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 decision to be made is to process all the 500 loan applicants, whether they are creditworthy or not, according the applicants data.

We can predict the creditworthiness of the applicant data by creating a model using all our past applications data

1. **What data is needed to inform those decisions?**

We would need data related with clients that can :

* + - 1. **Show that client is capable to pay back the loan:**

Example : Account Balance, Payment status of previous credit, length of current employment, most available assets, age , guarantors, value savings stocks, income etc.

* + - 1. **The risk client may have that can cause client may unable to pay the loan:**

For Example: Installment percentage, Duration of credit month, number of dependents, other current credit, job risk level etc.

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

We need a binary model with end result of predicting whether the applicant is creditworthy or not.

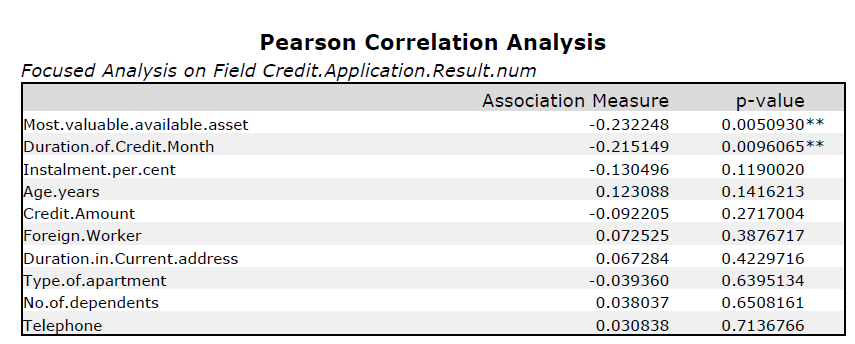
## Step 2: Building the Training Set

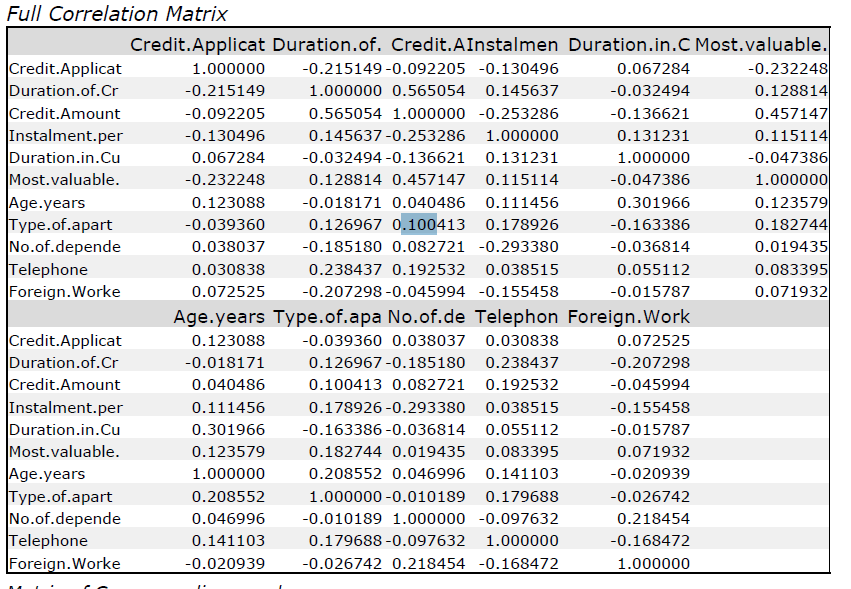
*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn’t* ***need to convert any data fields to the appropriate data types.***

*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”.*

*According to Pearson Correlation Analysis, there is no highly correlation (all less than 0.70)with the Credit result variable and with each other. So we can use all the numerical variable to take it to the next test (missing data & variability test)*

**

**

* *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*

|  |  |
| --- | --- |
| Duration in current address variable has 68,8% missing data out of 500 records, I decided to exclude this variable. |  |
| Age year variable has 2.4% missing data. We will include this variable and conduct imputation for the missing record, using average of age values in the data. |  |

* *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.*

|  |  |
| --- | --- |
| Occupation has uniform value, which mean it has no variability and it will have no impact on prediction analysis. So we will Exclude it | *Occupation:*  *Min : 1*  *Max: 1*  *Med : 1*  *Std.Dev : 0* |
| Concurrent Credit has uniform value, which mean it has no variability and it will have no impact on prediction analysis. So we will Exclude it |  |
| Foreign Worker variable has low variability and very skewed, so we will exclude it |  |
| No od Dependents variable has low variability and very skewed, so we will exclude it |  |
| Guarantors variable has low variability and very skewed, so we will exclude it |  |
| While the data is not skewed, Telephone variable still has low variability, and I don’t think number of telephone owned will relate much with the creditworthiness, so we will exclude the variable . |  |

* *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)*

*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.

