**Northeastern University**

**College of Professional Studies**



**ALY-6040 Data Mining Applications**

**TOPIC:** **Data Munging & Data Wrangling**

**TO: Prof. Justin Grosz**

**BY:** Snehal Keshav Bende

Khushboo Bharatkumar Oza

Shruti Sawale

Mohit Patel

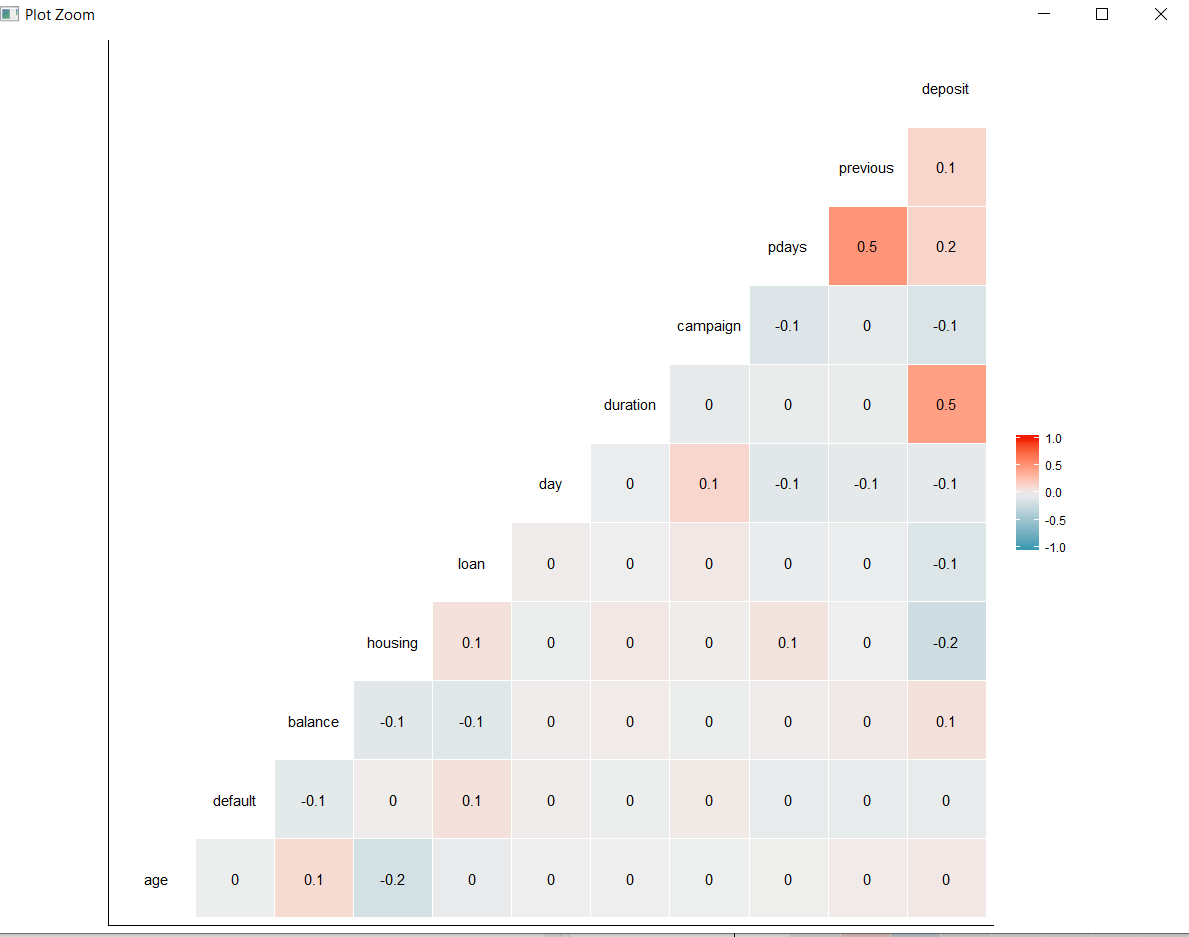
Rishi Barbhaya

**Quarter 2- WINTER 2020**

**Data Munging** is considered as the pre-processing of data where each variable in the data set is converted to the appropriate data type or mapping them to correct format as required for the further processing of the data for the purpose of analysis, in a way this is similar to cleaning and formatting of data. This process works like, Identify🡪 Extract 🡪 Prepare 🡪 Integrate 🡪Consume. In Identify phase the data is collected and required variables from te data are decided. Extract phase refers to regaining of any lost data or hidden content which can be found from any other resource. Next is Prepare step, in which actual cleaning of data is done. Lastly, this cleaned data is used for the purpose of analysis.

