Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- **What decisions needs to be made?**
  As a loan officer at a young and small bank, I need to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not.

  Due to a financial scandal that hit a competitive bank last week, a sudden influx of 500 new people migrated to my bank, I need to use a series of classification models to figure out the best predictive model and provide a list of creditworthy customers in the next two days, in order to don't miss this huge opportunity for the bank that I work for.

- **What data is needed to inform those decisions?**
  Data on all past applications:
  Credit Application Result, Account Balance, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Credit Amount, Value Savings Stocks, Length of current employment, Instalment per cent, Most valuable available asset, Age years, Type of apartment, No of Credits at this Bank

  The list of customers that need to be processed in the next few days:
  Account Balance, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Credit Amount, Value Savings Stocks, Length of current employment, Instalment per cent, Most valuable available asset, Age years, Type of apartment, No of Credits at this Bank

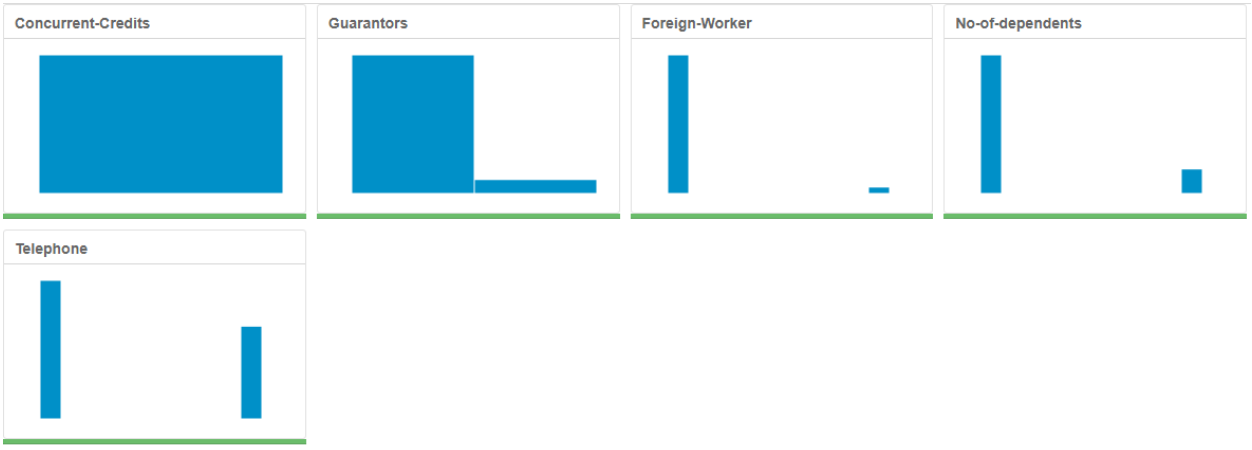
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

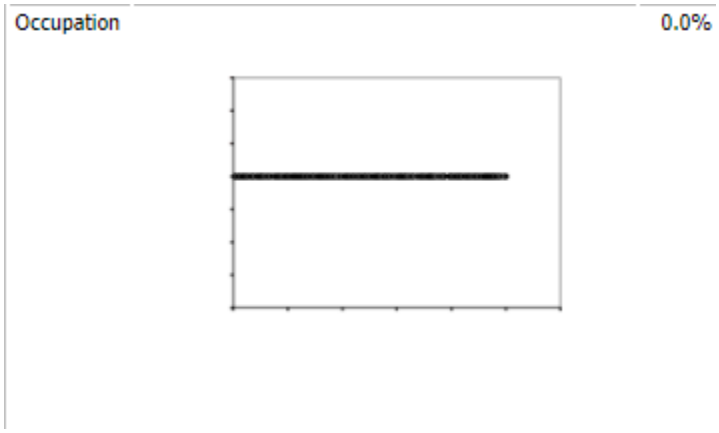
  As the answer to this problem is Binary (creditworthy / non- creditworthy), we need to build a model that best fit with it. In order to achieve it, I'll compare the following binary classification models and choose the one that performs best:
  - Logistic Regression.
  - Decision Tree.
  - Random Forest.
  - Boosted Model

