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*

1. What decisions needs to be made?

The company wants to classify new customers for loan approvals.

1. What data is needed to inform those decisions?

For Loan approvals, we usually need data like

Account balance

Age of the customers

Credit information

Total Debt information, previous loans from other banks

Income information

Investments in terms of stock, assets

Employment information

Surety information

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

Typically, when a Yes or No decision we use Binary models but for a two-variable classification we can use both Binary models and Non-Binary models with two variables.

## 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 average of the entire data field instead of removing a few data points. (100 word limit)*

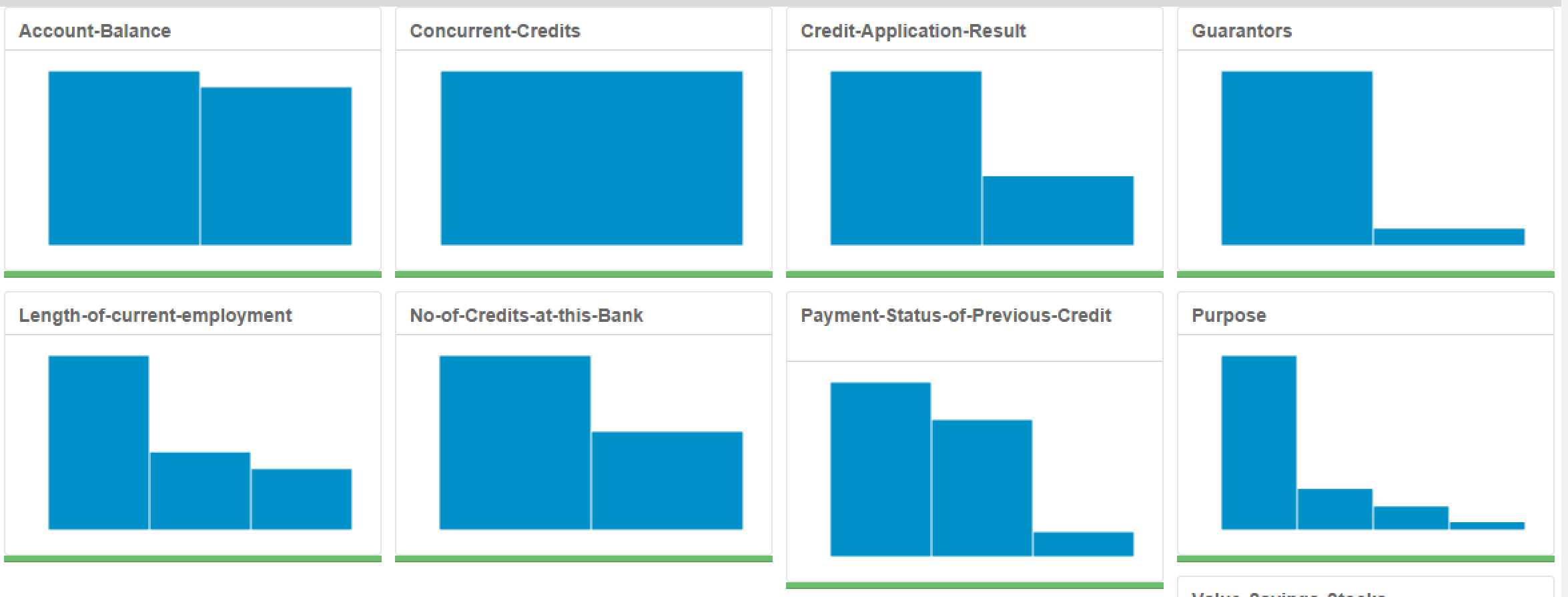
***Note****: For students using software other than Alteryx, please format each variable as:*

