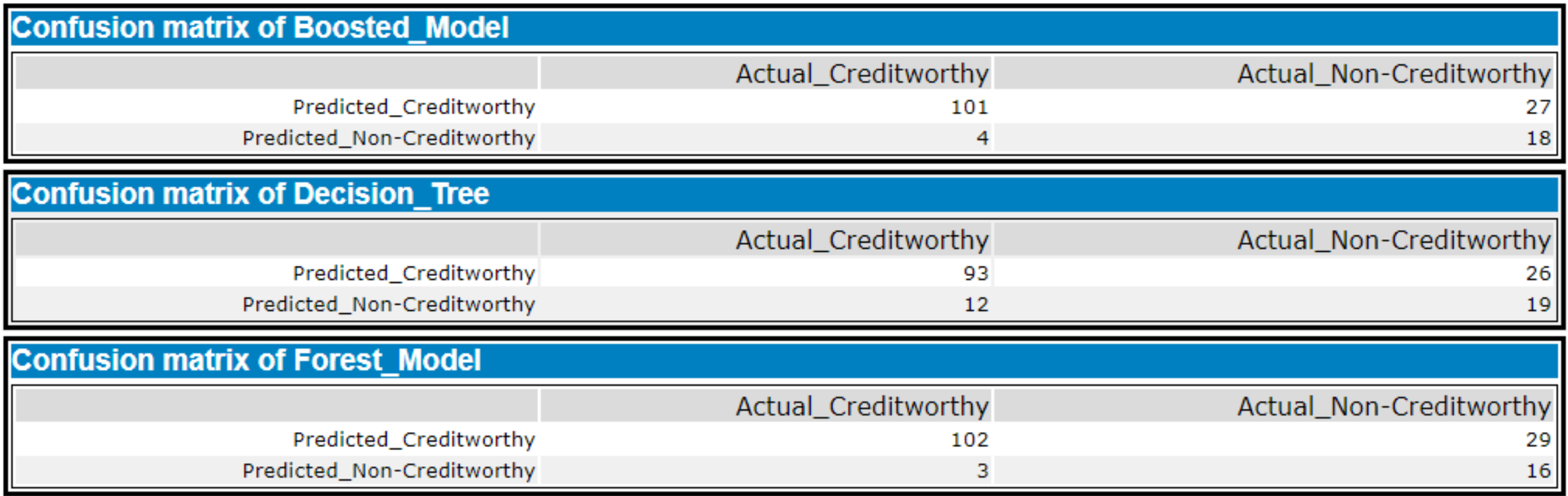
The final data set has 13 columns.

