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?
* *A small bank which typically get 200 loan applications per week and approved by hand suddenly received nearly 500 loan applications due to a financial scandal that hit a competitive bank last week. As a loan officer, my manager wants me to figure out how to process all these loan applications within one week. Based on the classification models that I learned recently, I need to systematically evaluate the creditworthiness of these new loan applicants and provide a list of creditworthy customers to my manager in the next two days.*
* What data is needed to inform those decisions?
* *Data on all past applications*
* *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?
* *We need to use the Binary model to help make these decisions because we want to identify people who qualify and do not qualify for loans for this problem.*

# 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.*

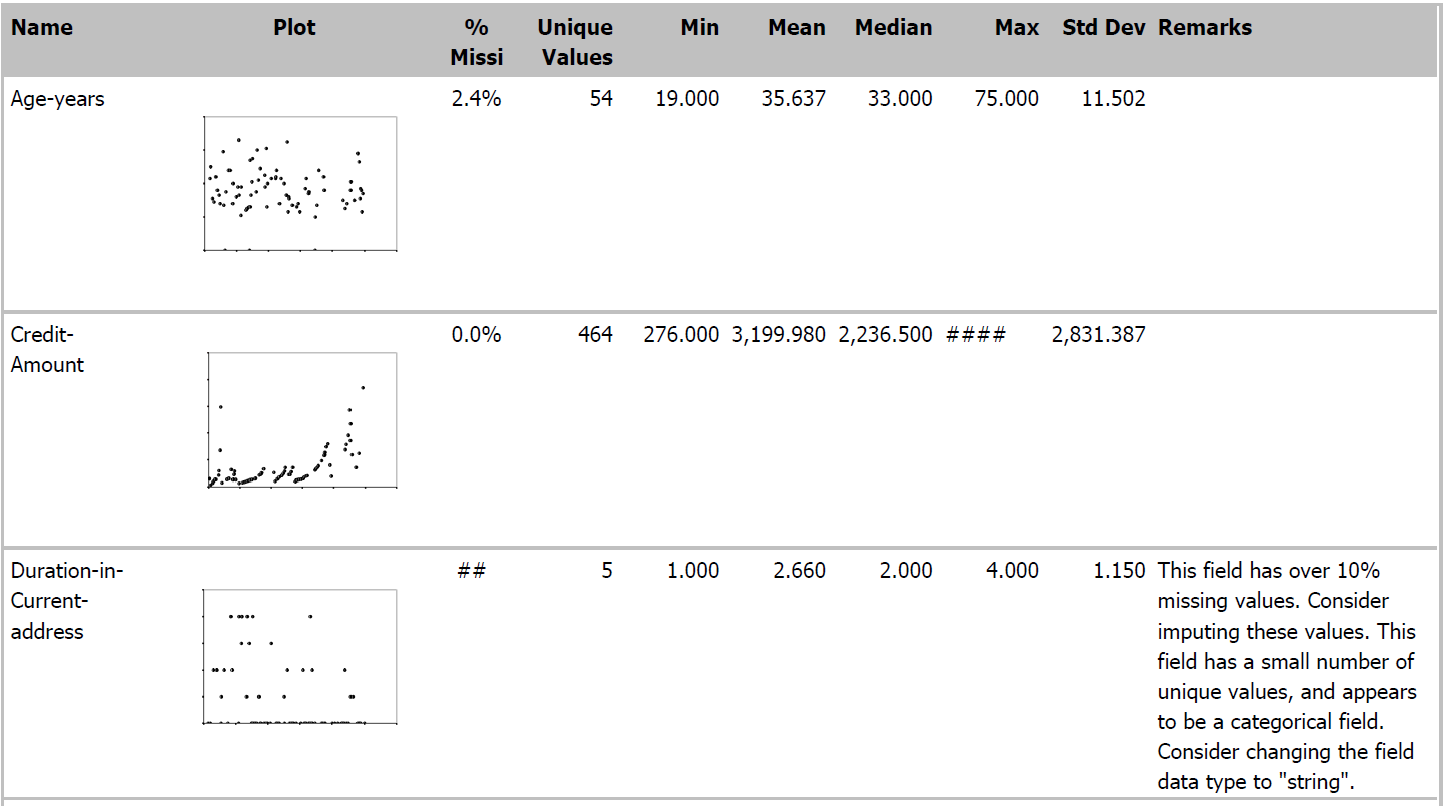
*Answer this question:*

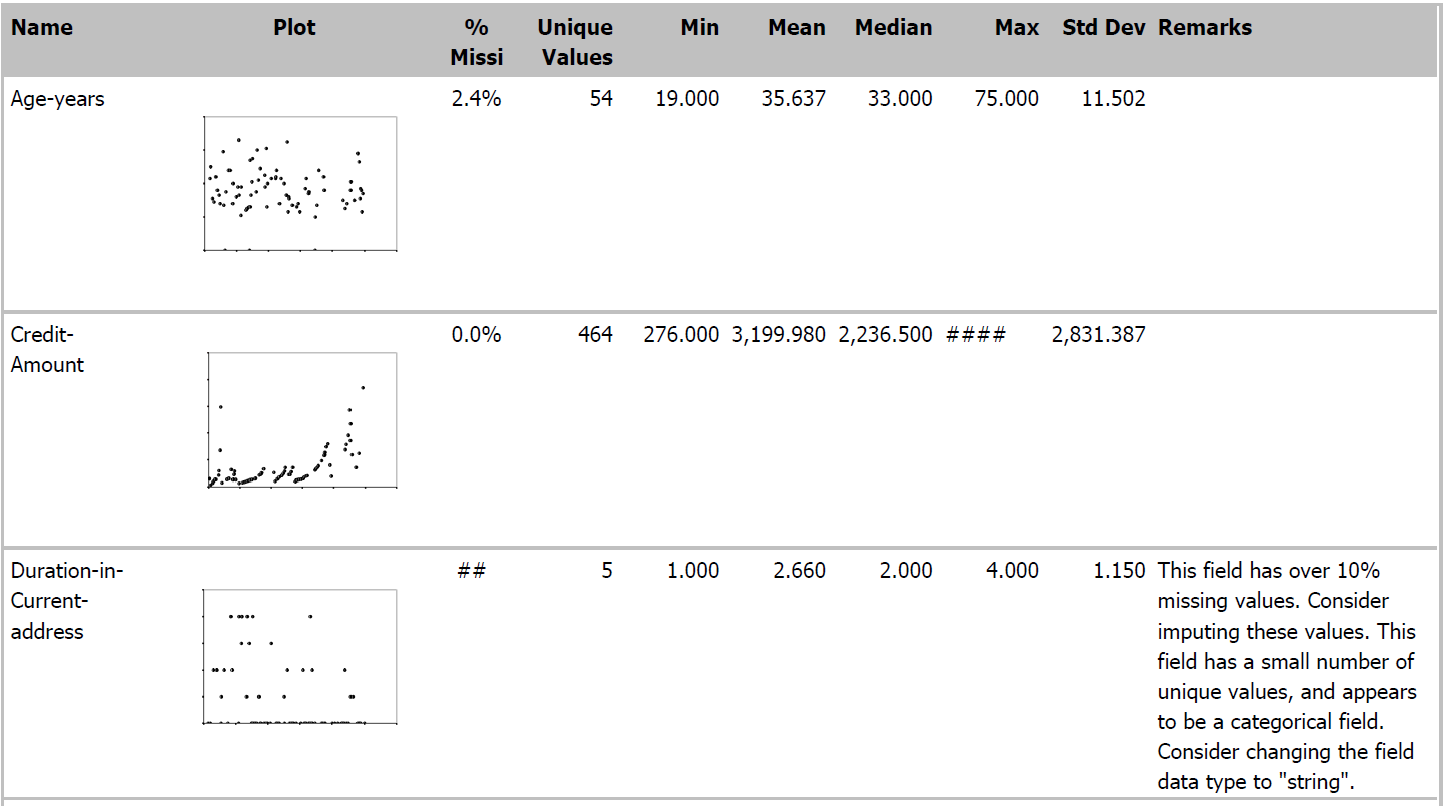
* In your clean-up process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

***Data clean-up***

*Generate the summary of all the fields and analysed based on the presence of null values and the measure of variability. Below are the columns are focused for removal/imputation:*

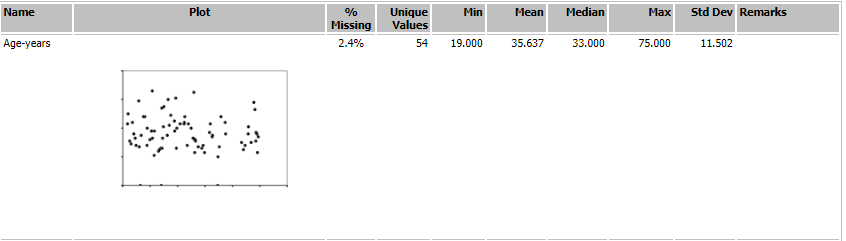
1. ***Duration-in-Current-address****: We have 68.8% of data as NULL values which is very high so we will drop this column.*