Cleaning of data was done in the last week, this week we proceed by changing the data type of variables like Loan, Housing, Default and Deposit to Numeric from character as it is necessary for model creation. Later, the data is divided into the test and train after which models are created and their accurracy can be measured.

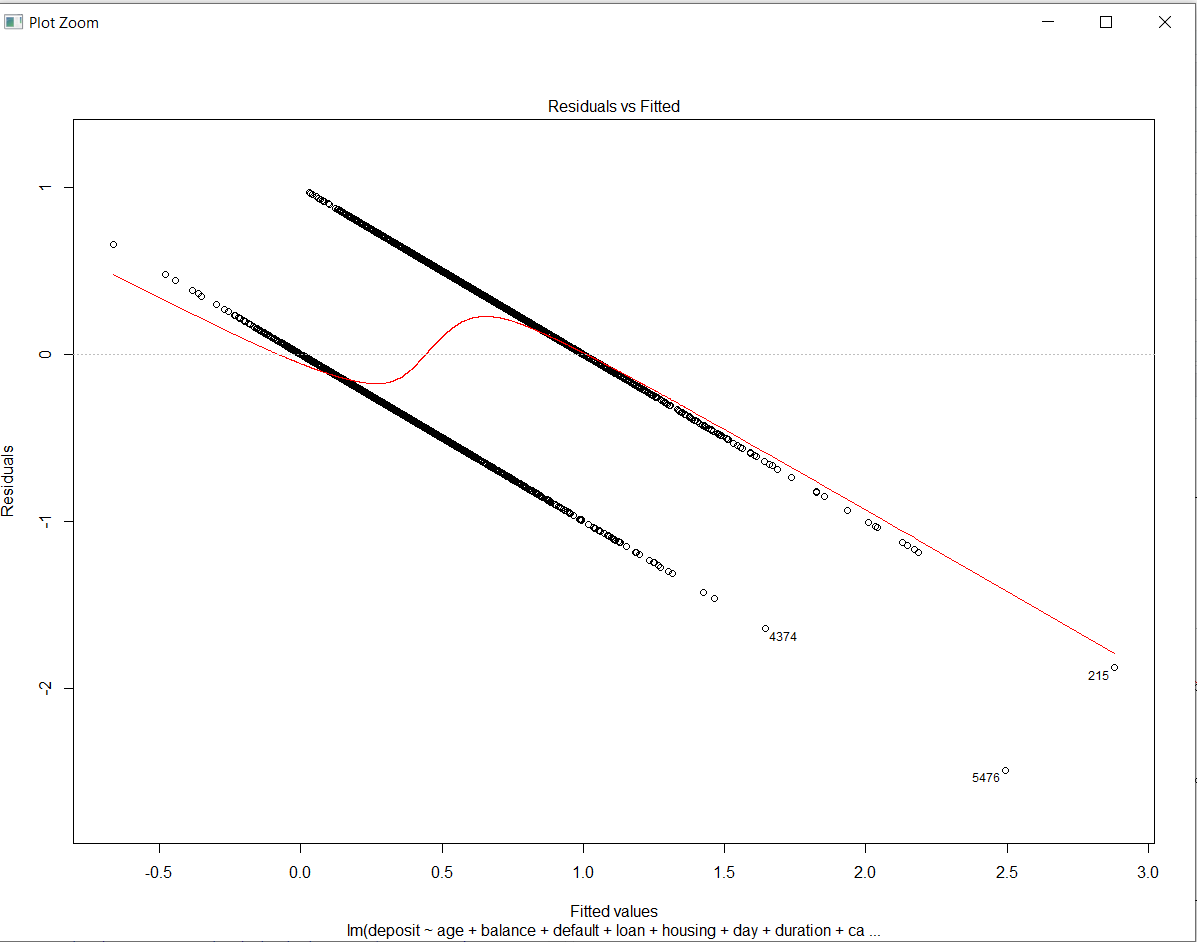
We begin by finding the significance of variables, the chart below demonstrates the same where balance, pdays, duration and previous have considerable positive significance and that of negative is observed in default, loan, housing and campaign.