The variable that I decided to exclude are:

* + - 1. Duration in Current Address - Too many missing data at 68%
      2. Concurrent Credit - Low variability, the data is entirely uniform
      3. Occupation - Low Variability, the data is entirely uniform
      4. Guarantors - Low variability
      5. Foreign Worker - Low variability
      6. No-of-Dependents - Low variability
      7. Telephone - Low variability

## 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.*

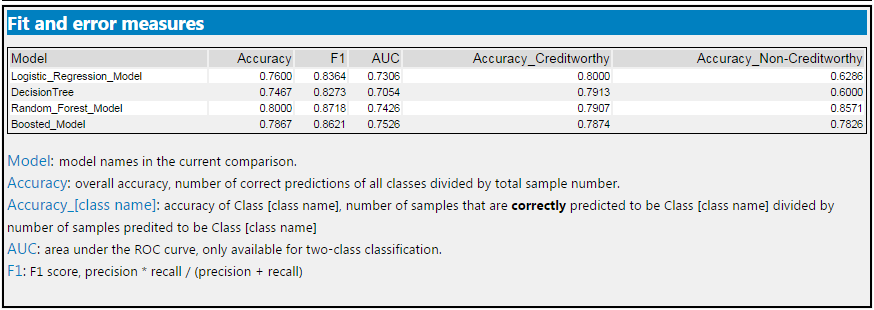
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

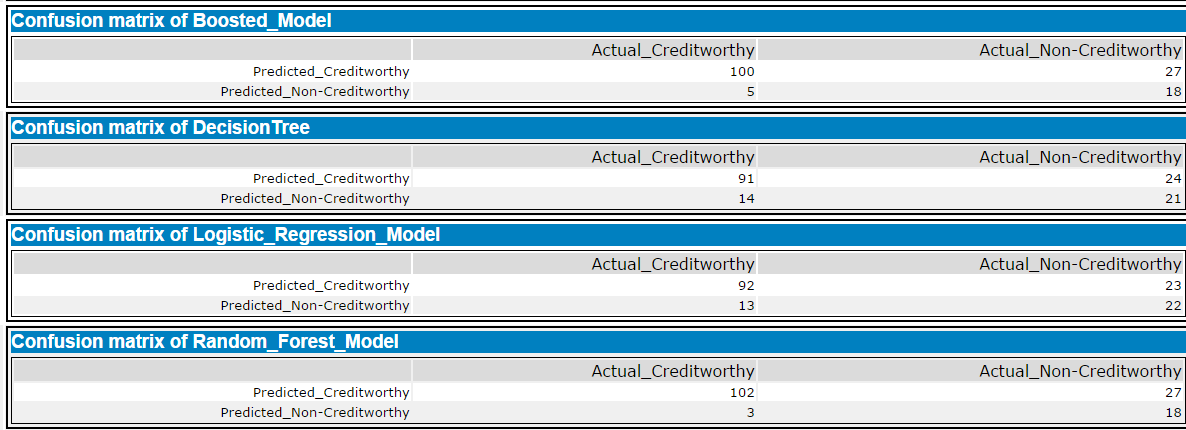
*Answer these questions for* ***each model*** *you created:*

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.
2. 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)*