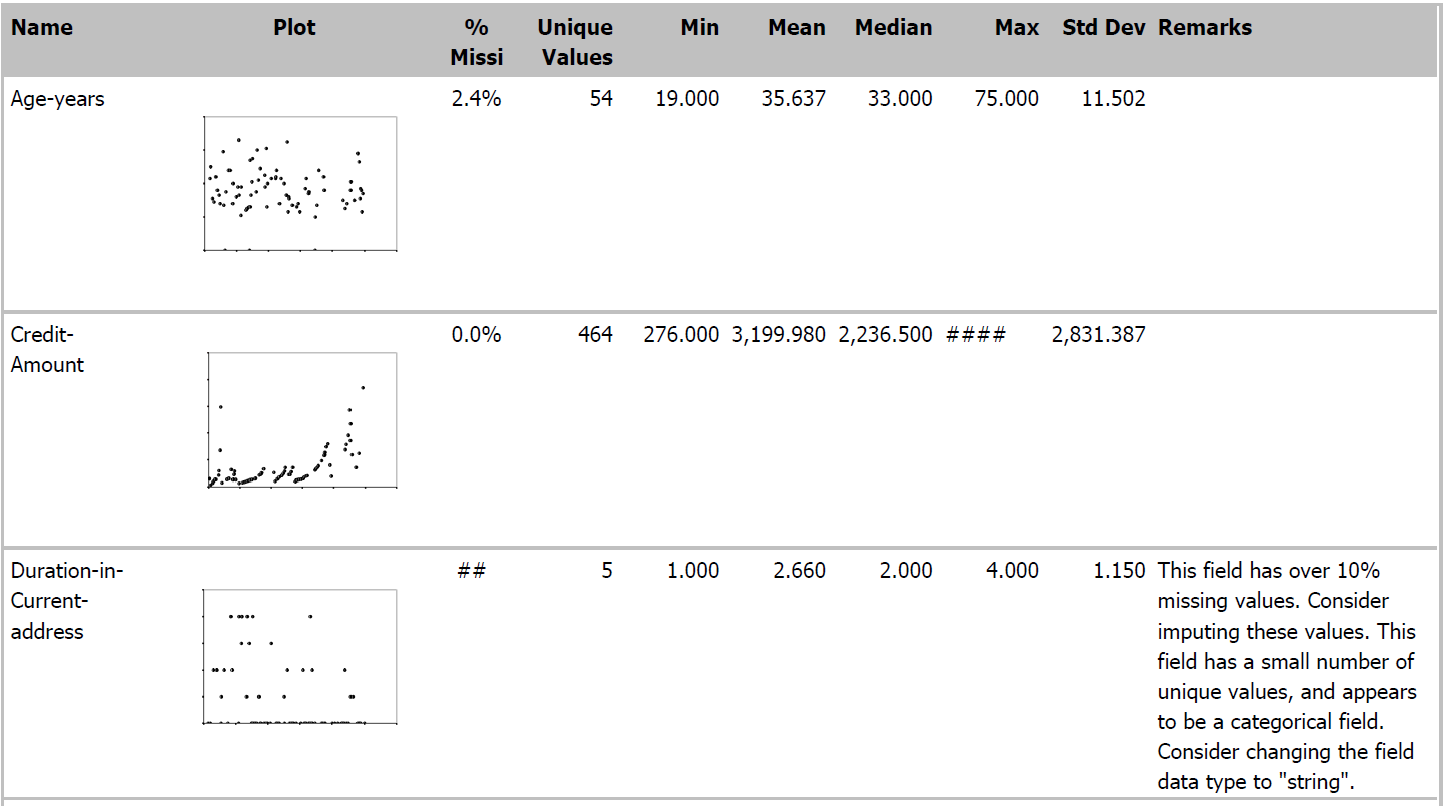


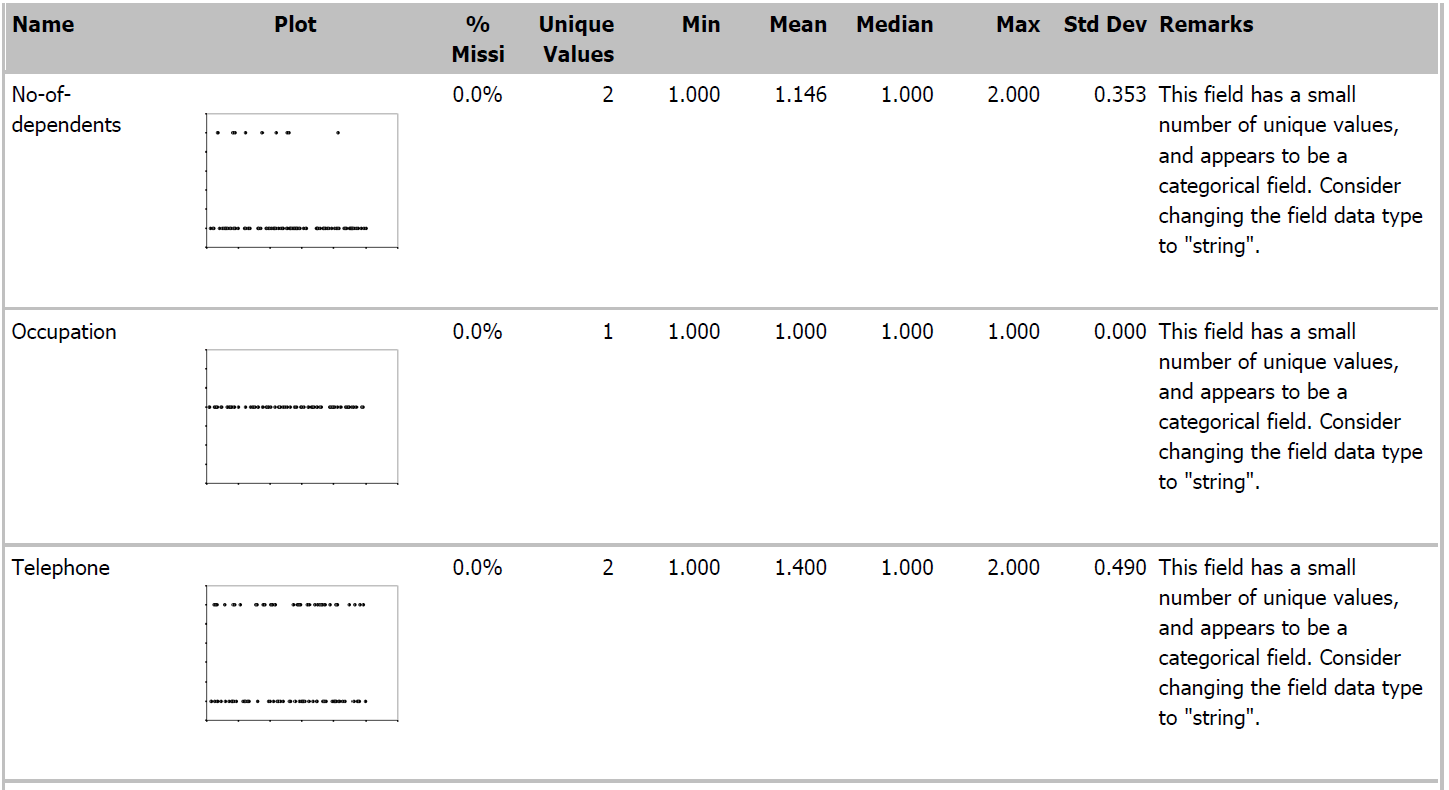
1. ***Age-years:*** *We have 2.4% of data as NULL values so we will impute with Median=33. Because Age should be rounded values and imputing with average getting us a decimal value and rounding can cause an impact in the data. Also, Median is preferable to use to*

*impute for nominal categorial variable and not affected by outliers.*

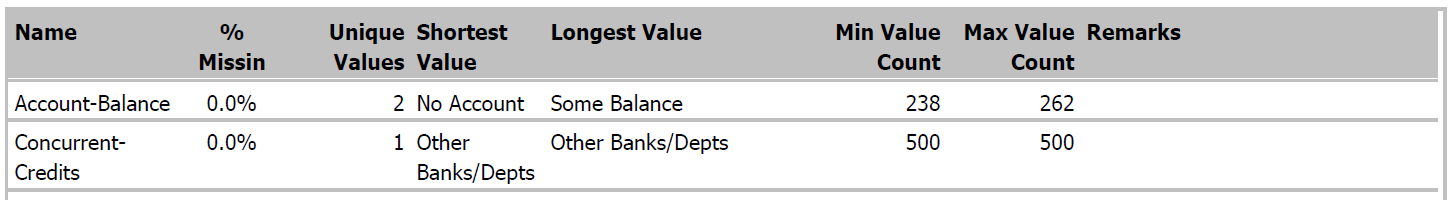


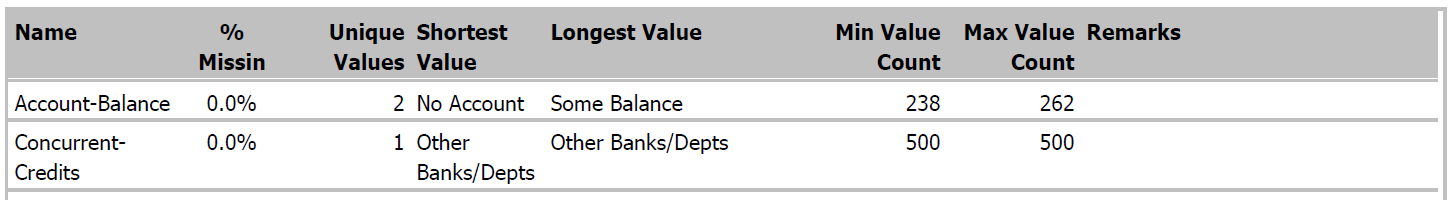
1. ***Occupation:*** *This field has only value in all the rows and there is no variance present. Hence, would not be useful for our target variable and can be dropped.*



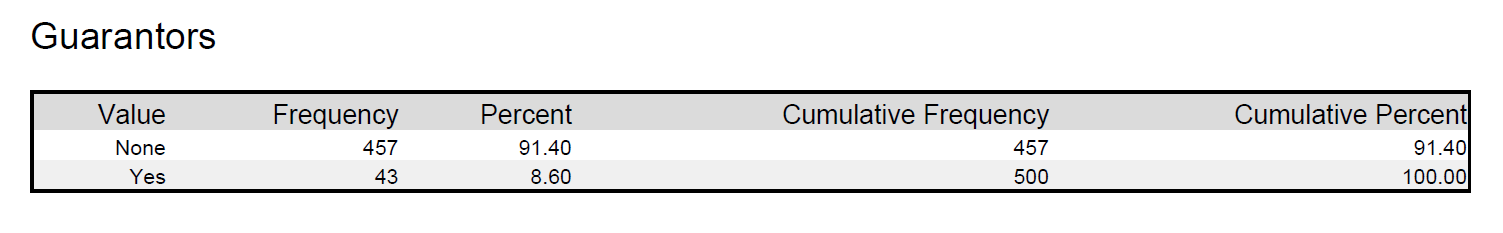


1. ***Concurrent-Credits****: Only one value is present throughout the column as Unique values = 1. Like Occupation, this field can be dropped.*