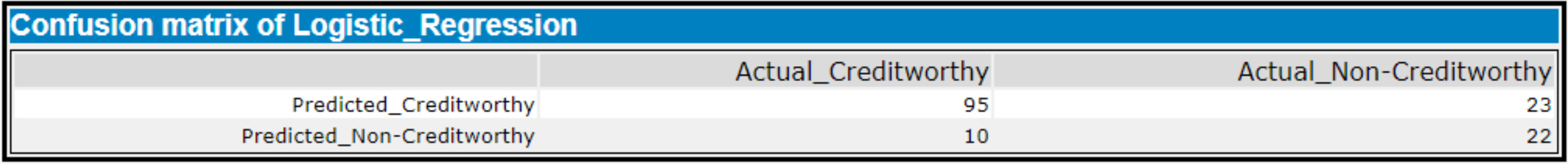


## Step 3: Train your Classification Models

### Create Estimation and Validation samples

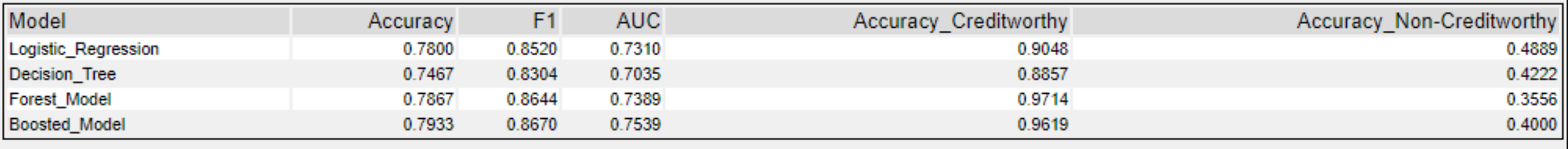
### Confusion matrices





### Overall accuracy

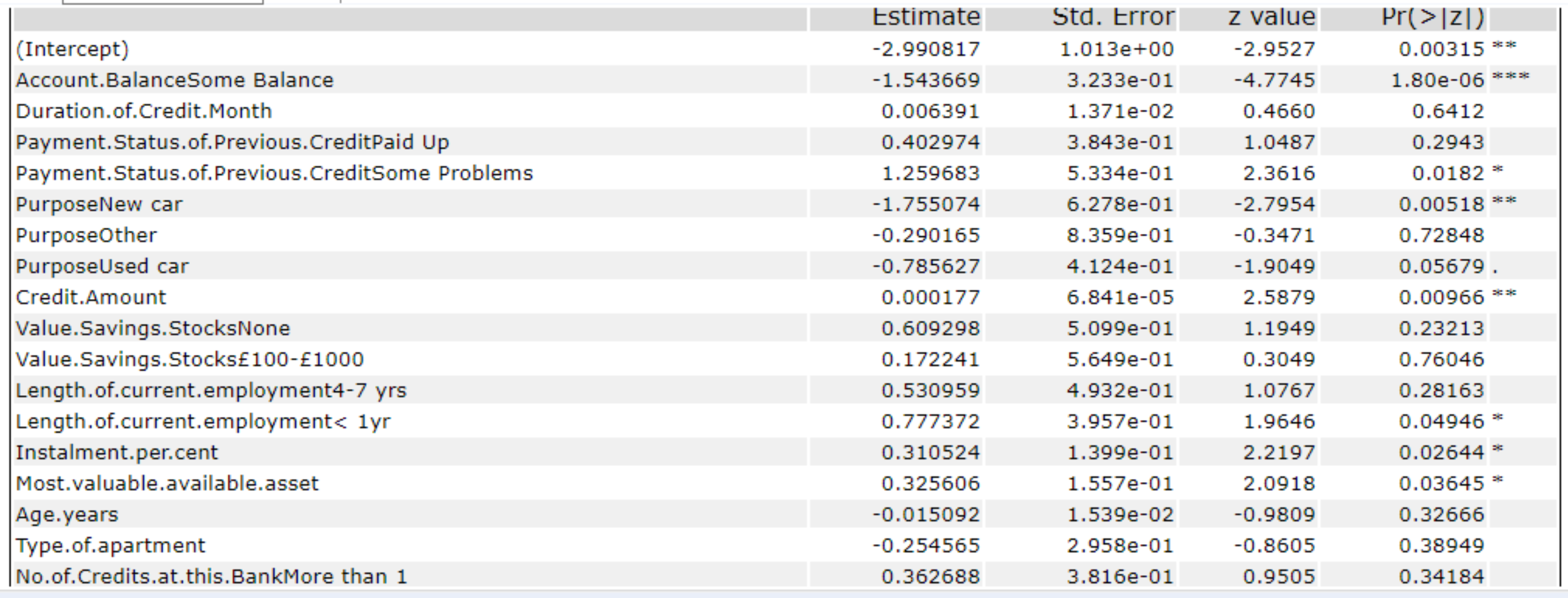
Among the 4 models, Boosted Model has the highest overall accuracy agains the Validation set (0.7933). Forest Model is the least biased with PPV = 0.78 and NPV = 0.84.



