Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project>

# 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?

Decision on whether new customer can be approved for a loan or not needs to be made.

* What data is needed to inform those decisions?

We need the data for the previous customers who have their loans approved along with information like Account balance, Duration of Credit month, Payment status of previous credit, Purpose of loan, Credit amount, Value of stock savings, Length of current employment, Installment per cent, Most valuable available asset, Age, Type of apartment, No of credit at the bank

We also need above data of new customers who applied for the loan.

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

We need to predict whether the customer is Creditworthy or not to give loan , it is more like True or False situation , so Binary Model should be used.

# 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.***

*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:*

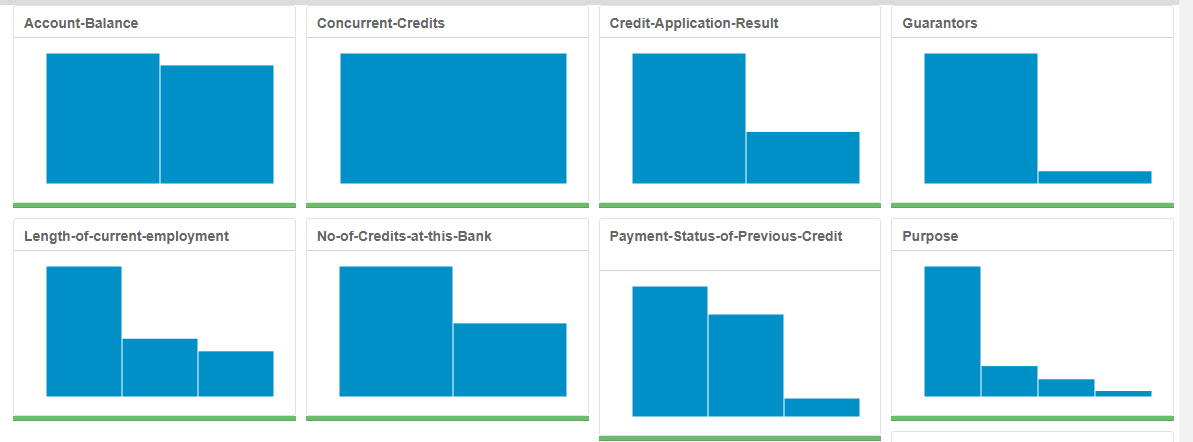
|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Credit-Application-Result | String |
| Account-Balance | String |
| Duration-of-Credit-Month | Double |
| Payment-Status-of-Previous-Credit | String |
| Purpose | String |
| Credit-Amount | Double |
| Value-Savings-Stocks | String |
| Length-of-current-employment | 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 |

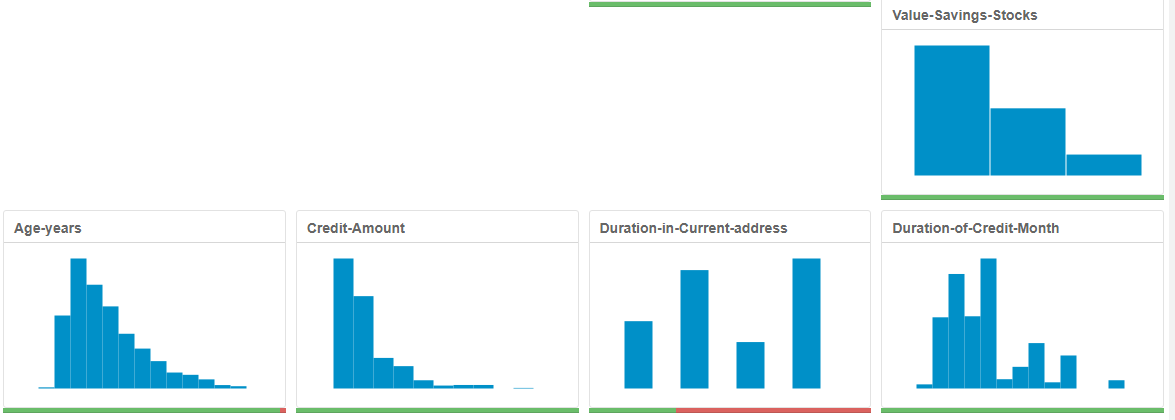
*To achieve consistent results reviewers expect.*

*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.