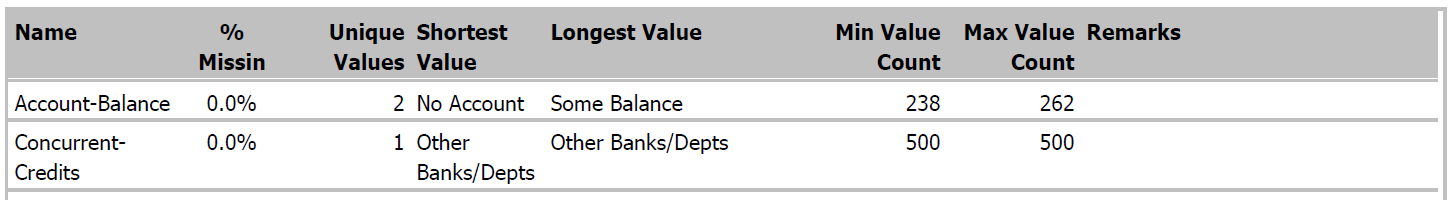


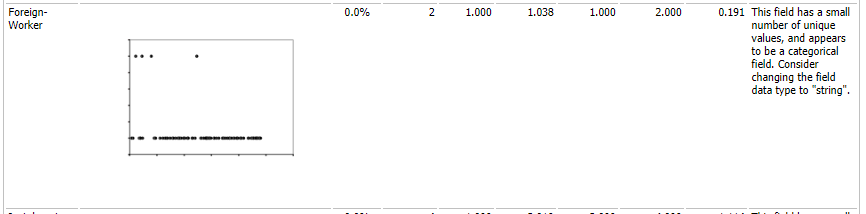


1. ***Guarantors:*** *We have skewness in this column data. We have only 8.6% values as “Yes”. Hence, this column can be dropped.*

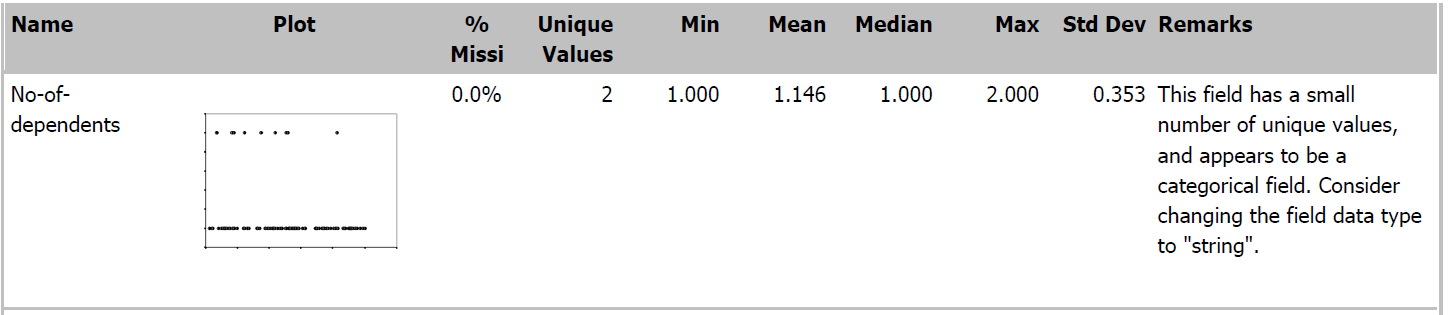


1. ***Foreign-Worker: T****his variable shows that most of the data is skewed toward “1”. Hence, this column can be dropped.*





1. ***No-of-dependents:*** *We have skewness in this column data towards “1”. Hence, this column can be dropped.*



1. ***Telephone:*** *This**field should be removed as well because telephone number is not having any correlation with the creditworthiness of the applicant.*

***Summary – Data Clean-up***

* + *We will remove total 7 columns:*

*Duration-in-Current-address, Occupation, Concurrent-Credits, Guarantors, Foreign-Worker, No-of-dependents, and Telephone.*

* + *Perform Imputation in 1 Column: Age-years with median value.*

# 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.*

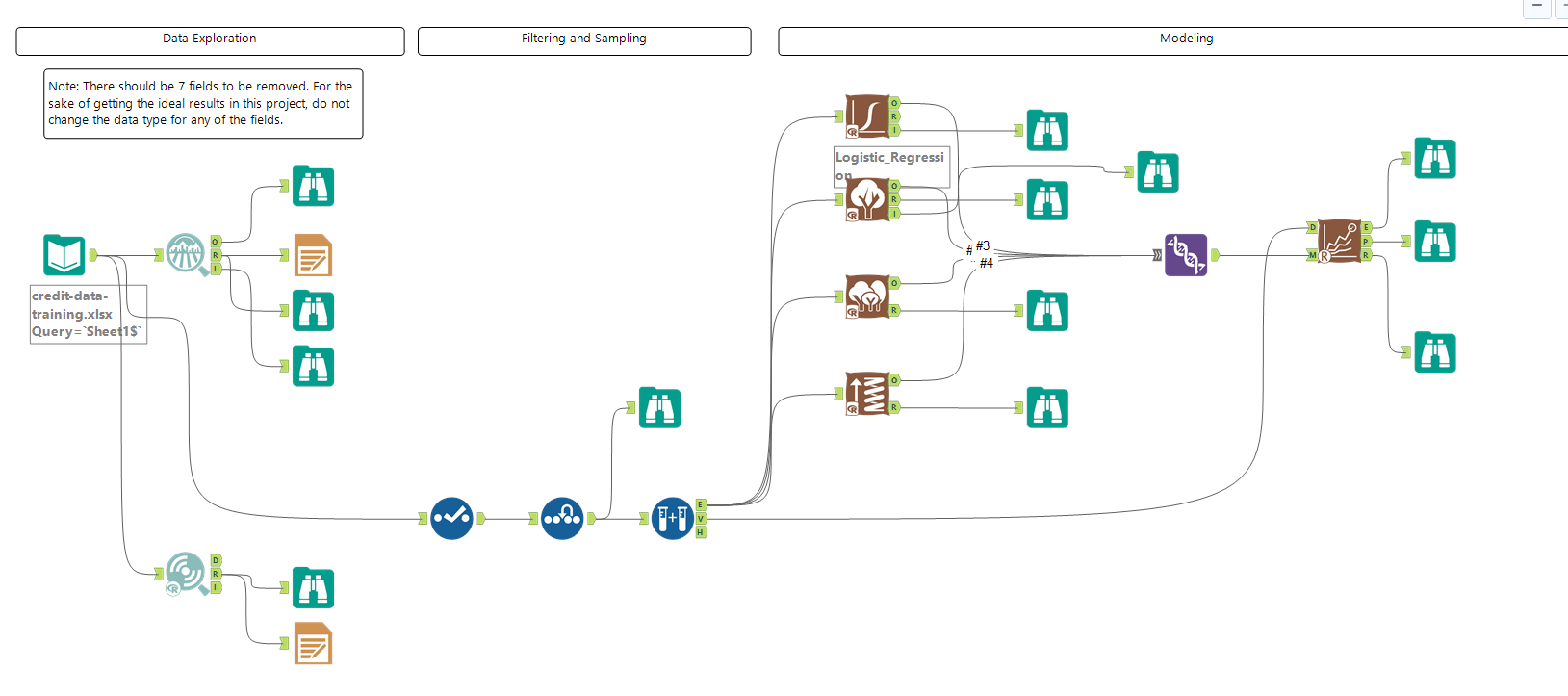
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

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

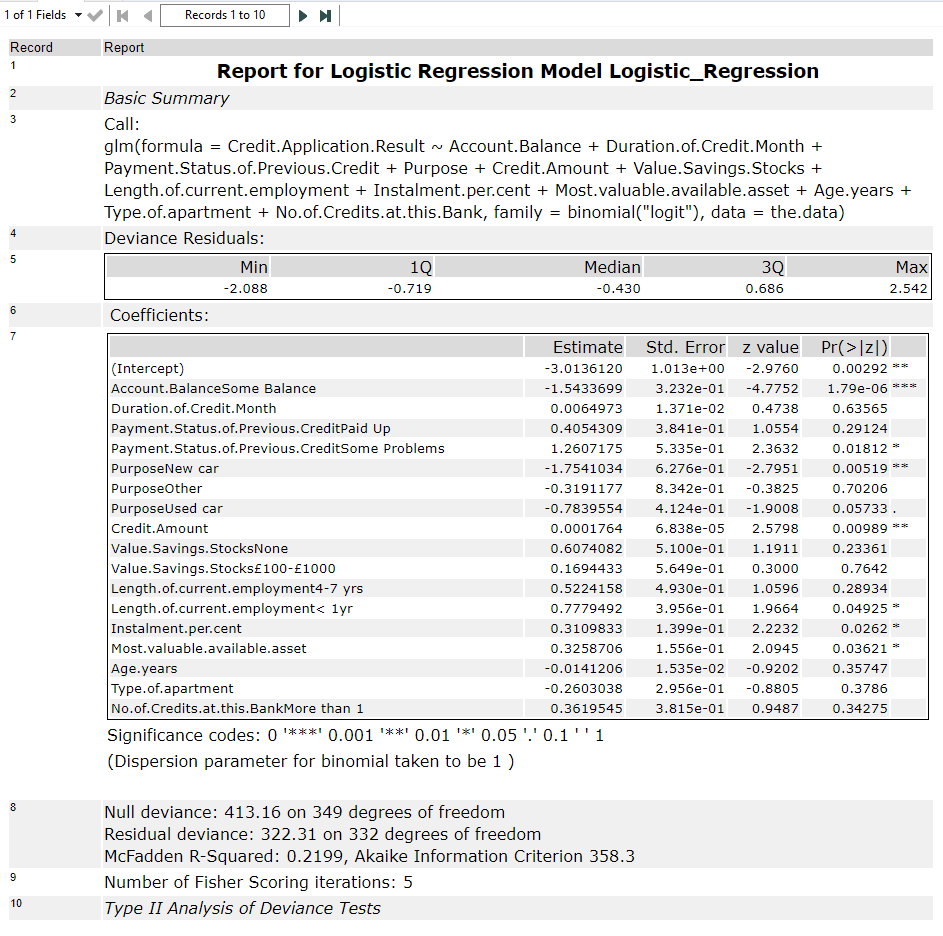
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