Step 2: Building the Training Set

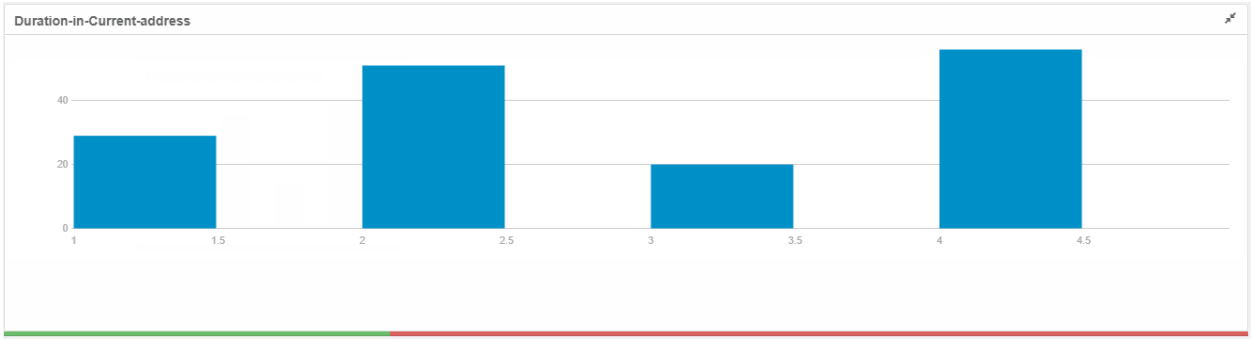
2.1. Data Exploration

- **In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.**
  To the cleanup process, I started with Data Exploration, visualizing the data distribution and identifying which fields could be removed because of its “Low Variability”. All the following graphs were classified as “Low Variability” and was removed:





Because the “Duration-in-Current-address” field has a lot of missing data (69%) I also removed this field.

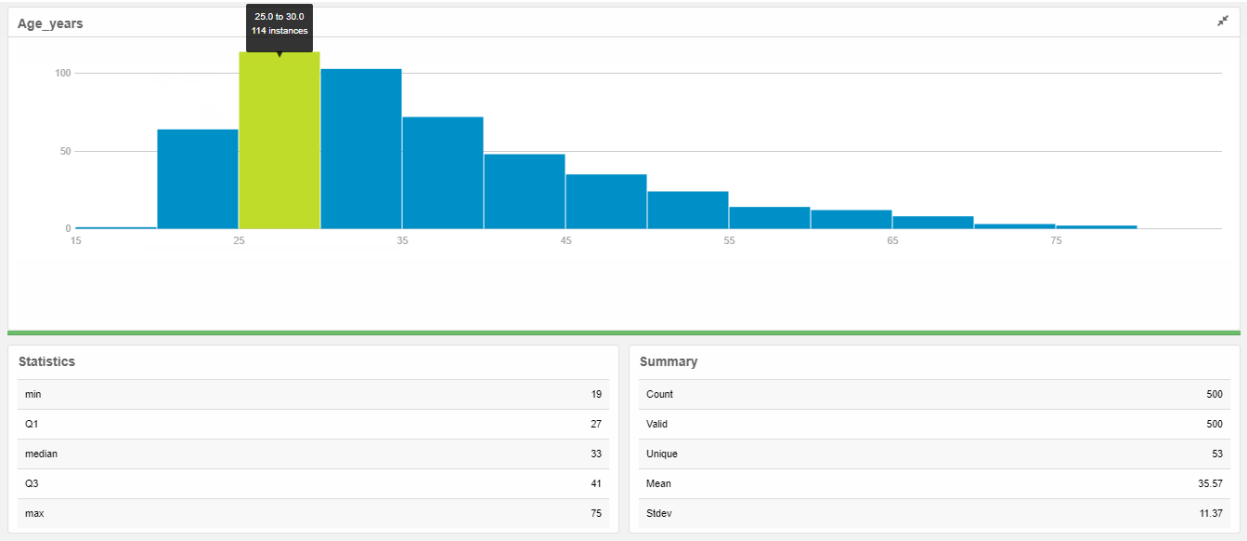


Now looking for the remaining fields can be identified some missing data in the “Age-years” field. As the missing data is only a few parts of the total (2%) and this field seems to be a high candidate to be an important variable to our prediction, I imputed the missing values with the median of the Age-years.

Although there are better techniques of value-imputation of the missing values, I used an imputation of median-values, which inputs less bias in the dataset than the average-values instead.