I have done a field summary analysis and below are some of the output







Based on the above analysis, I have removed the following fields

1. Concurrent Credits - it has a low variability and its removed (all values are unique)
2. Duration in Current address – Since percentage of data not available is high, this field has been removed
3. Guarantors - this category field doesn’t have equal split up and it has more instances in one criteria, so this field has been removed
4. Occupation – it has low variability and it’s been removed (all values are unique)
5. Number of Dependents – Doesn’t have uniform distribution and there is wide gap in the two instances and it’s been removed
6. Foreign Worker – Doesn’t have uniform distribution and there is wide gap in the two instances and it’s been removed
7. Telephone – Telephone fields will not be proper value used for predicting, since mostly this will be used during collection process

Imputed the Age field -It had uniform distribution and only 3% of the values are missing , Median of Age has been taken and added to the null values

# Step 3: Train your Classification Models

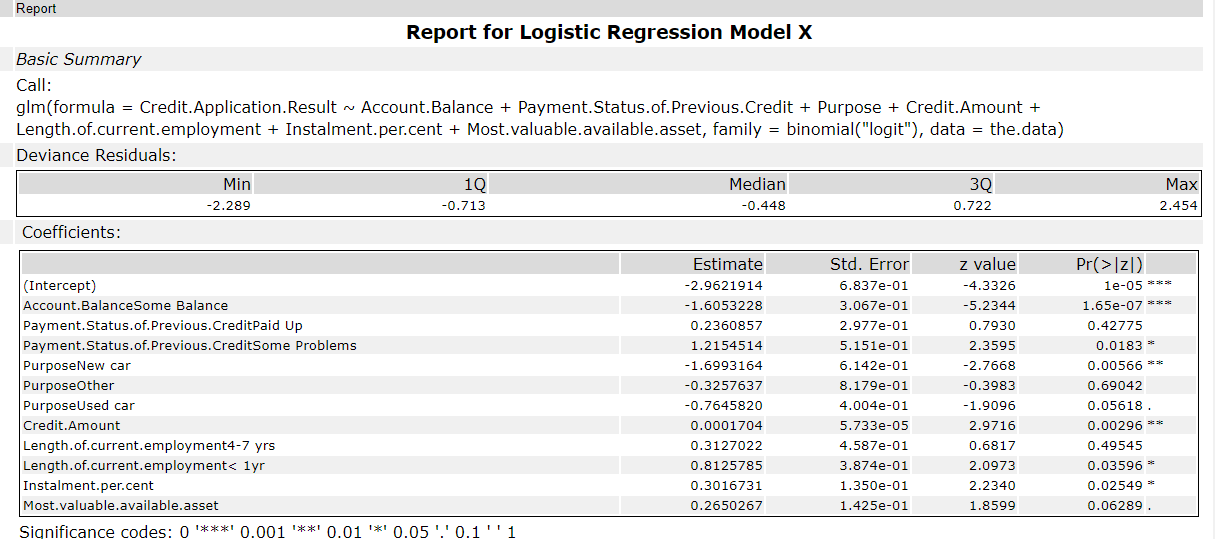
*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.*

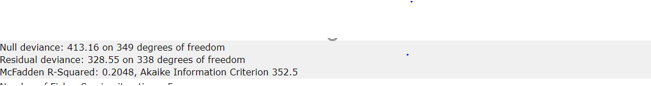
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