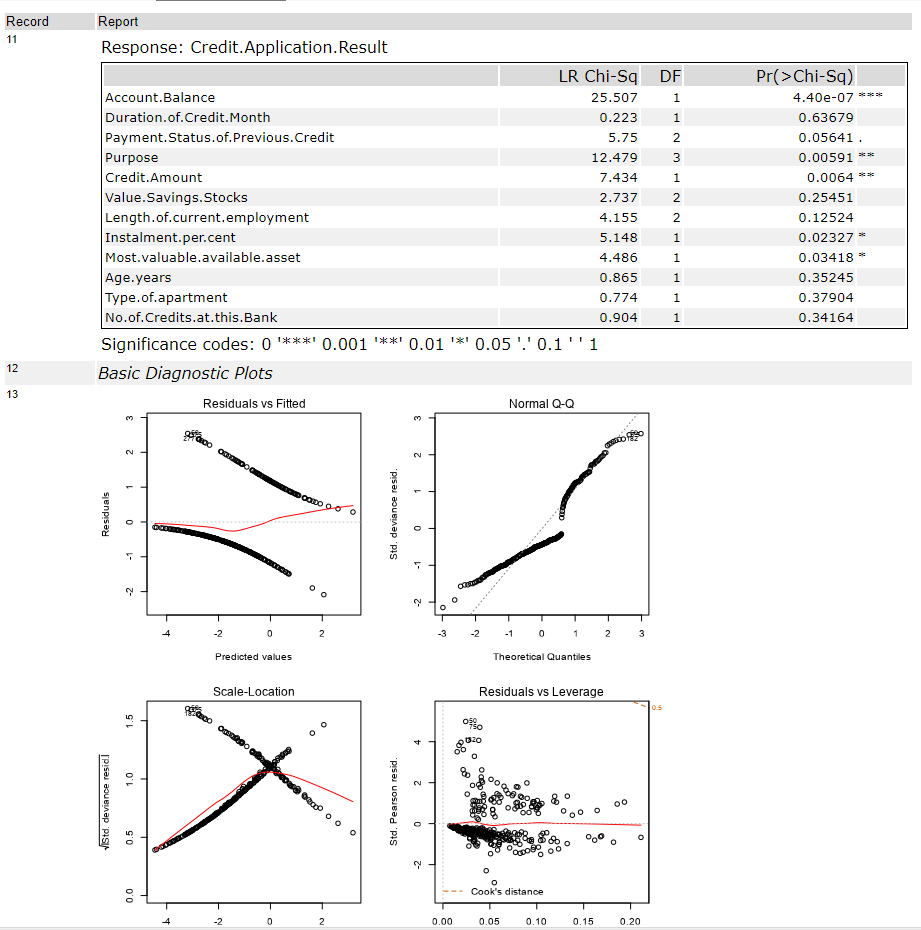
*You should have four sets of questions answered. (500 word limit)*

***Model Training:*** *After data clean-up and preparation, we are ready to move to Model training. Target variable – “Credit-Application-Result” is a binary variable so we would perform classification with the below 4 algorithms:*



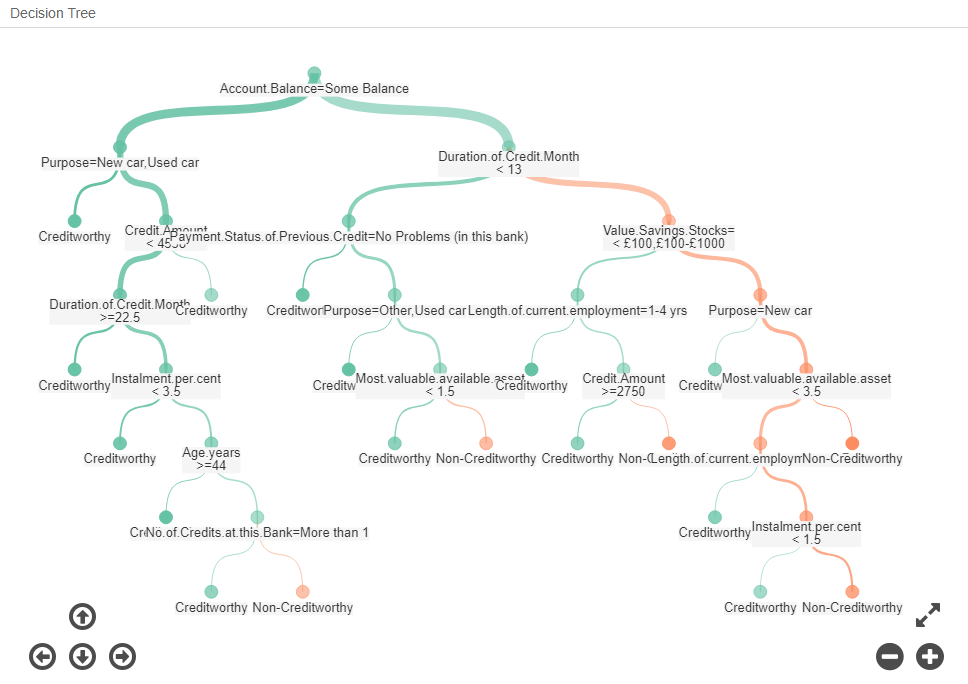
1. *Logistic Regression: Model details are given below*



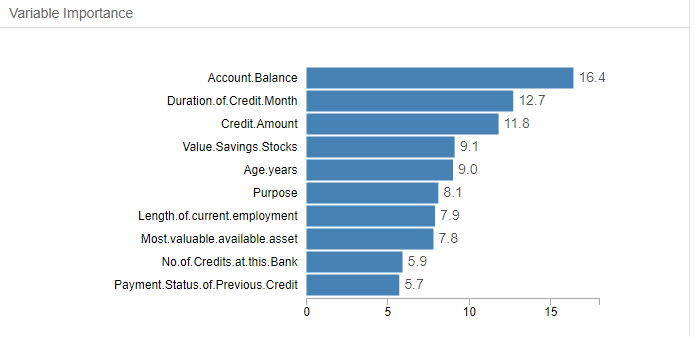


*Based on the Model generated and p-value for each column, we can conclude that the most important variables are Account Balance, Purpose, Credit.Amount, Instalment.per.cent, Most.valuable.available.asset*

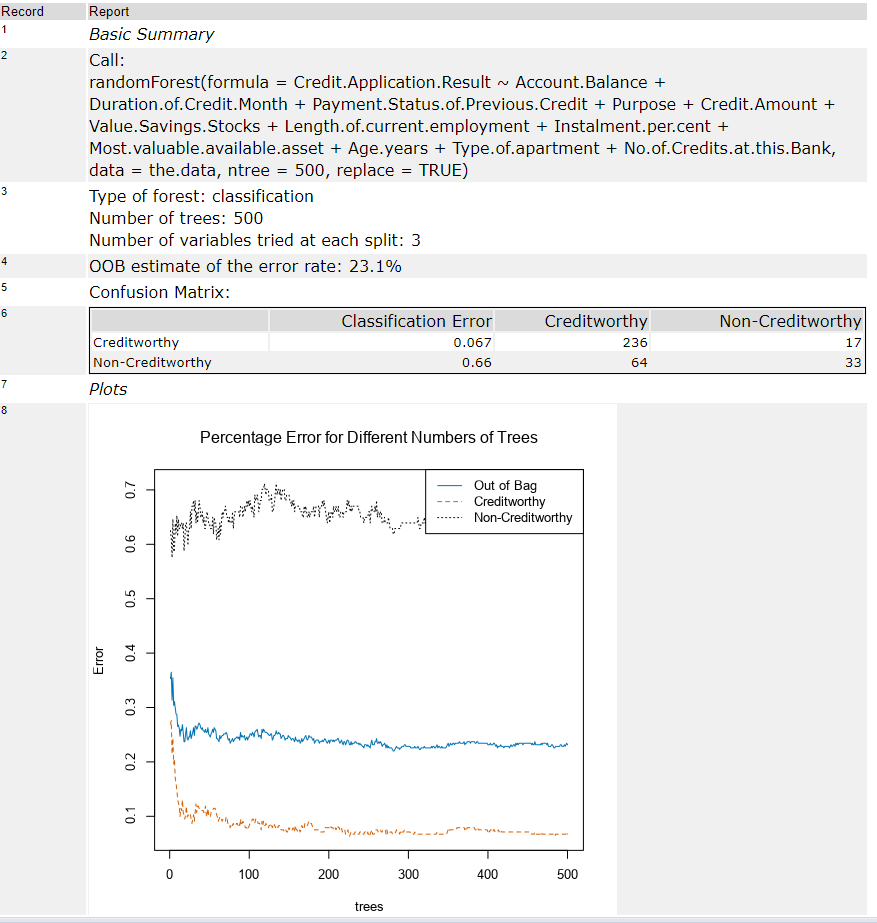
1. *Decision Tree*



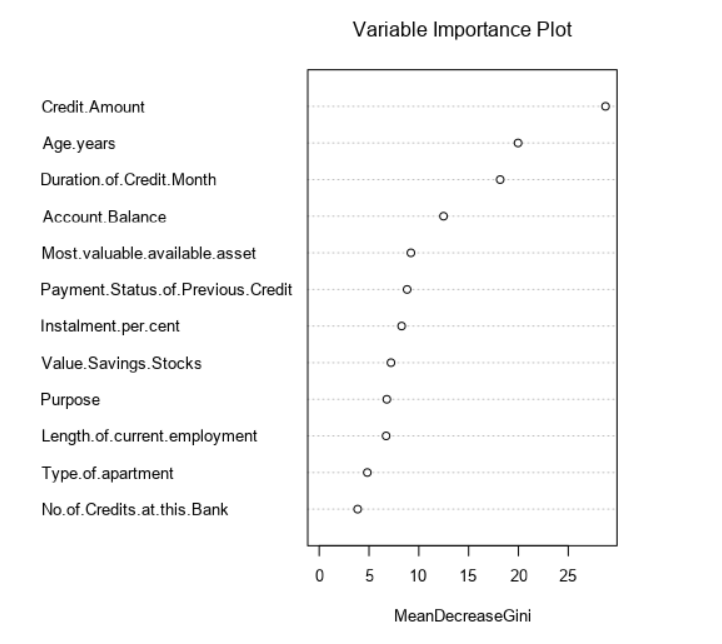
***Variable importance:*** *Report below shows that the top 5 predictive variables are Account Balance, Duration of Credit Month, Credit.Amount, Value.Savings.Stocks and Age.years.*