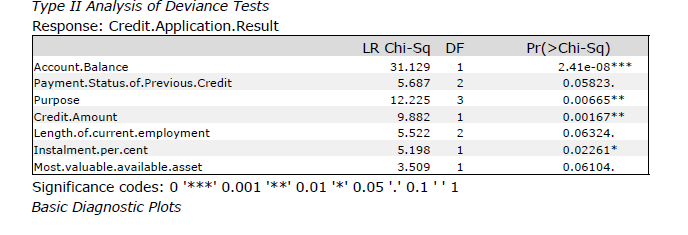
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Forest Model** | **Boosted Model** |
| Important Variable | Account Balance\*\*\* | Account Balance\*\*\* | Credit Amount\*\*\* | Account Balance\*\*\* |
| Purpose \*\* | Duration of Credit Month\*\*\* | Age Year\*\* | Credit Amount\*\* |
| Credit Amount \*\* | Value Savings Stock\*\*\* | Duration of Credit Month\*\* | Duration of Credit\* |
| Instalment percent\* |  | Account Balance\* | Payment Status of Previous Credit\* |
| Payment Status of previous Credit. |  |  | Purpose\* |
| Most valuable available asset |  |  |  |
| Accuracy | 76% | 74.67% | 80% | 78.67% |

******

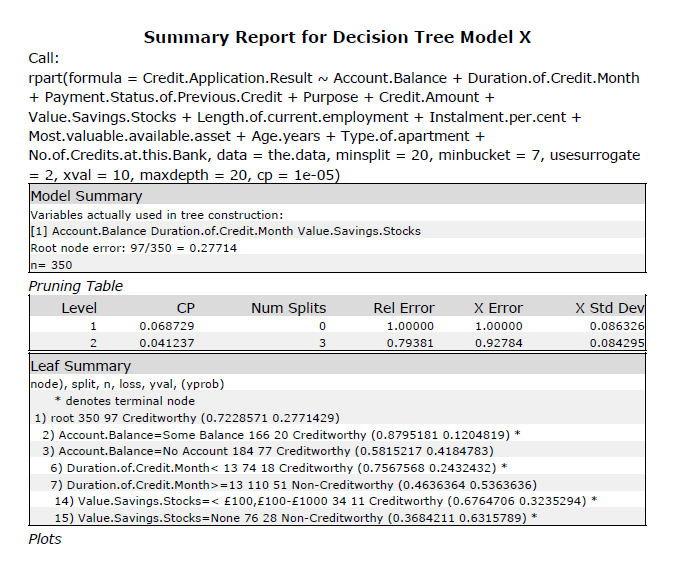
******

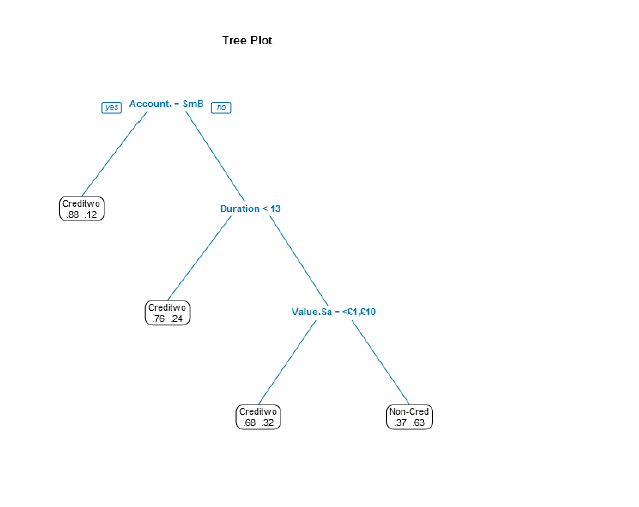
***Based on validation against the current models, the highest accuracy goes to Random forest model. Therefore, we are using Random Forest Model to test the 500 new loan applicants***