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

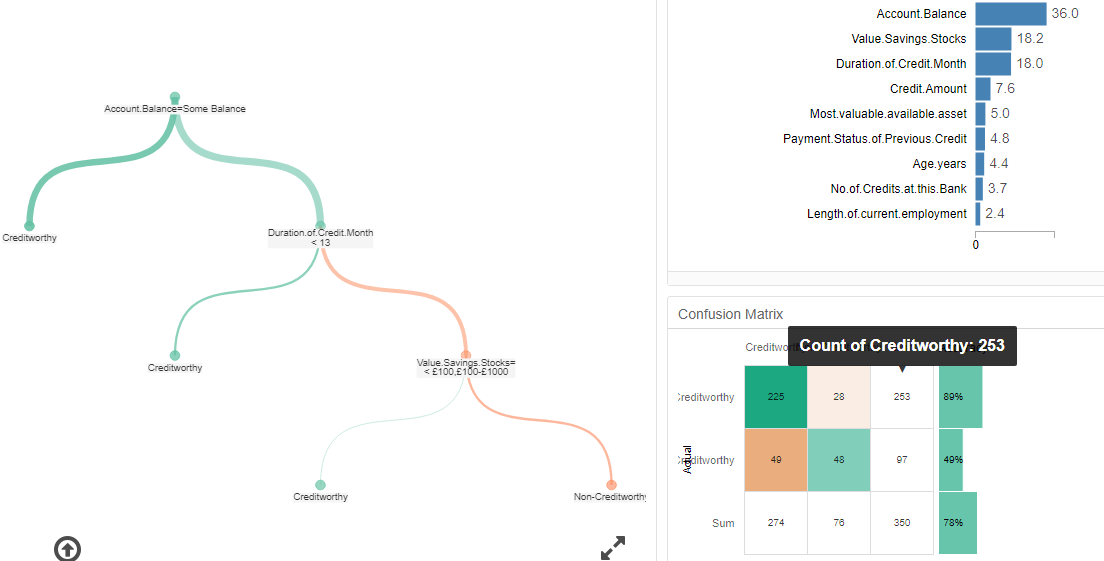
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

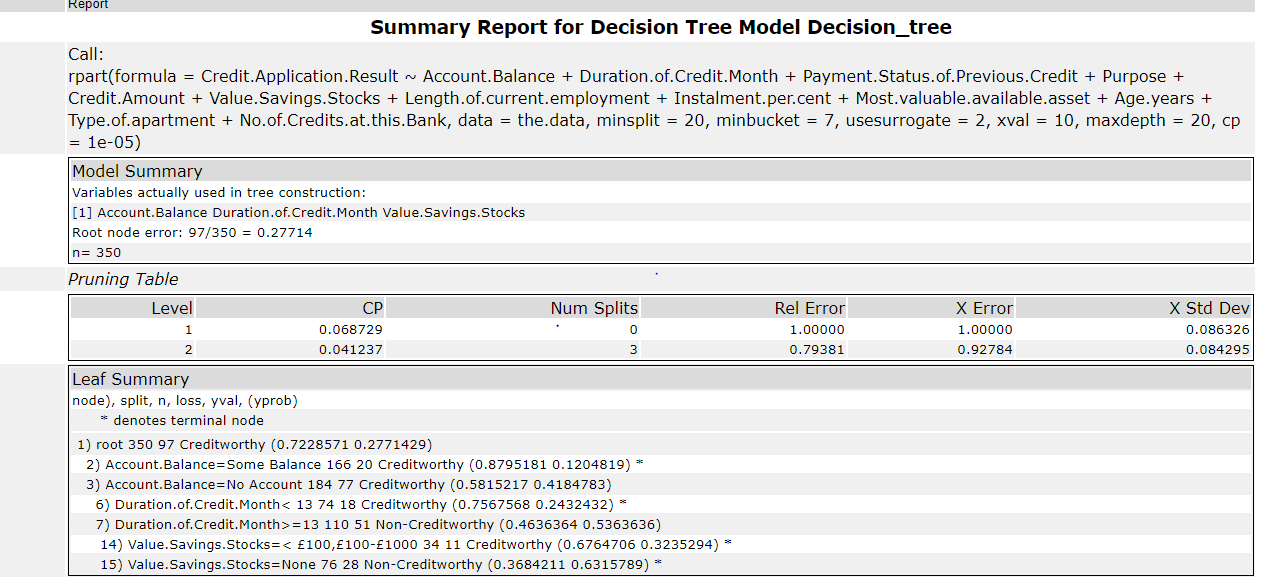
1. **Logistic Regression Model** – Account balance, Payment status of Previous Credit , Purpose, Credit amount , Length of current employment , Installment Percent , Most valuable asset