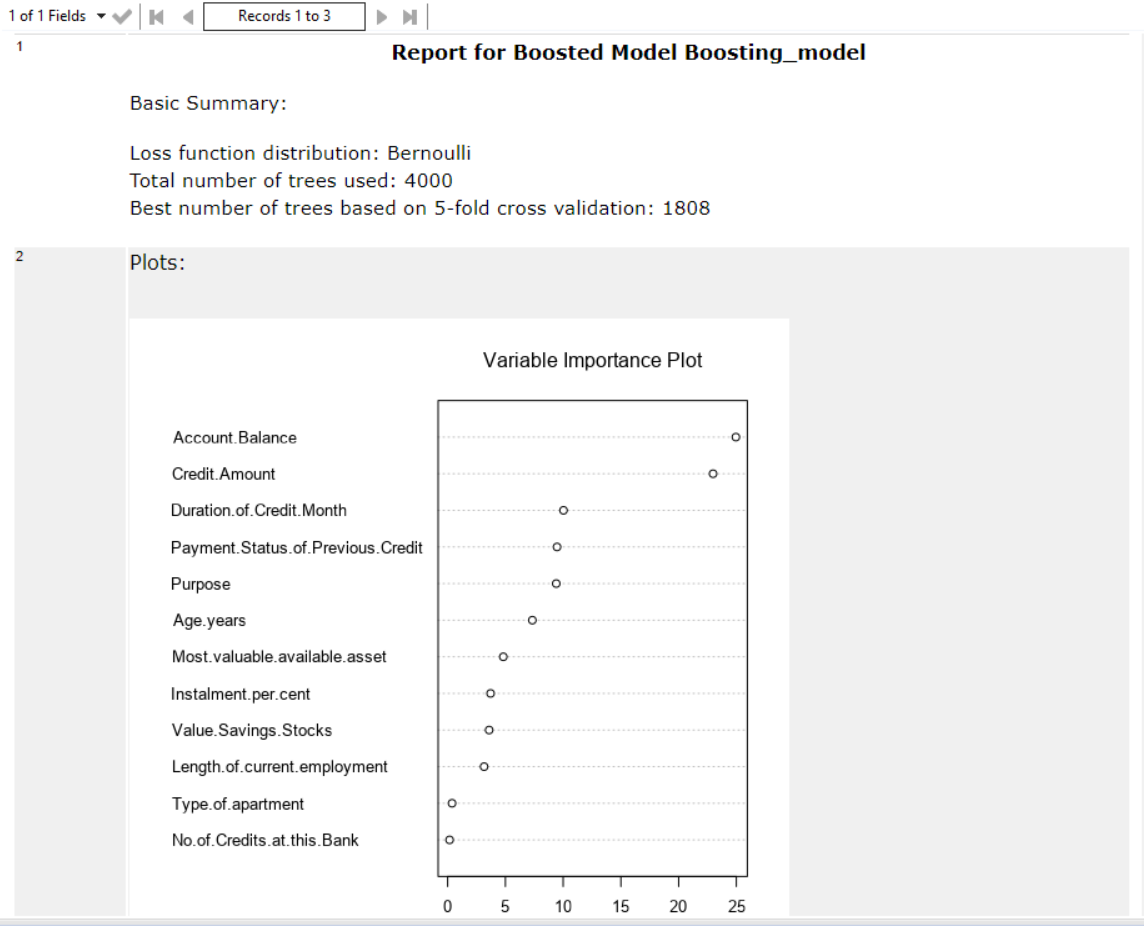
1. ***Forest Model:*** *Below Random forest model is generated:*

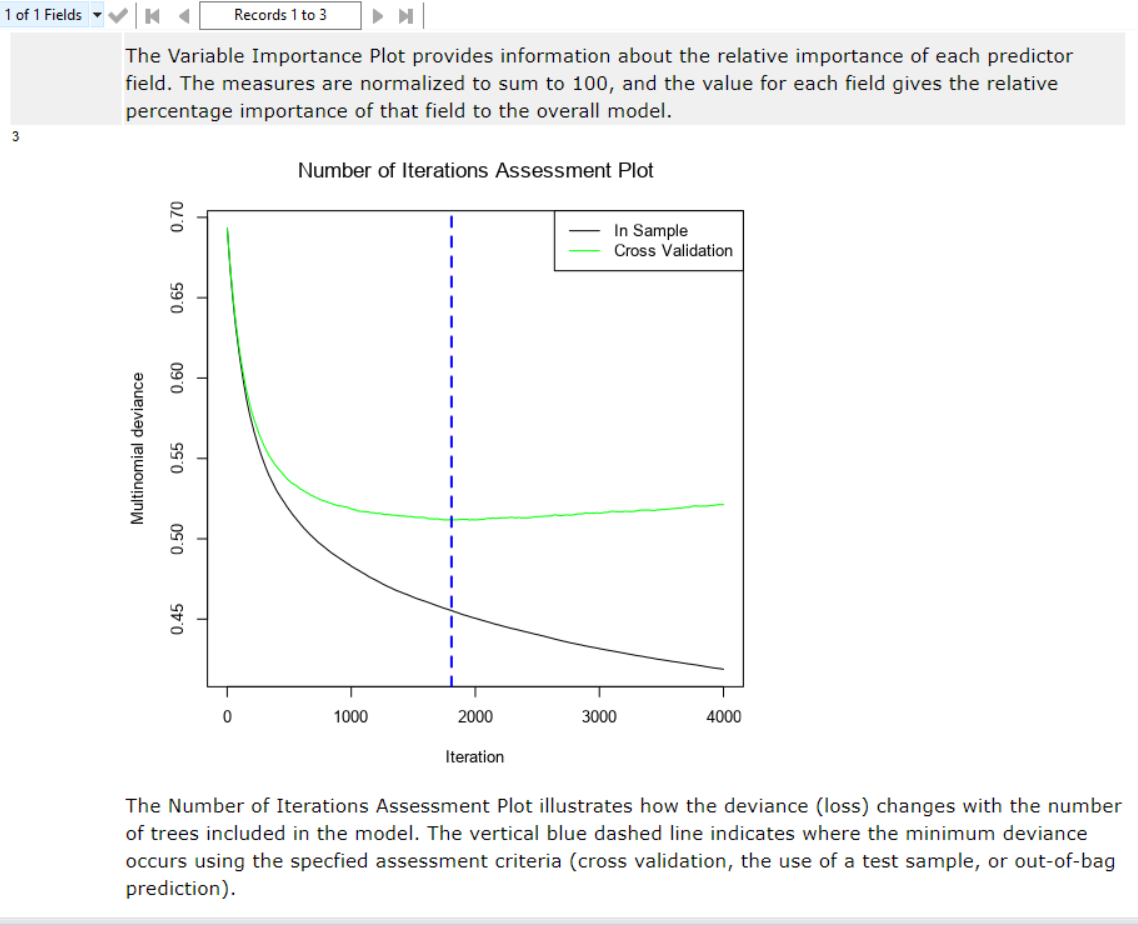


***Variable Importance:*** *Based on the plot below, the top 5 predictive variables are Credit.Amount, Age.years, and Duration.of.Credit.Month, Account.Balance and Most.valuable.available.asset.*