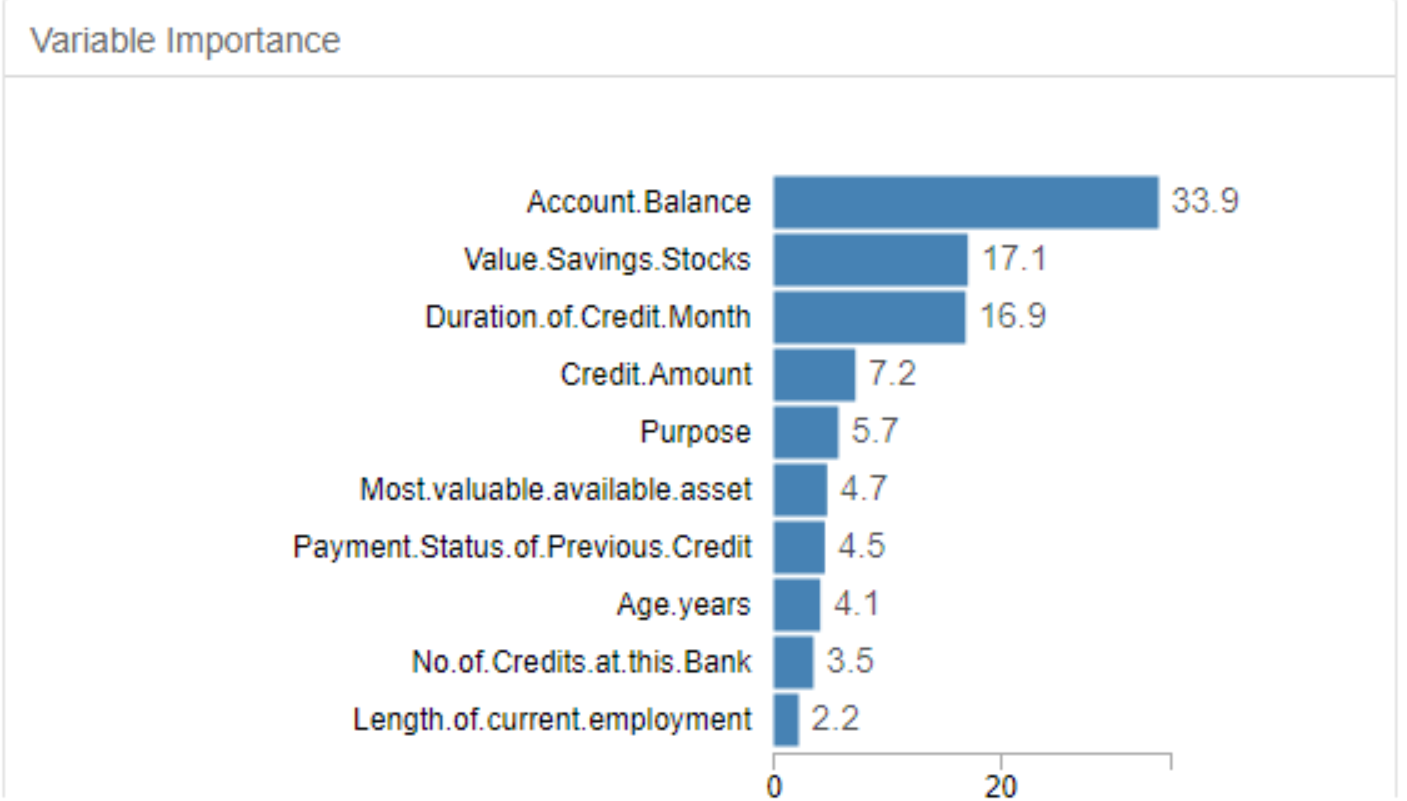
### Plot variable importance

Most important predictors: - For Logistic regression: Account\_Balance, Purpose, Payment\_Status\_of\_Previous\_Credit (Paid up), Length\_of\_current\_employment (<1yr), Instalment\_per\_cent, and Most\_valuable\_available\_asset. - For Decision tree: Acocunt\_Balance, Value\_Savings\_Stocks, Duration\_of\_Credit\_Month, Credit\_Amount, and Purpose. - For Forest Model: Credit\_Amount, Age\_years, Duration\_of\_Credit\_Month, Account\_Balance, and Most\_valuable\_available\_asset. - For Boosted tree: Account\_Balance, Credit\_Amount, Duration\_of\_Credit\_Month, Paymen\_Status\_of\_Previous\_Credit, and Purpose. The difference is not remarkable among the fields.