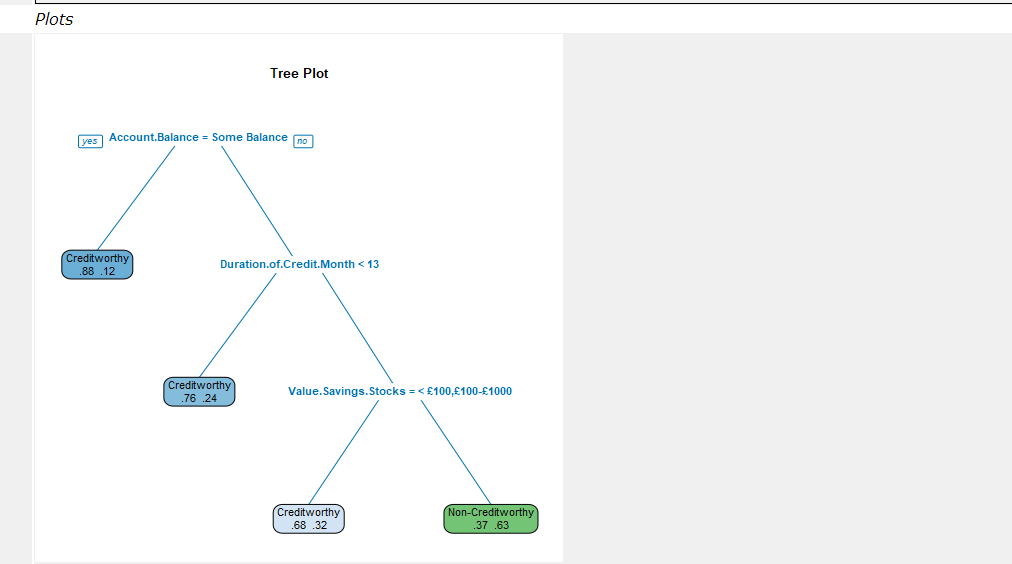




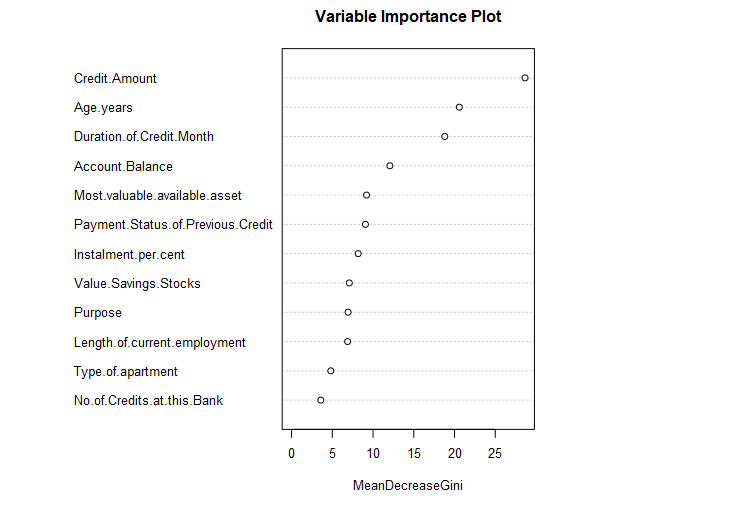
1. **Decision Tree –**

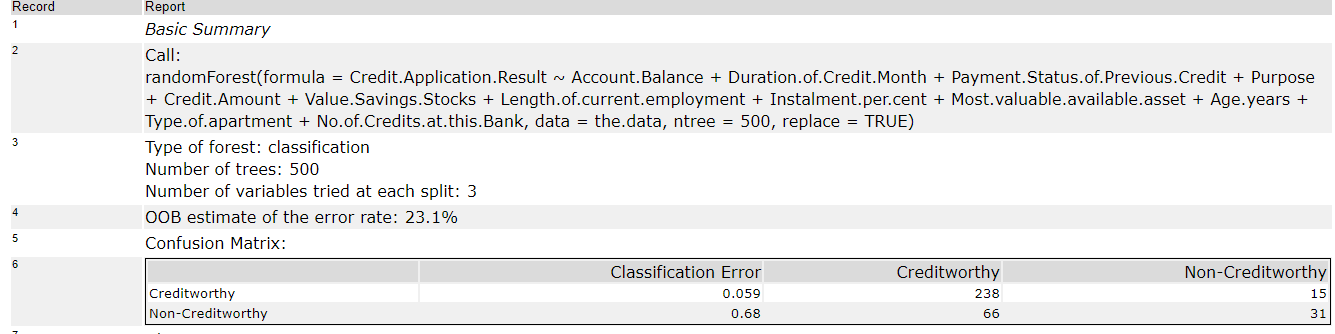
49



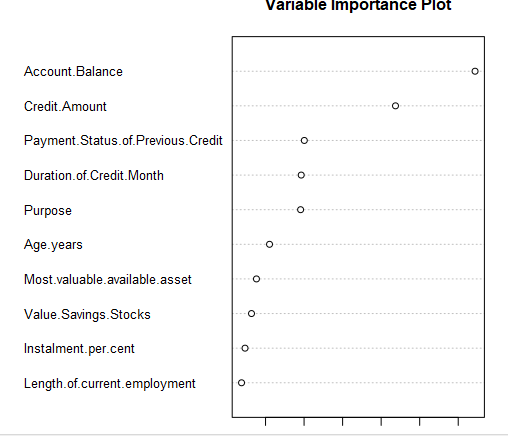


1. **Forrest Model**



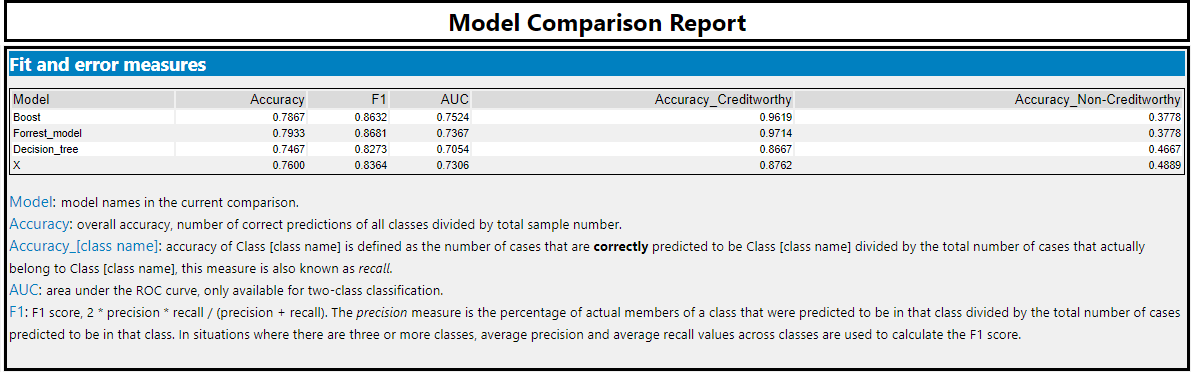


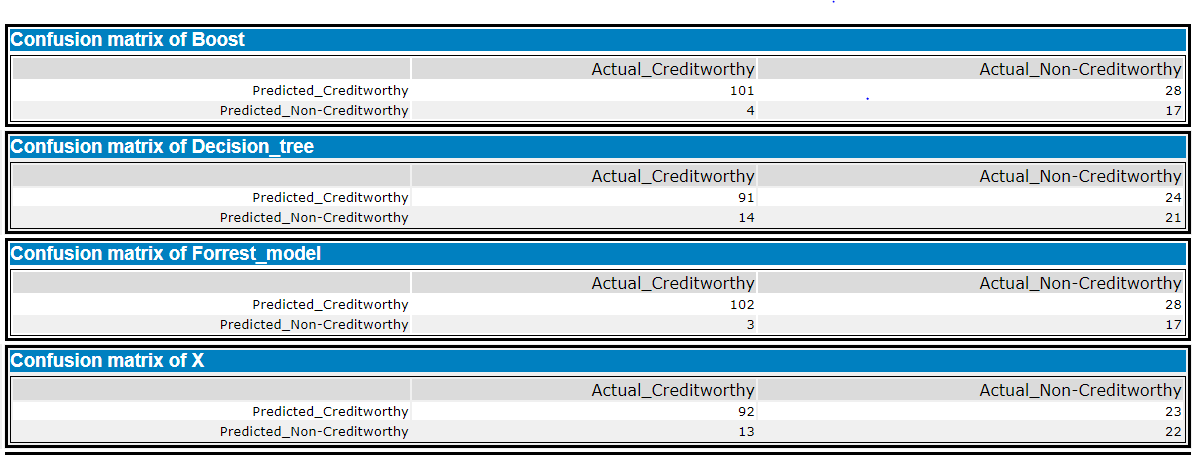
1. **Boosted Model**



* 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?

**Comparison of All models**





Based on the above matrix there is a bias across each model ,

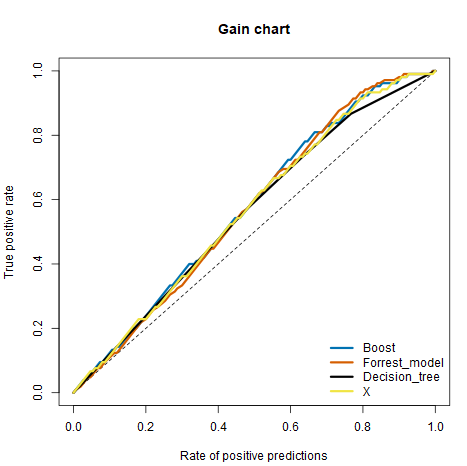
Boost model – has a accuracy of predicting Creditworthy at 96% and Non Creditworthy 38%

Forrest Model – has a accuracy of predicting Creditworthy at 97% and Non Creditworthy 38%