1. ***Boosted Model:*** *Model generated has below details:*

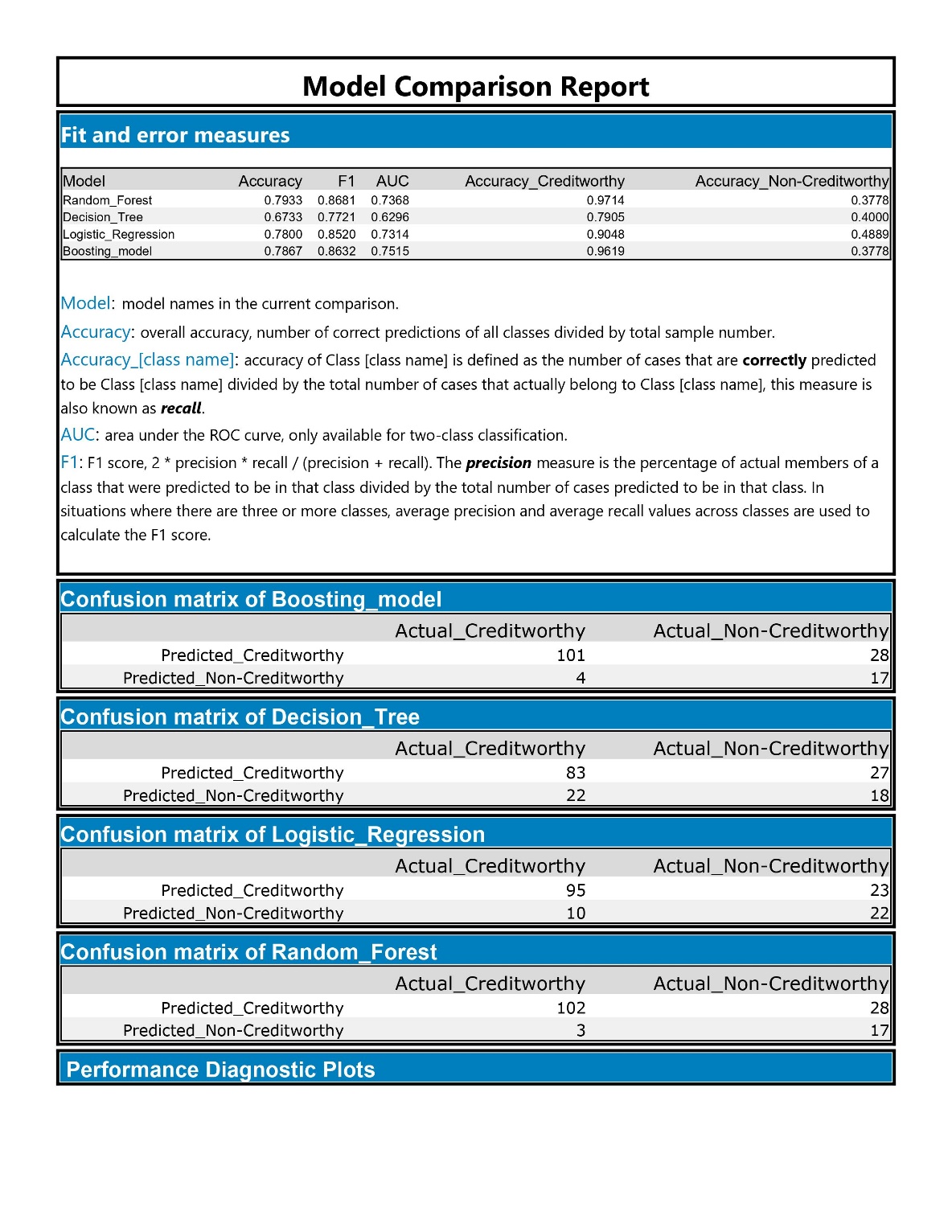




***Variable Importance:*** *Based on the model above, 5 most important variables are Account.Balance, Credit.Amount, Duration.of.Credit.Month, Payment.Status.of.Previous.Credit and Purspose.*

***Model Comparison***

*We have imported a macro “Model\_comparison” to compare the 4 models we build and generated the report.*

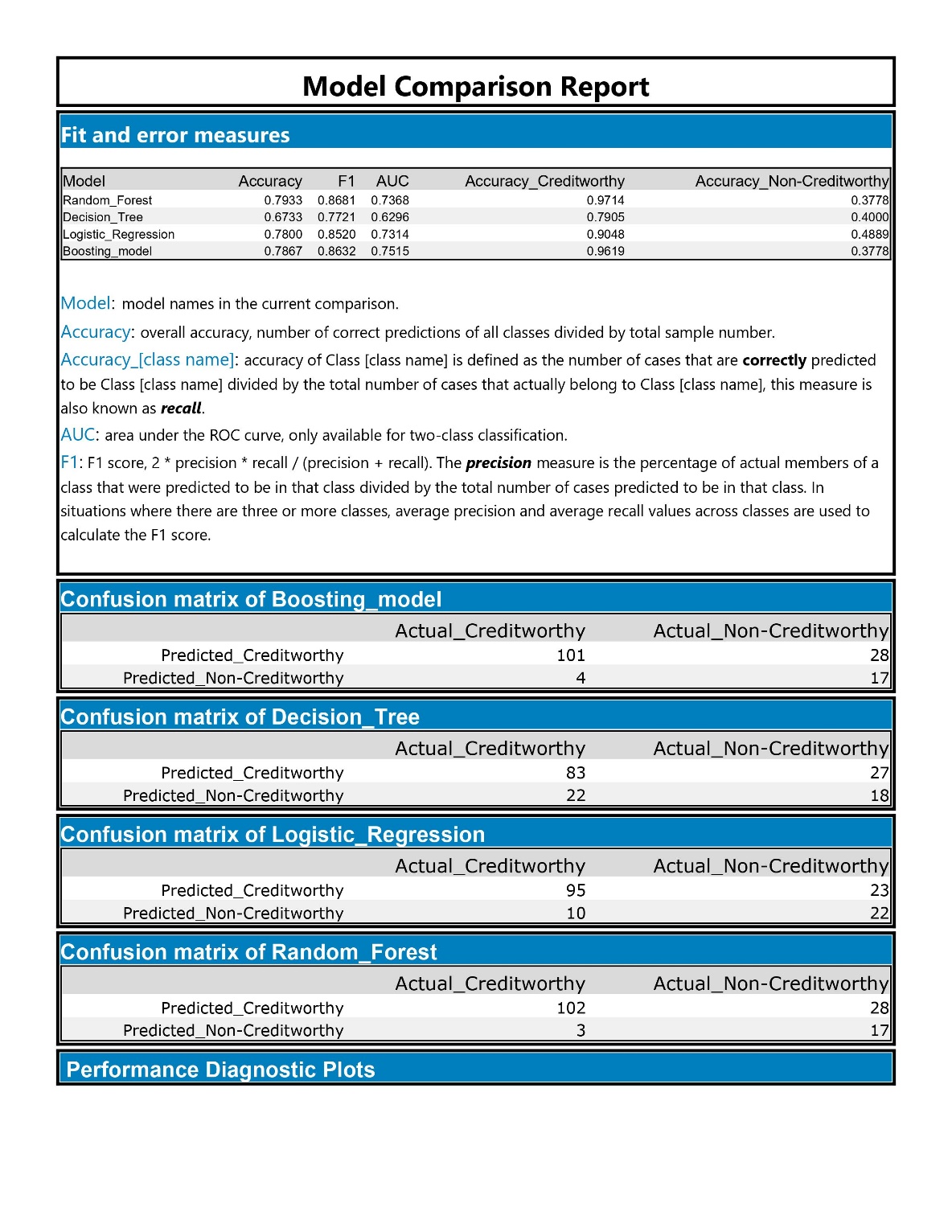


*Above report suggest that the overall accuracy for*

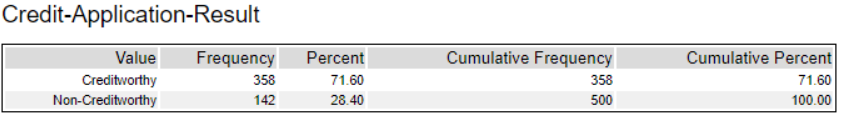
***Random forest > Boosting > Logistic regression model > Decision tree***