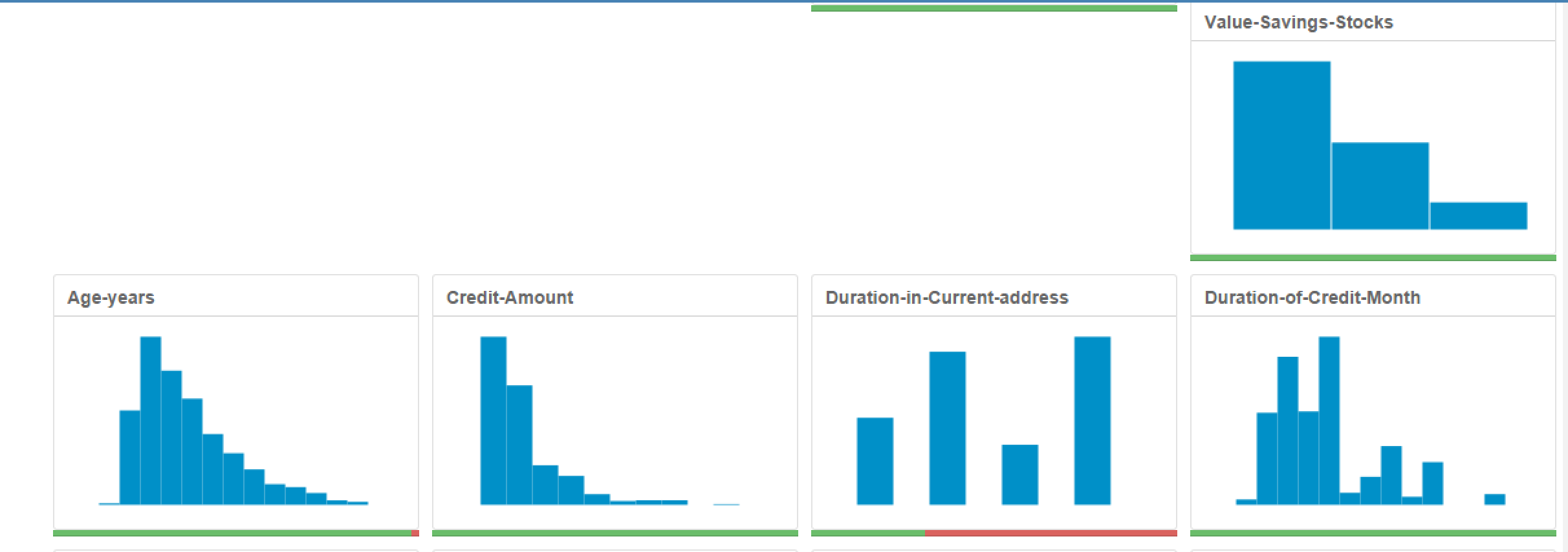
|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Credit-Application-Result | String |
| Account-Balance | String |
| Duration-of-Credit-Month | Double |
| Payment-Status-of-Previous-Credit | String |
| Purpose | String |
| Credit-Amount | Double |
| Value-Savings-Stocks | String |
| Length-of-current-employment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

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







Variables Removed –

Concurrent Credits: Reason – Low variability.

Guarantors: Reason – Low Variability.

Duration in Current Address: Reason – 70% of the data missing.

Occupation: Reason – Not an insightful variable.

No of dependents: Reason – Low Variability

Telephone – Low Variability

Foreign Worker – Low Variability

Imputation –

Age – 2% of the data was missing. Because the data was less than 5% we can remove those fields. But I have imputed this 2% of this data with the average of the Age field.