Decision Tree – has a accuracy of predicting Creditworthy at 86% and Non Creditworthy 46%

Logistic Regression – has a accuracy of predicting Creditworthy at 87% and Non Creditworthy 48%

So all these models predicting Creditworthy more accurately than non creditworthy. So there is a bias across all the models



*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

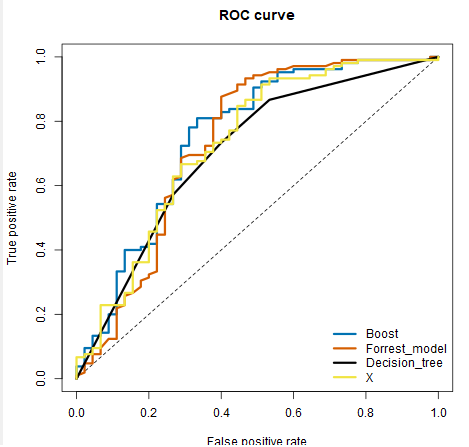
Forrest Model has overall accuracy of 79% which is higher than the other model and this model has been selected

Based on the ROC curve below Forrest model has the highest true positive rate compared to other models, hence forrest model has been chosen among other models.

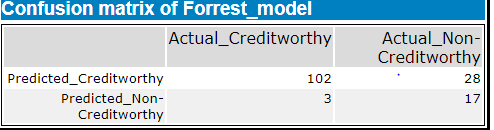
* + Accuracies within “Creditworthy” and “Non-Creditworthy” segments

Accruacies within “Creditworthy” is at 97% and “Non-Creditworthy” is at 3%

* + ROC graph



* + Bias in the Confusion Matrices



102 of Creditworthy and 17 Non-Credit worthy has been predicted correctly

28 are predicted as Creditworthy which is actually a Non-Creditworthy

3 are predicted as Non-Creditworthy which is actually a Creditworthy

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

* How many individuals are creditworthy?

408 Individuals are Creditworthy

**Before you Submit**

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/265/view) here. Reviewers will use this rubric to grade your project.