***LOGISTIC REGRESSION***

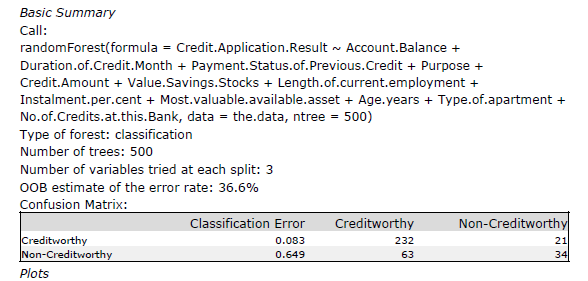
**

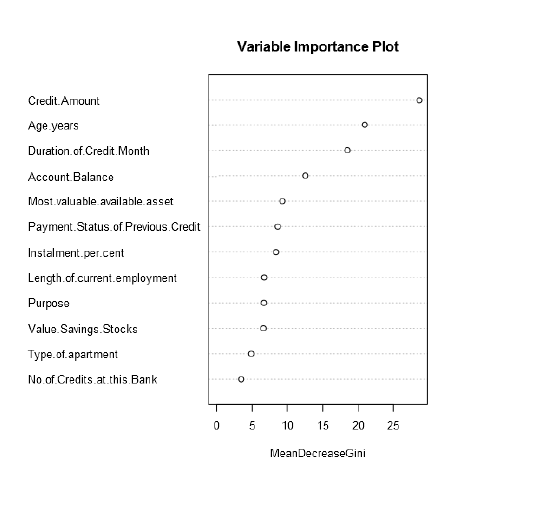
***DECISION TREE***

**

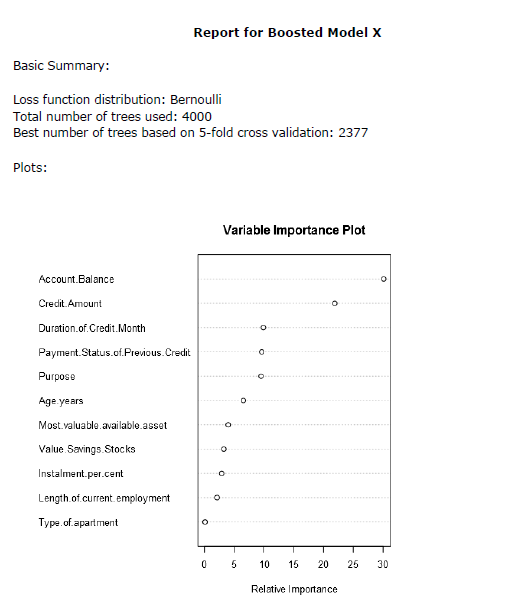
**

***FOREST MODEL***

**

**

***BOOSTED MODEL***

**

## 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”*

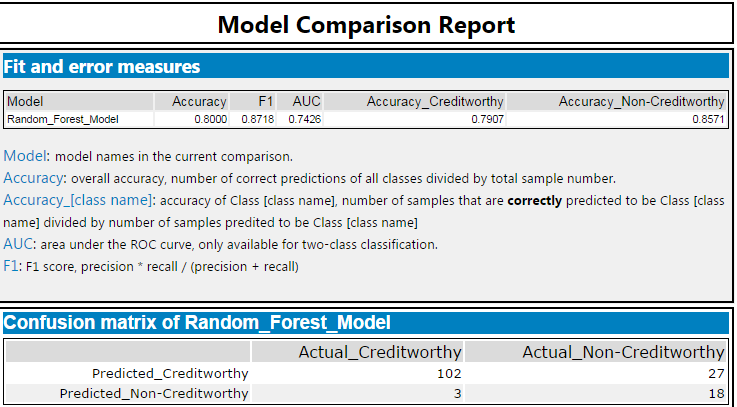
*Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)*

*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?

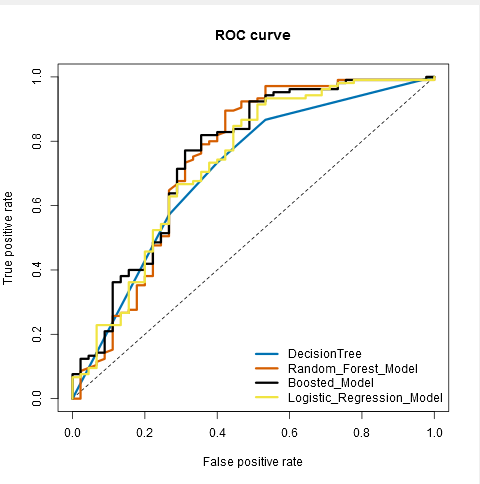


Using random forest model, the accuracy is 80% against validation data set

With 79.07% Accuracy within “Creditworthy”

And 85.71% Accuracy within “Non-Creditworthy”

The ROC graph is as follow:



And using formula , if possibility of creditworthy > possibility non-creditworthy , then it is considered credit worthy, we will get **413** approved clients

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.