## Step 3: Train your Classification Models

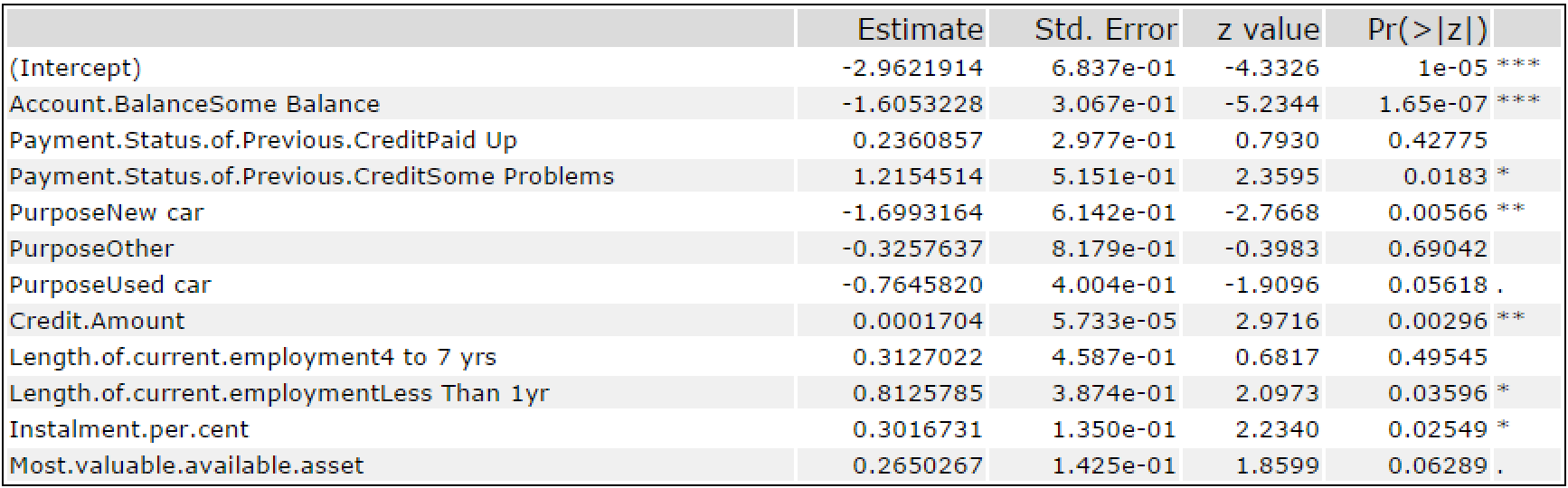
*First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.*

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

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

1. 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 Significant Variables –



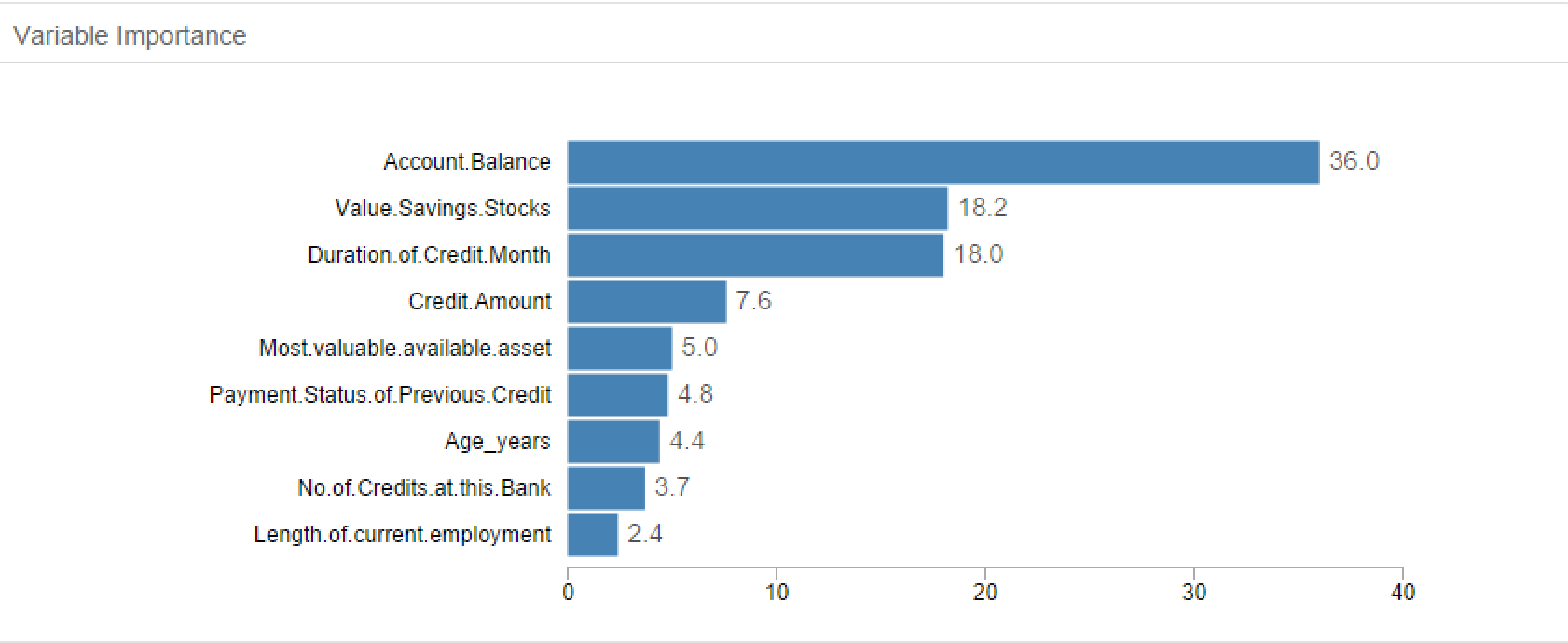
Account balance

Purpose – New Car, Other

Credit Amount, Installments

Current Employment – Less than 1 year.