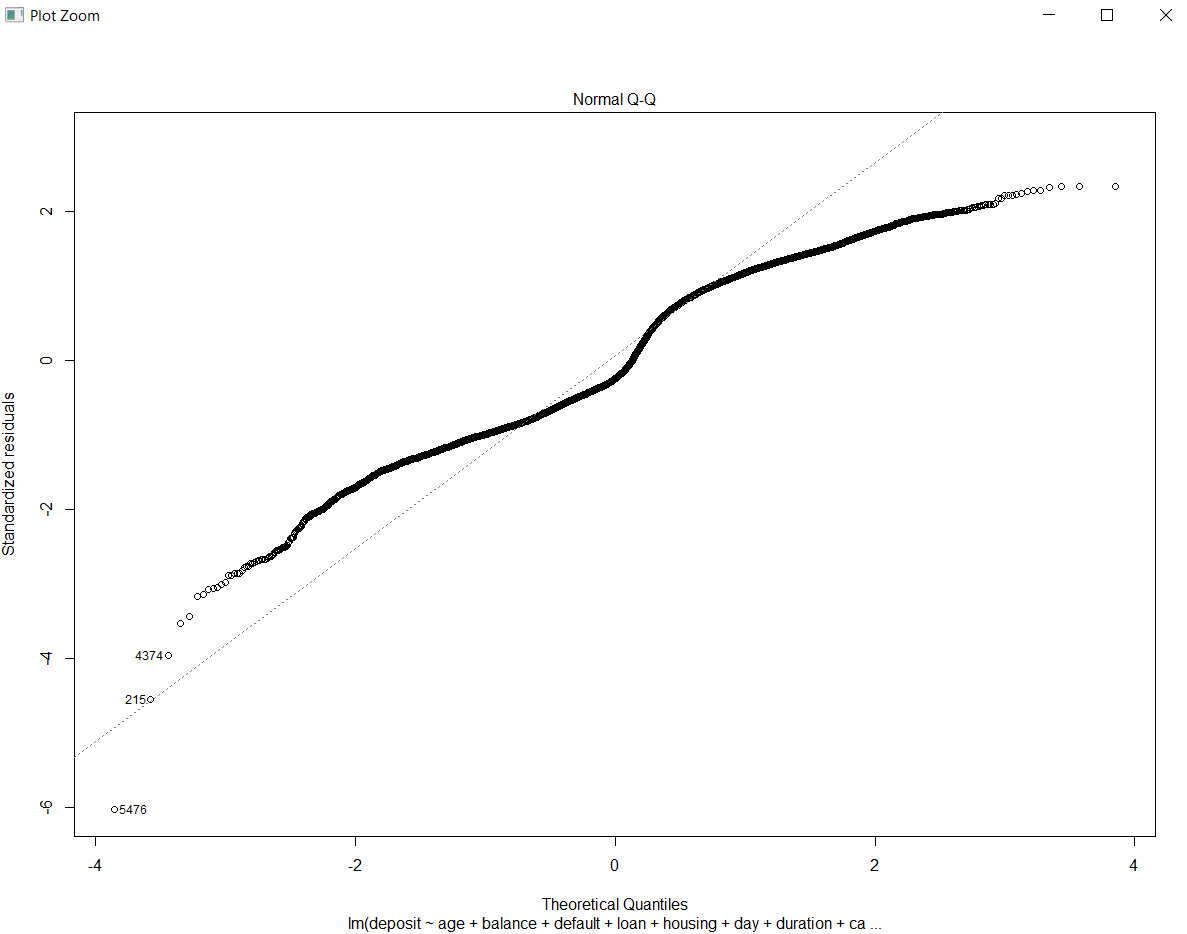


1. **Linear model:**

Linear model demonstrates the relationship between 2 variables where one variable is dependent and the other is explonatory varaiable.