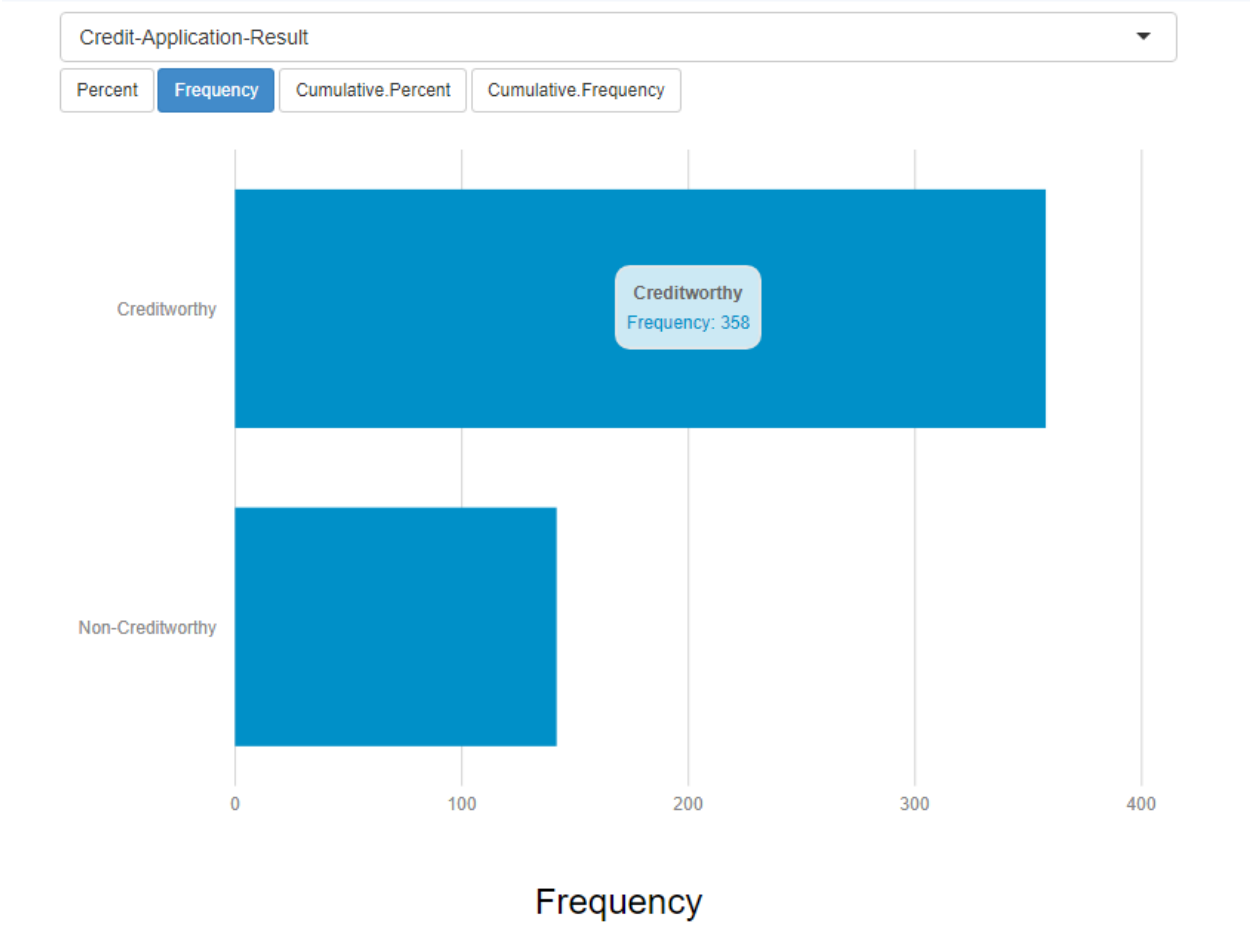
Using the "average values" could insert a skewed graph in our model based on the age of the current clients, in this case we would be inputting an error/bias in our model and it could return a biased analysis, for example: if we train/teach our model that people with 50+ years are the most reliable/creditworthy, most of the new customers that have less than 50 years probably wouldn’t be accepted in our bank, but this concept of reliability was based in our current customers and this couldn’t be (and probably isn’t) an indisputable truth, especially when compared with a large amount of data.

When we’re using the median of age-years, we’re inputting the median value between the oldest and youngest, representing the age group that we attend in our bank. Even though this method input some errors in our dataset, at least our dataset isn’t biased, and these errors tend to be lower than the average method when applying our model with a large age group.



2.2. Frequency of target-variable

In order to analyze the frequency of the creditworthy and non-creditworthy in our dataset, we could see that our bank grants a lot of credit, probably because it's a young and small bank.