*And same pattern is present with F1 score.*

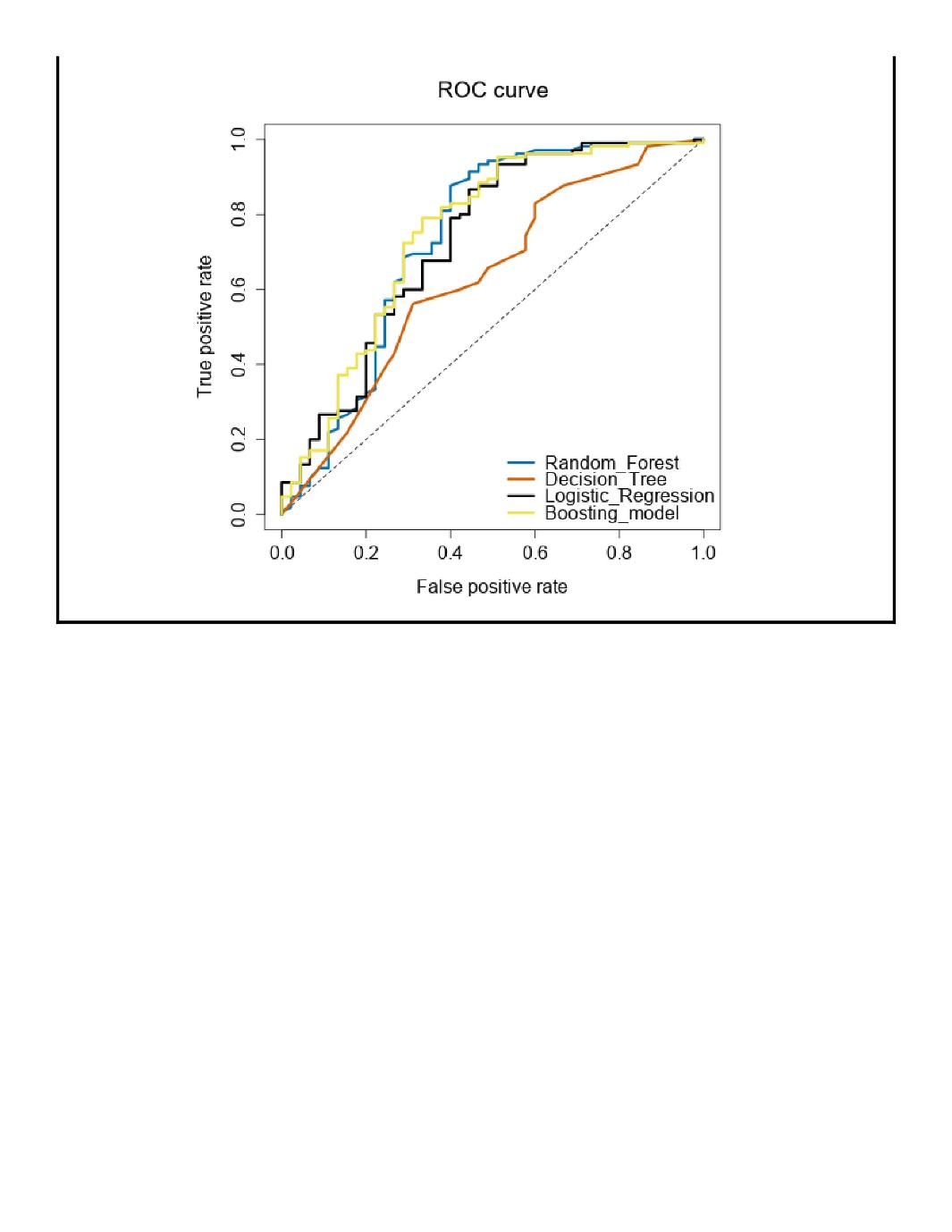
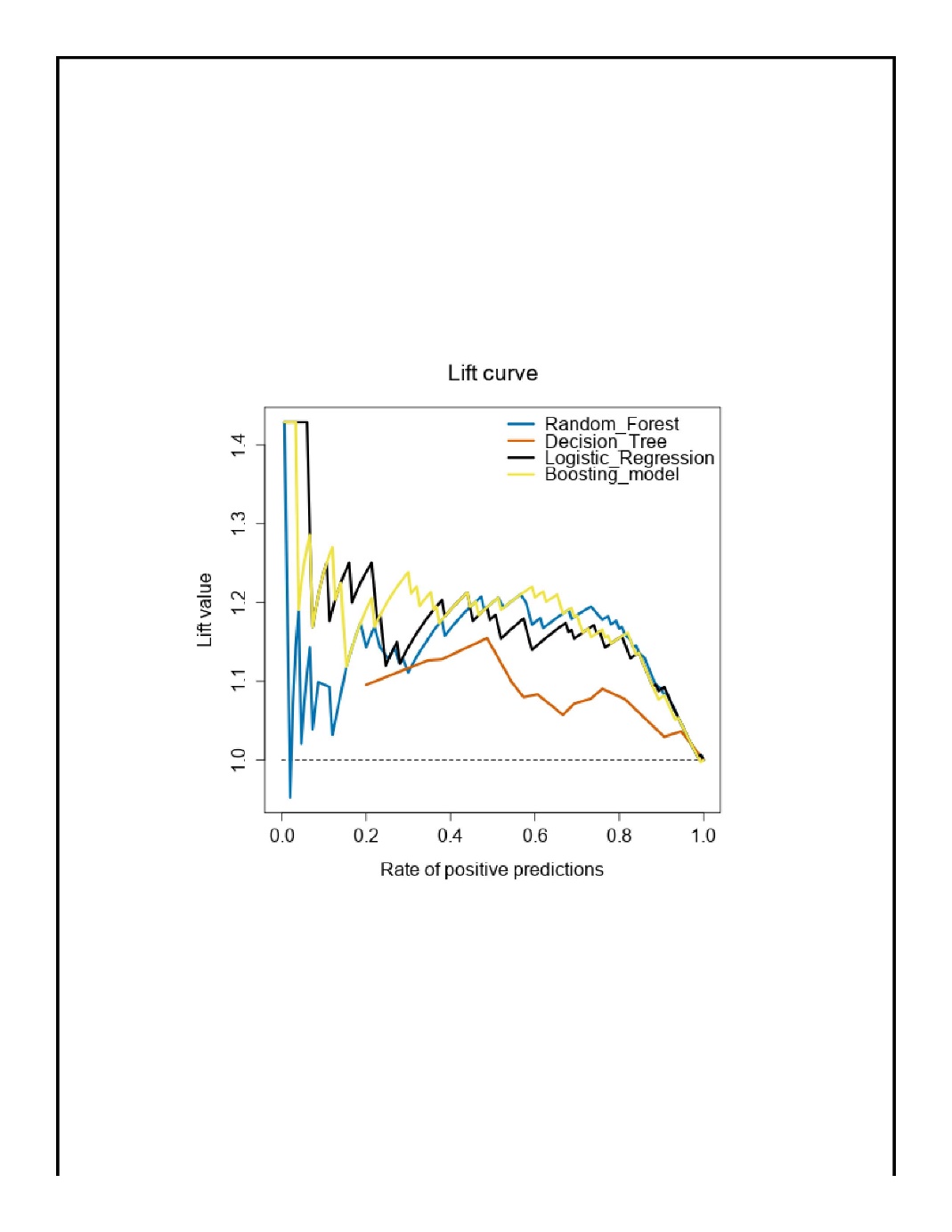
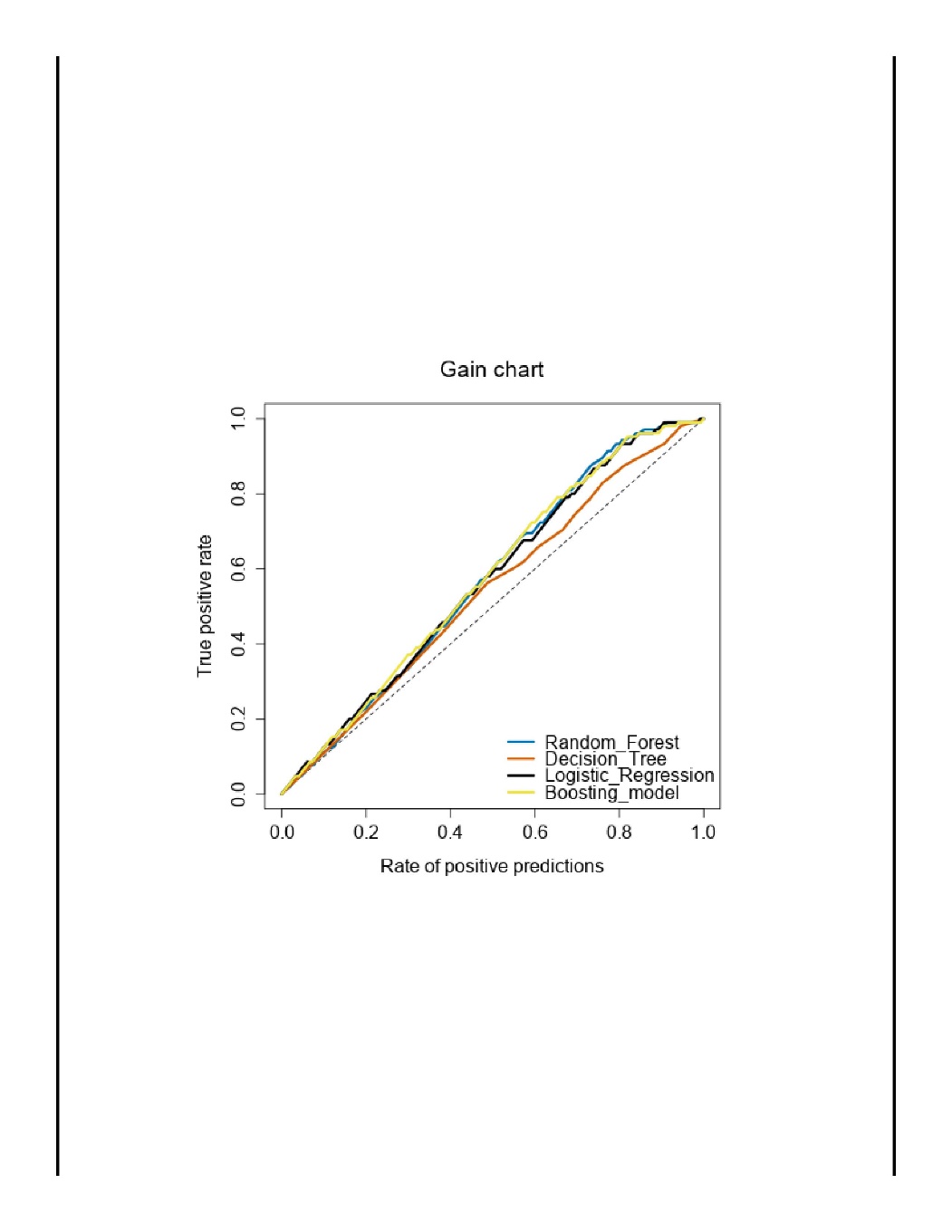
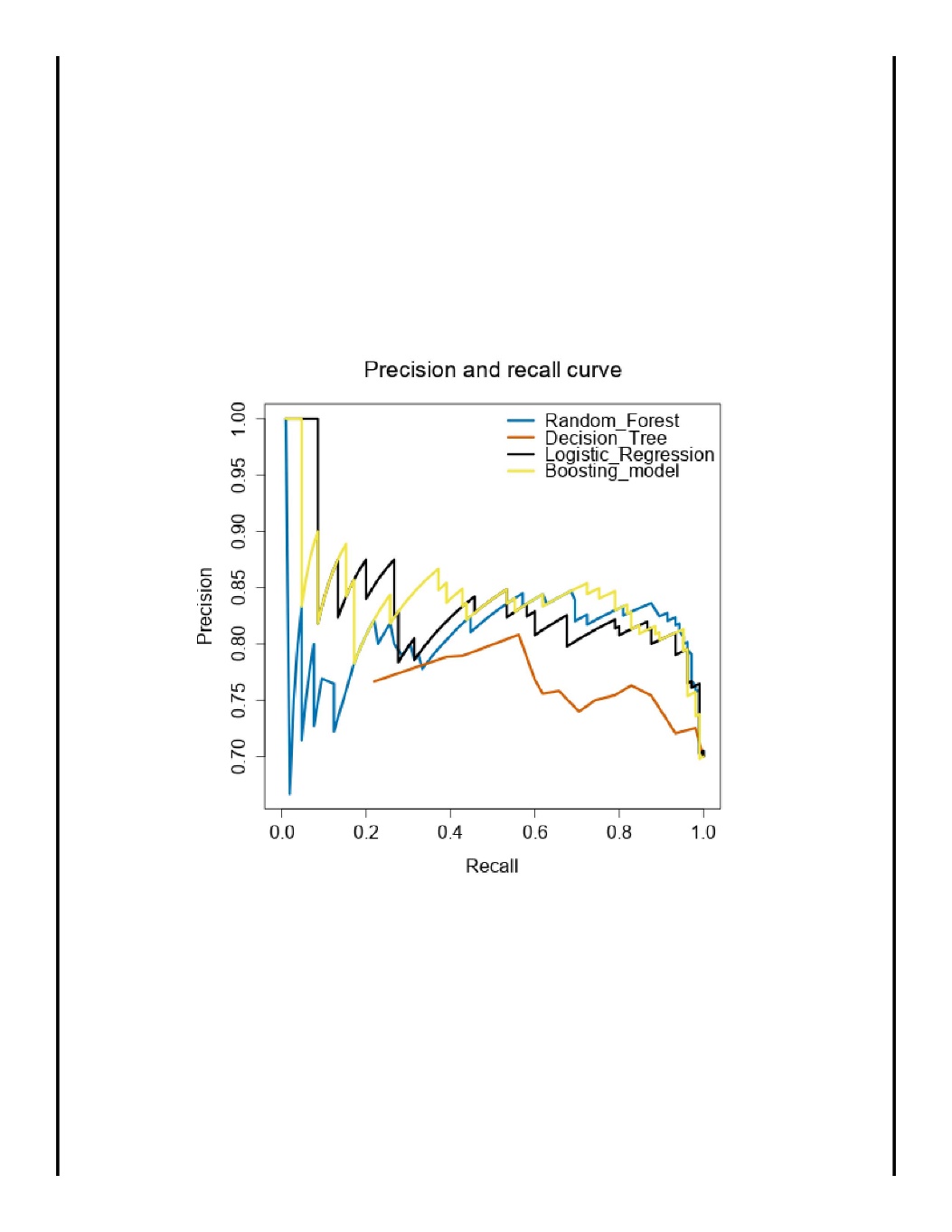
*Confusion matrix for all these models:*