The performance of linear model is not good enough, with the accuracy of 31.2%

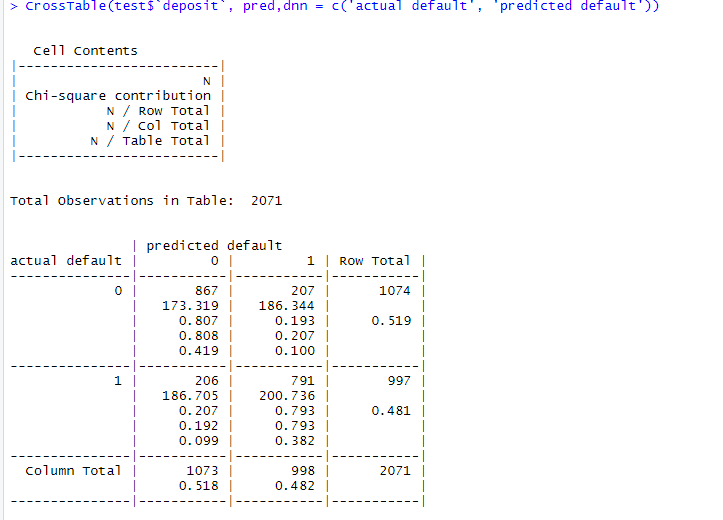




1. **Random Forest:**

Random forest is generally most accurate model and do work efficiently for large data sets which can work on thousands of input variables without any deletion. Random forest do estimate the missing values effectively for and hence generating the more accurate model in such cases. It uses multiple decision trees and each decision tree is applied seperately on different data samples od same dataset with replacement.

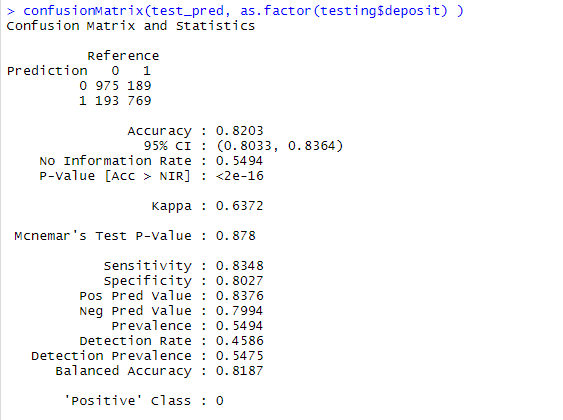
The accuracy of random forest is 80.06%.



1. **SVM Model:**

Support Vector Machine (SVM) is a linear model for regression and outlier detection problems. In SVM, the data is divided into classes based on the algorithm. SVM takes data as input and generates 2 lines as output which separate the 2 different classes. Using SVM algorithm, a support vector is found by getting the line closest to both the classes. The distance between the line and the support vector is known as margin and the main aim is to maximize to margin.

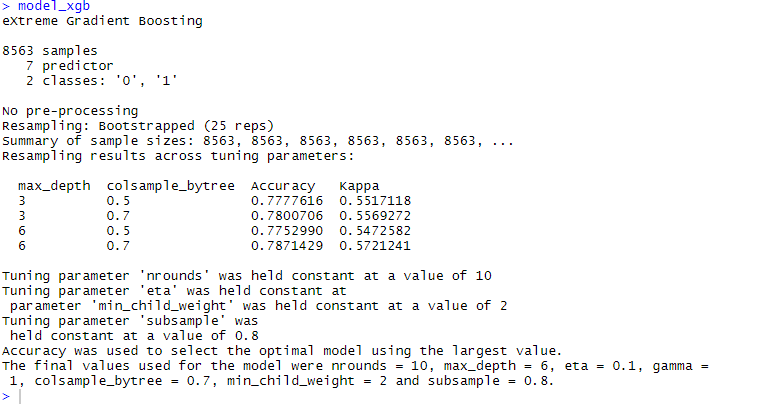
Maximum accuracy is observed in this model with an accuracy of 82%.



1. **Extreme gradient boosting:**

This is termed as XGBoost, XGBoost is a decision-tree-based method which use gradient boosting framework, so more boosting is the iterative approach where more trees are created rather then using a single model. This iterative approach is benificial as the mistakes made in the prior models can be corrected in the new models.

The accuracy obtained by using this method is 78% which is considerable.



The variables used for creating models were selected based on the significance found. Deposit, Age, Balance, default, loan, housing, duration, campaign, pdays and previous are the variables found to significant. Out of the four models generated, SVM turned out to be the one with maximum accuracy followed by random forest and linear regression was the one with minimum accuracy. Deposit is output variable and it is observed that greater the duration of the call, higher the chances of deposit. Greater the age, lesser the chances of an individual having the housing loan. Greater the number of campaigns, greater the chances of deposit. Next, the models need to be optimized. Few questions need to be answered like campaign relation with age, customers account balance influence on campaign, potential candidates for term deposit, right time for marketing activities, etc.