Step 3: Train your Classification Models

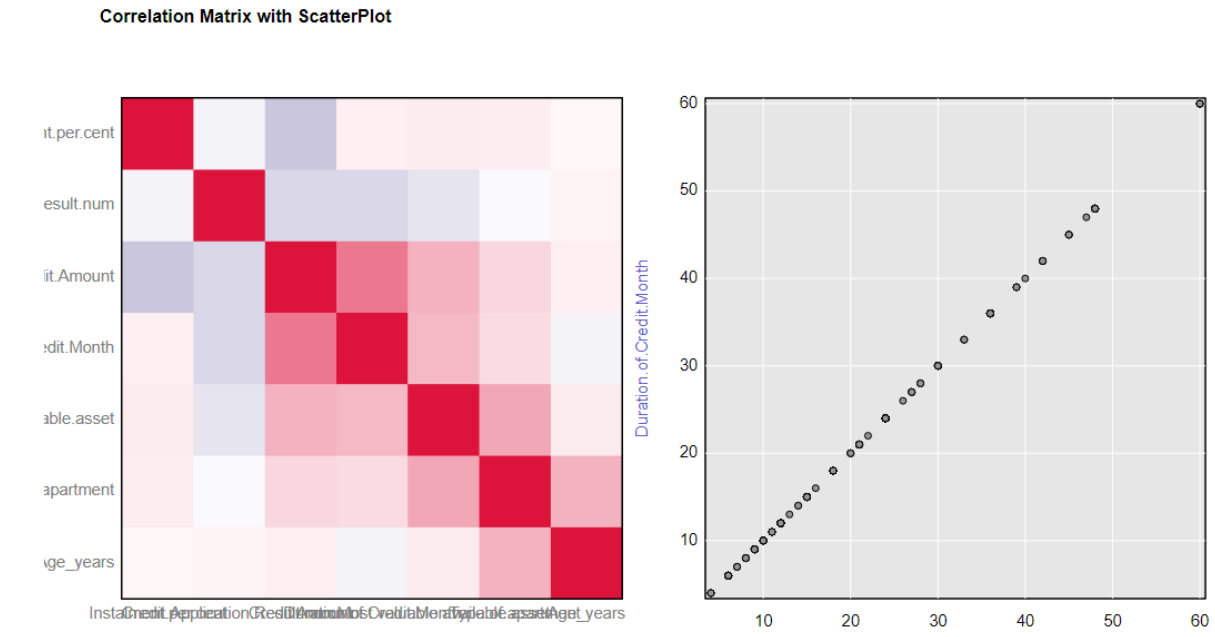
- **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**

1.3. Predictors variables – Correlation

After applying logic to check if our list of potential variables has duplicates variables, now we will check the correlation between remaining variables.

I checked for correlations between my predictor variables to see if there is any possibility of multicollinearity in my dataset. Below is a table that shows the correlations between the different predictor variables:

Field Name	Association Measure	p-value
Duration.of.Credit.Month	-0.202504	5.0151e-06***
Credit.Amount	-0.201946	5.3311e-06***
Most.valuable.available.asset	-0.141332	1.5334e-03**
Instalment.per.cent	-0.062107	1.6556e-01
Age_years	0.052914	2.3758e-01
Type.of.apartment	-0.026516	5.5417e-01



- 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?

  In order to obtain a reliable model which could predict/classify our customers, we need to train our model. First, I created an Estimation and Validation samples where 70% of the dataset were to Estimation and 30% of the entire dataset were reserved for Validation.

  1. Logistic Regression

     The first model we'll test is Logistic Regression. To let the Stepwise tool, decide which predictor variables are significant, we chose all the variables in the Logistic Regression tool, other than the target variable.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183	*
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566	**
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042	
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596	*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

According to the Logistic Regression model, the Predictor Variables that are significant are:
Account-Balance
Payment-Status-of-Previous-Credit
Purpose
Credit-Amount
Length-of-current-employment
Instalment-per-cent