Decision Tree Significant Variables –

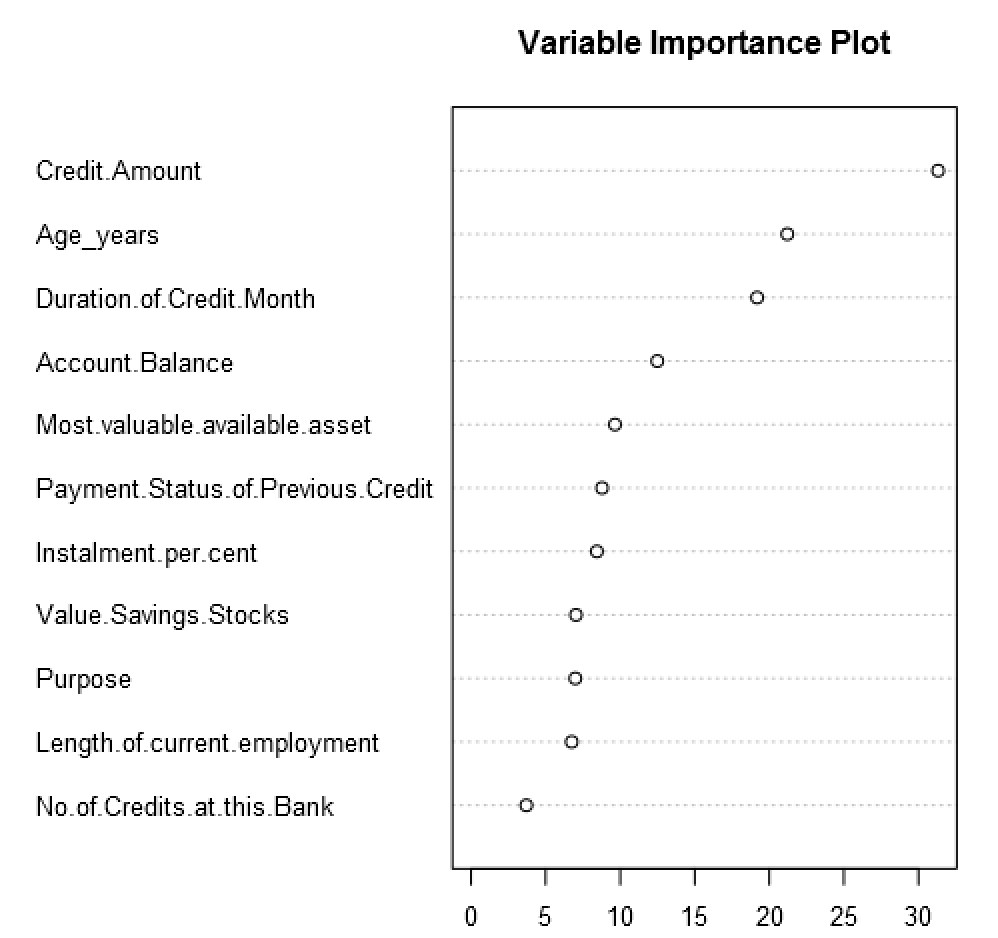


Account Balance

Stocks

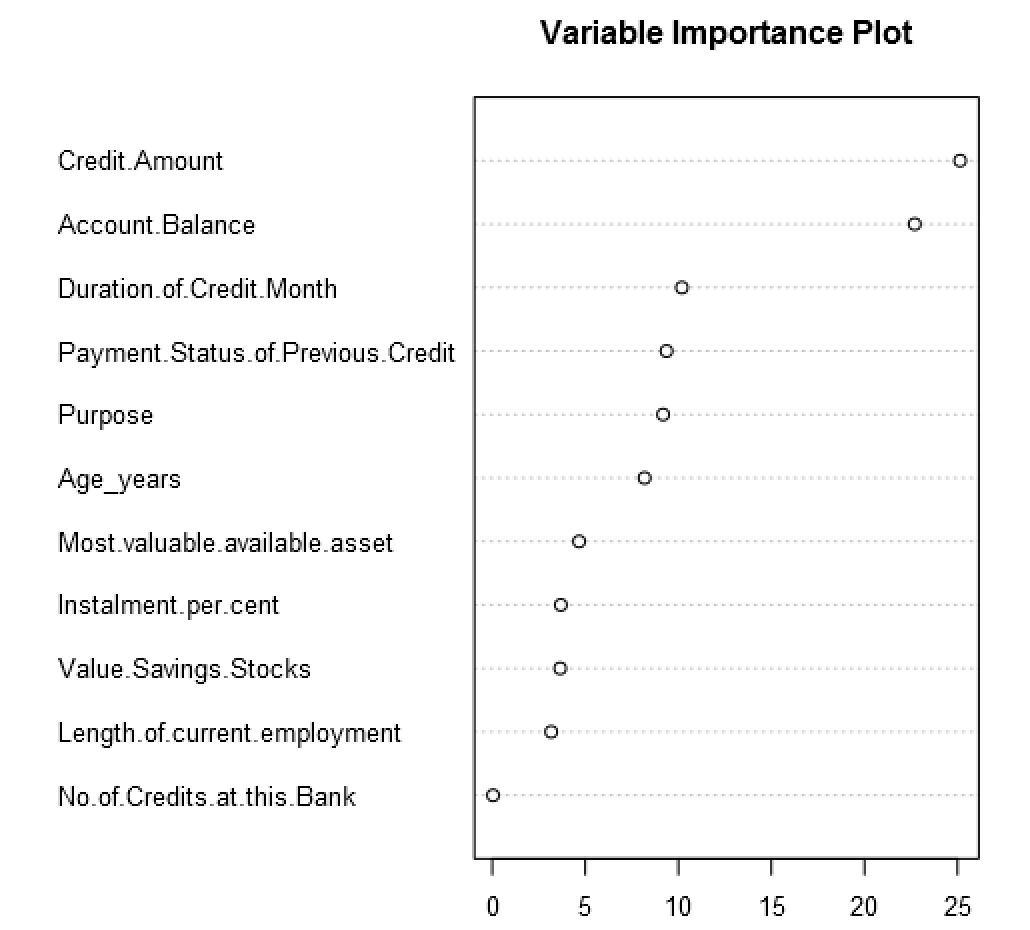
Credit Months

Forest Model Significant Variables –



Credit Amount, Age, Credit Months, Account Balance.