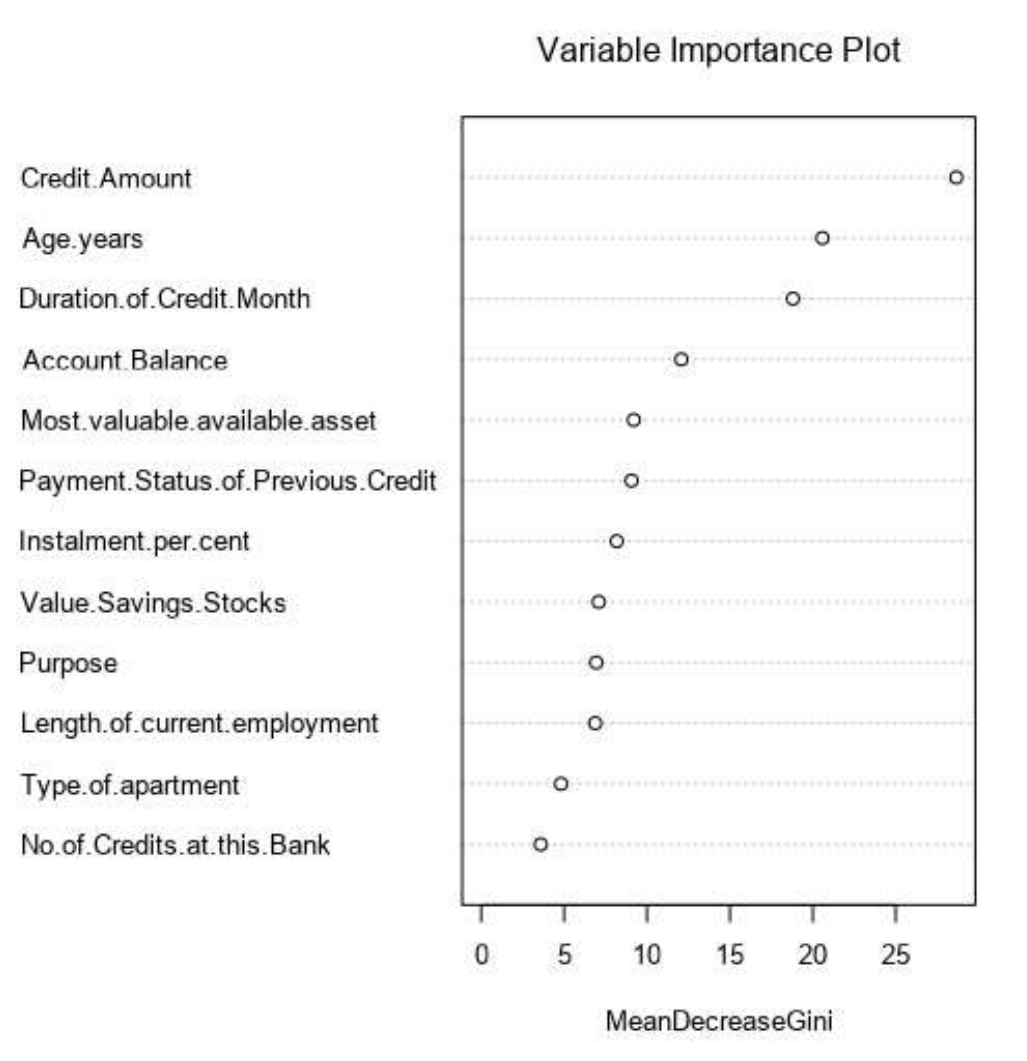
- Logistic Regression:



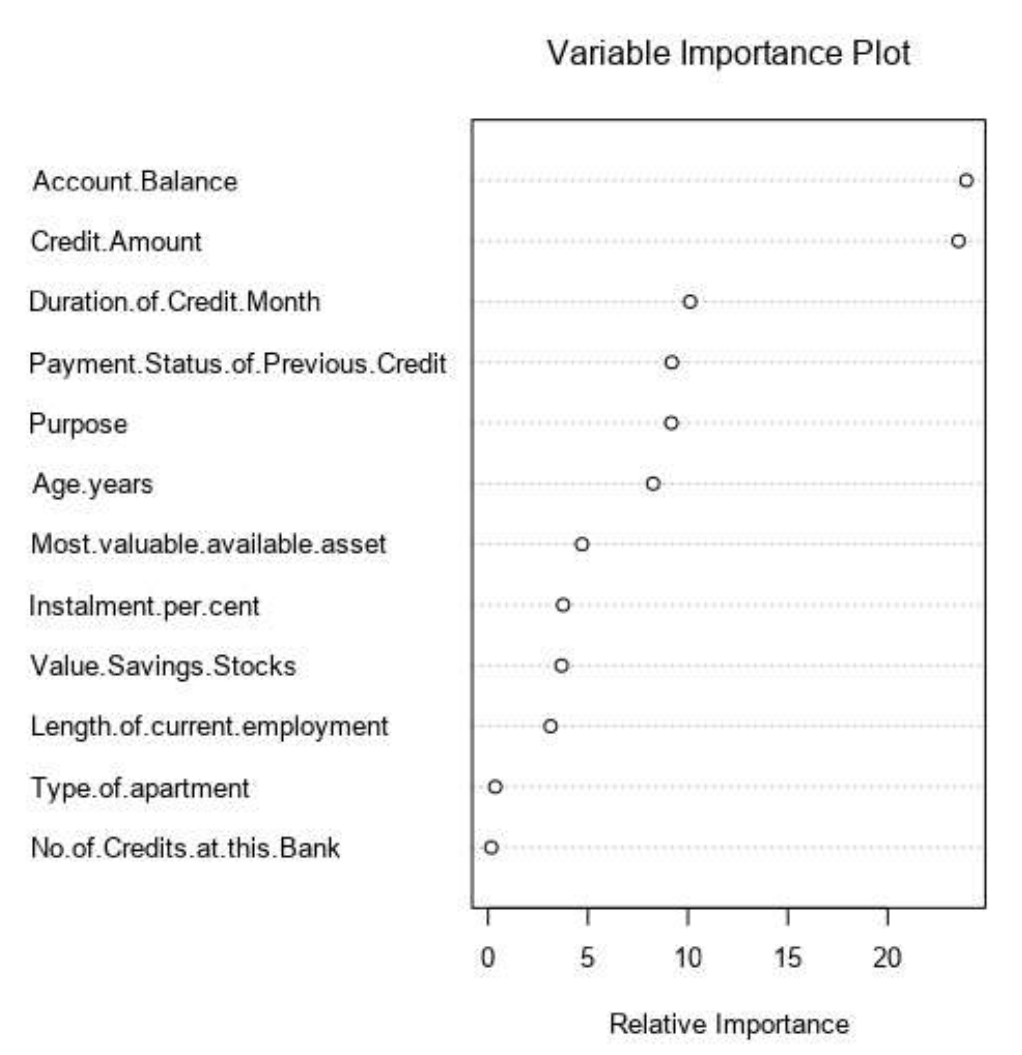
- Decision Tree:



- Forest Model:

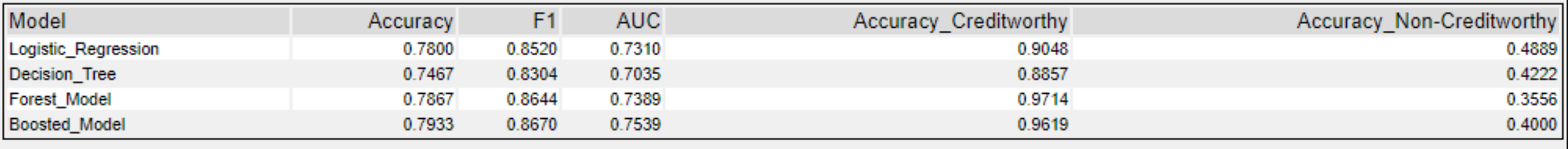


- Boosted Model:

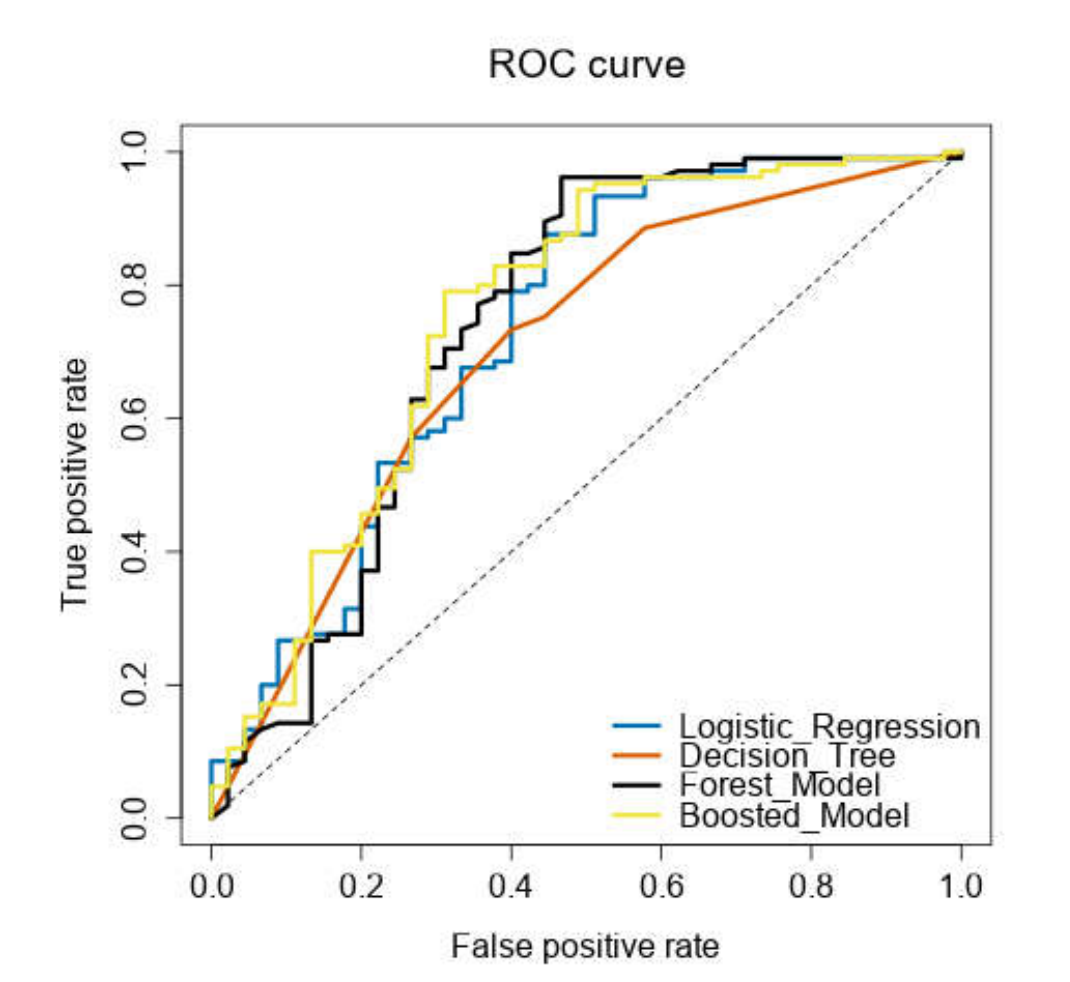


## Step 4: Writeup

Accuracy table



Plot ROCs



Forest Model is the least biased and has the nearly highest overall accuracy against the Validation set. Although its accuracy within Non-creditworthy segment is not high, its respective index in Creditworthy segment is the highest. Forest Model also has the most optimal ROC which reaching the top quicker than Boosted Model. Therefore, I choose to use Forest Model. Applying the chosen model, the number of individuals are creditworthy is 412.