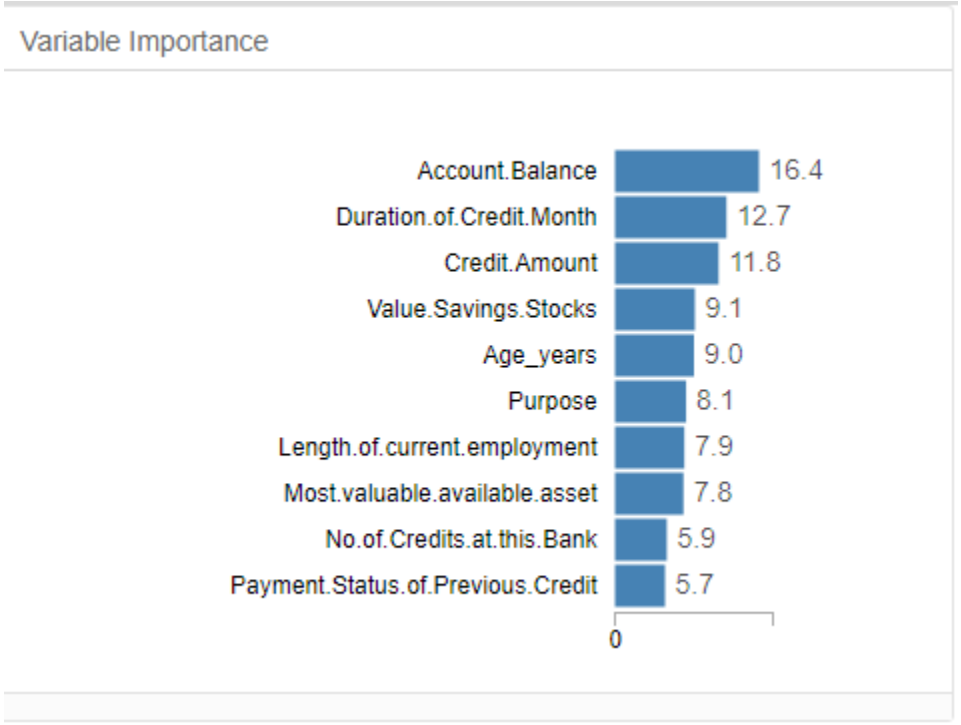
2. Decision Tree

Testing the Decision Tree model into our dataset we could see that even though the Root node error is quite high it still under 28%, which considered as an acceptable error.

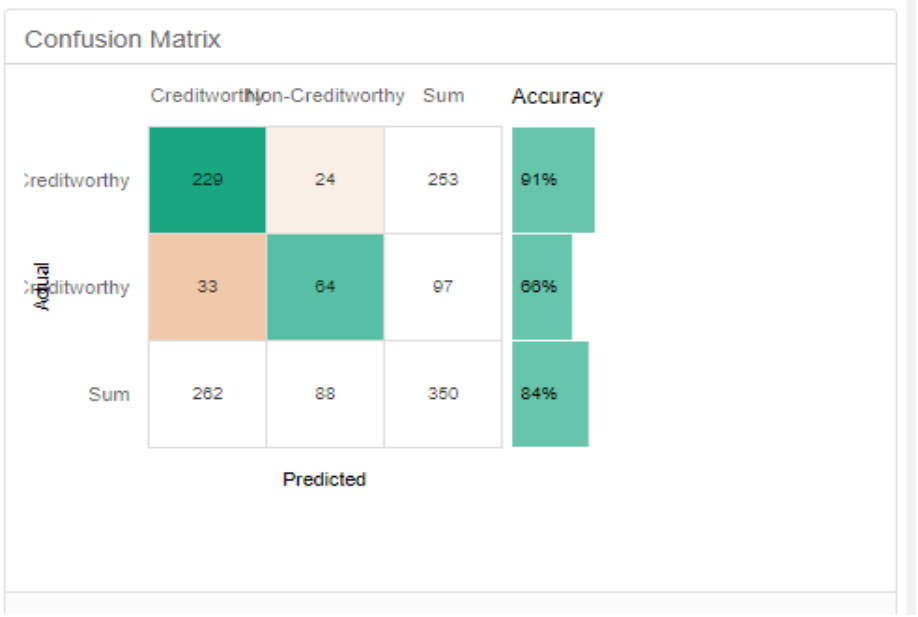
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age_years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

The Variable Importance Plot indicates that the most important predictor variables into this model are:

Account-Balance
Duration-of-Credit-Month
Credit-Amount



When we are validating our model against itself with the Confusion Matrix, we can see that the Sum of Accuracy is 84%, classifying it as a reliable model.