*In the training dataset, we have an imbalanced dataset in terms of target variable.*



*There are 71.6% creditworthy applicants while 28.4% are non-creditworthy. Validating and comparing the models just based on the accuracy would not suffice as we may have high bias in the positive class so we would also use F1 score to compare models and choose accordingly.*



*Looking at the report above, we can say that there is a bias present in the prediction towards creditworthy which is easily justified as our training data also contains the similar bais.*

# 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”*

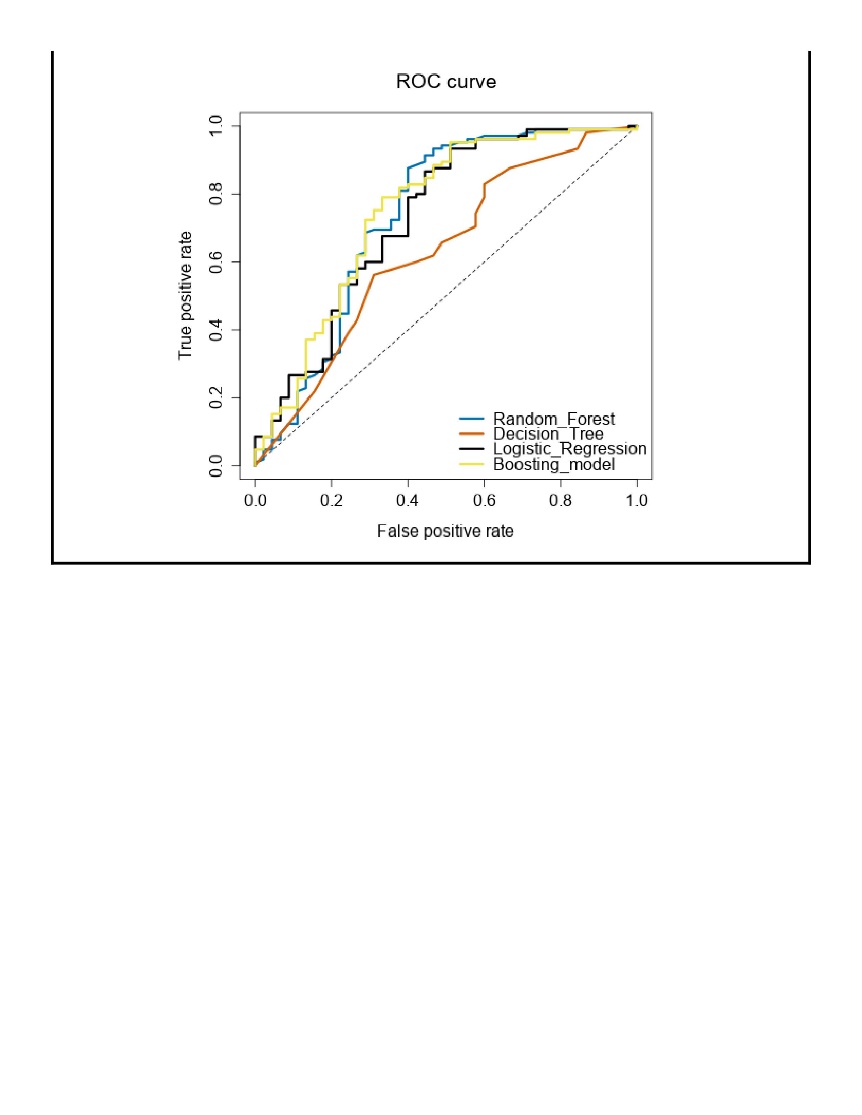
*Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)*

*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

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

* How many individuals are creditworthy?

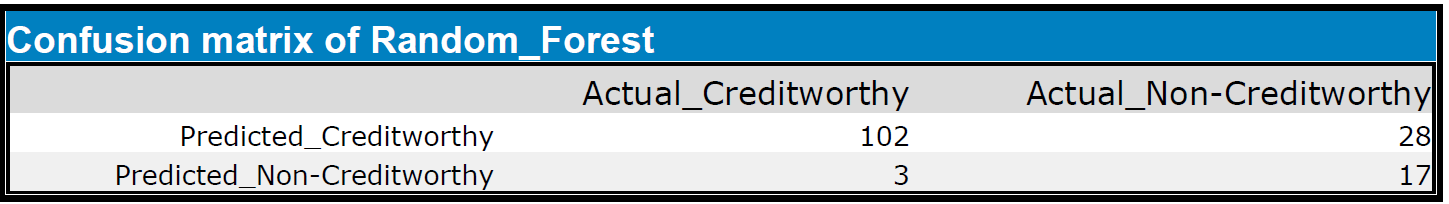


***Model Selection and Prediction:***

*Comparing the Accuracy and F1 score of all the models we have built, we would select the model “Forest Model” which has highest F1 score as .8681 and highest accuracy of 79.33%.*

*Looking at the ROC curve Boosting model and Forest model show good performance but combing above, we may keep forest model as the best model for this dataset.*

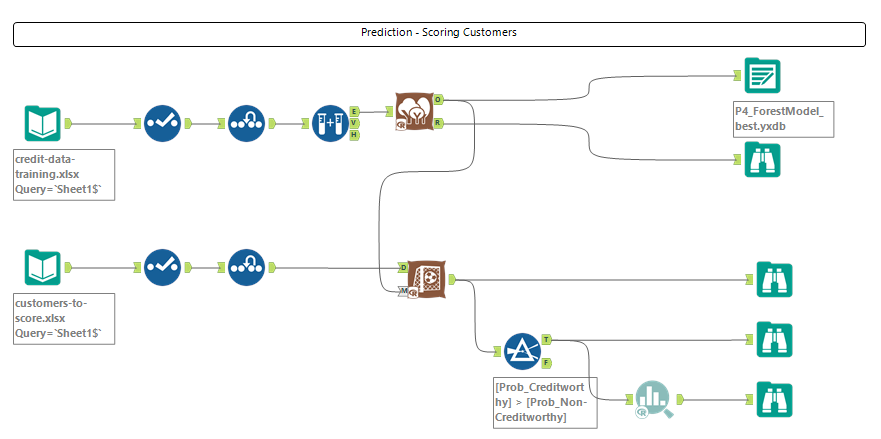
*We would now understand the confusion matrix for Random Forest (Forest Model):*



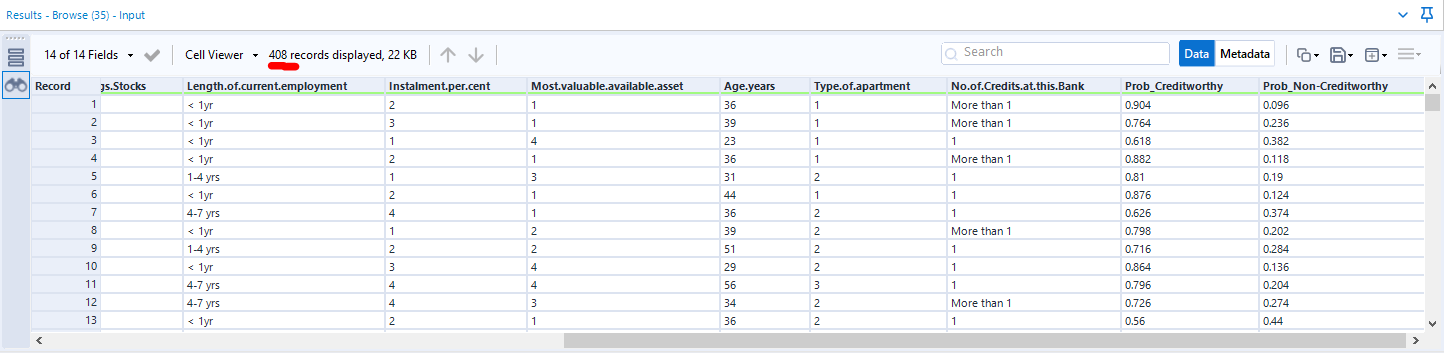
*We have correctly predicted 102 cases as Creditworthy and 17 cases as non- creditworthy.*

*Overall, we can also see that higher ratio of predictions are observed in creditworthy cases hence we have a bias present.*

*To perform the prediction on the new data/cases, we have built a workflow using the Forest model.*



*Apply the trained forest model on the new customer data and when probability of creditworthy is higher than probability of non-creditworthy, we mark the customer profile as Creditworthy.*



*We get total 408 customers who 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.