Boosted Model Significant Variables –

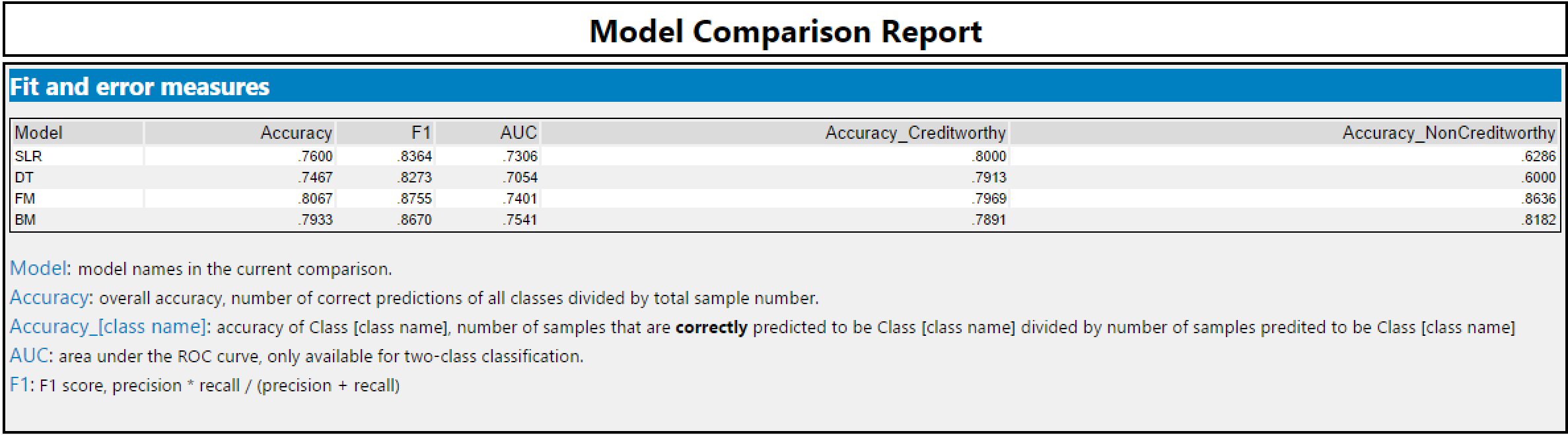


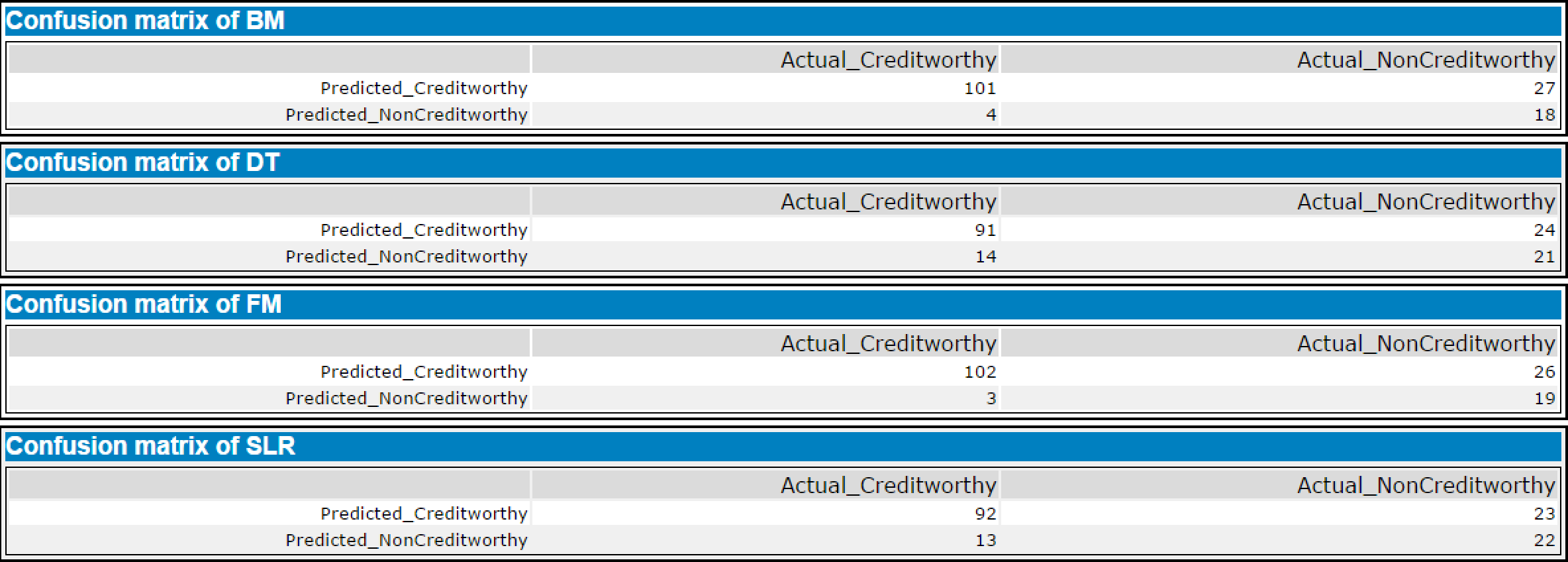
Credit Amount

Account Balance

1. 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?

The below image shows the overall percent accuracy. I see bias in all the models towards the creditworthy variable.





*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:*

1. Which model did you choose to use? Please justify your decision using only the following techniques:
   1. Overall Accuracy against your Validation set
   2. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
   3. ROC graph
   4. Bias in the Confusion Matrices

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

1. How many 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.