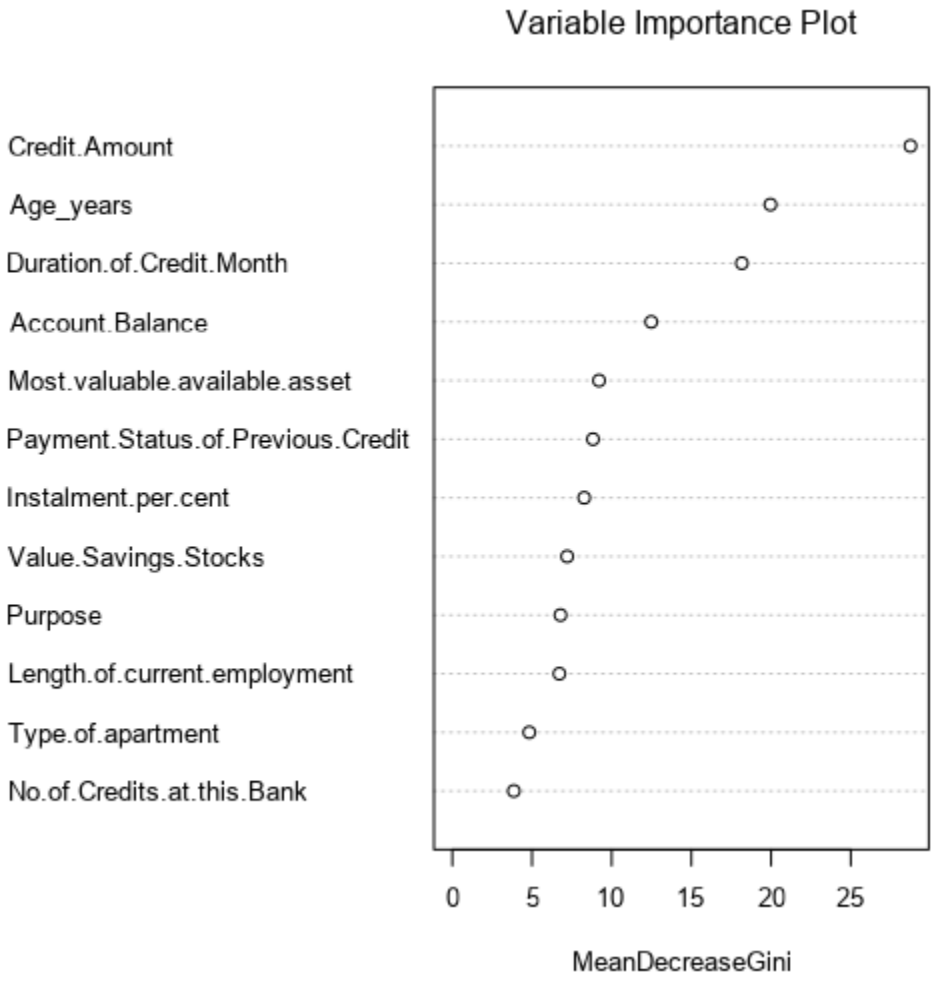


3. Forest Model

   Looking at the Confusion Matrix of the Forest Model, we can see that the accuracy of this model (trained with only Estimation Data) is the best until now, with an overall value of 78%

Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 23.1%
Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.067	236	17
Non-Creditworthy	0.66	64	33

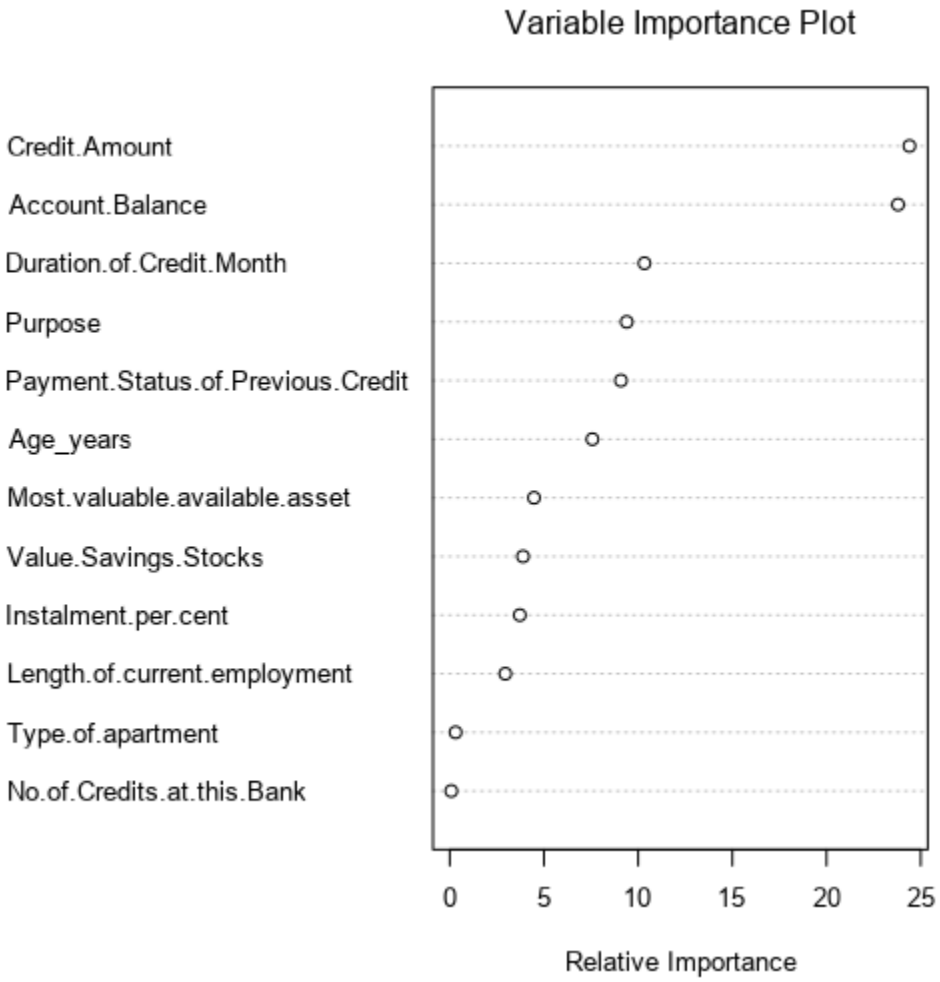


4. Boosted Model

Basic Summary:

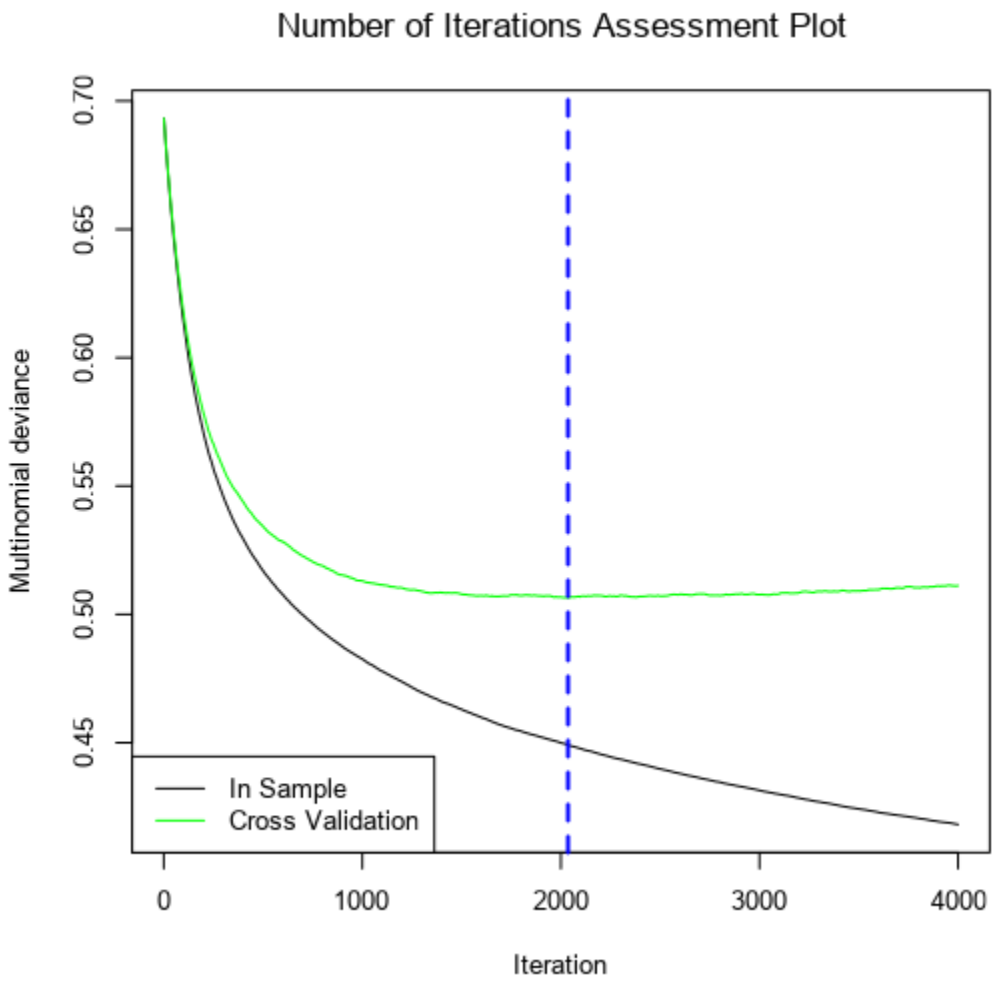
Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum

to 100, and the value for each field gives the relative percentage importance of that field to the overall model.



The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).

Step 4: Writeup

- 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

Choosing the best Model

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthy.	0.6733	0.7721	0.6296	0.7905	0.4000
Ft_Creditworthy.	0.7933	0.8681	0.7377	0.9714	0.3778
BM_Creditworthy.	0.7867	0.8632	0.7524	0.9619	0.3778
stepwise_Creditworthy.	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of BM_Creditworthy.

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DT_Creditworthy.

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Ft_Creditworthy.

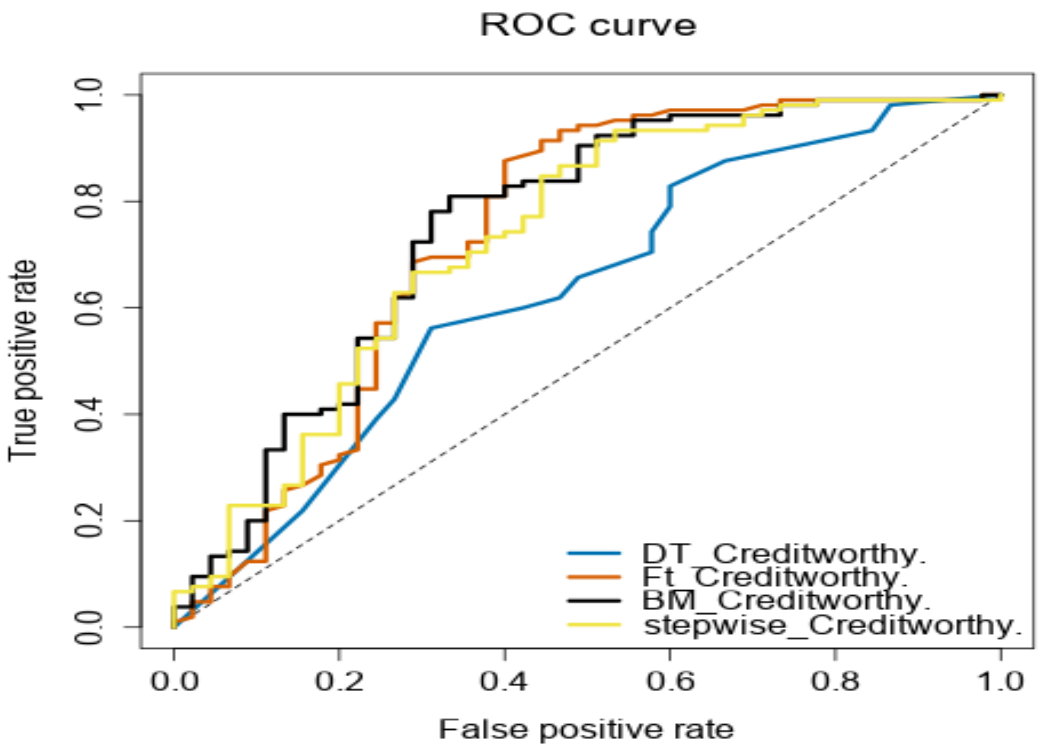
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of stepwise_Creditworthy.

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Taking over the overall accuracy to predict the best fit model, we came with the two bests models: Forest Model with the highest overall accuracy of 79.33%, as well we can see that Forest model has the highest Accuracy Creditworthy at 97.14%. We’re going

further onto this analysis focusing only on these models because the most important on this analysis is how accurately we can identify people who qualify for loan.
When we see at ROC graph, we can say that Forest model hugs the topmost true positive side of the graph.



- How many individuals are creditworthy?

  With an overall accuracy of **79%** and **97%** of accuracy in predicting Creditworthy individuals, we can predict that **408** of the new customers can be classified as **Creditworthy**.