John Prashanth Gnaniah M (18314752)
Msc Computer Science – Future Networked Systems
Data Analytics Final Semester Assignment – CS7DS1

Objective

The dataset contains a binary response of an experiment conducted. The objective is to find out how well can the response be predicted and to identify which set of independent features are highly predictive of the Response using various predictive analytical methods. R programming Language (R 3.4.0) have been used throughout this data analysis process.

Data Understanding

The given dataset contains 296 observations and 17 variables, and the data dictionary for the same is given in Table 1. Since ID column is just a random number we removed it for further processing. Also, there was a duplicate data point which was removed resulting in a total of 295 final observations. The dataset was not complete, that is there were missing values in the dataset. The percentage of missing values with respect to each features is visualized as a bar plot in Figure 1.

Sl.no.	Variable	Description
1	ID	Unique Id for each data points
2	Response	The binary response of the overall experiment (categorical)
3	Group	The experimentation group (categorical)
4	X1, X2, X3, X4, X5, X6, X7 (**X**)	Actual results of experimentation/tests conducted (Continuous)
5	Y1, Y2, Y3, Y4, Y5, Y6, Y7 (**Y**)	Obtained from respective X's (categorical)

Table 1: Data Dictionary

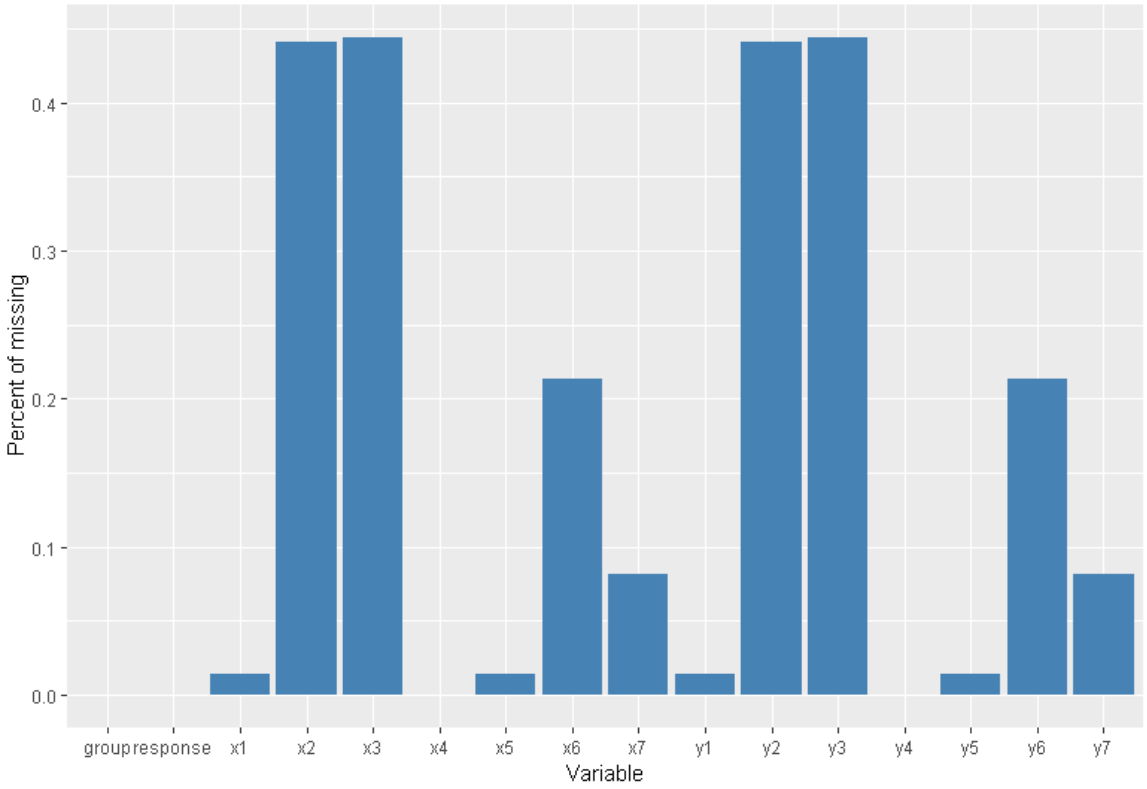

Figure 1: Percentage of Missing Value

Data Description

Univariate and Bivariate graphs were plotted using ggplot2 to understand each variable visually. Figure 2 shows all the correlation of the variables. From the Correlation plot we could understand that some of the X's (eg X1 and X2) are highly correlated with each other. Building a linear regression model to impute the missing values would be a very good option. Frequency bar plot of categorical features and box plot of Continuous variables have been show in Figure 3 and Figure 4. From the graphs we understand that the response variables have been more or less equally distributed which make the sample non biased and accuracy can be a good evaluation metric. The continuous variables (X's) are not normally distributed and there are many outliers (Figure 4) which are not good for fitting linear models. In order to treat the outliers 90% Winsorization/Capping is done where the values beyond 95% and below 5% are replaced with the 95^{th} and 5^{th} quantile respectively. Also square root transformation along with scaling have been performed for data stability of these continuous variables(X's). The box-plots after the treatment is in Figure 5. Bivariate analysis is performed using bar plots in Figures 6,7 to understand the relationship between Response and the independent features. Based on the plots it is conspicuous that Y's and X's are very much related to the response. There is a clear pattern between X's and Response, the distribution ad median of the values of X are different for 0 and 1. Also Y have a higher probability of response 1 when Y is 0 and response 0 when Y is 1. Finally Figure 8 shows the relationship between X's and Y's respectively and we could clearly see that Y's are derived from X's and building a Logistic model would be easy to impute the missing values. In a nutshell, the missing value imputation for X's and Y's could be done using Multivariate Imputation by Chained Equations(MICE). The imputation has to be done separately for X's and Y's. Tree based model and regression models could be used to model the feature space into predictive models and these have methods to interpret the relationship between the response and the inputs variables

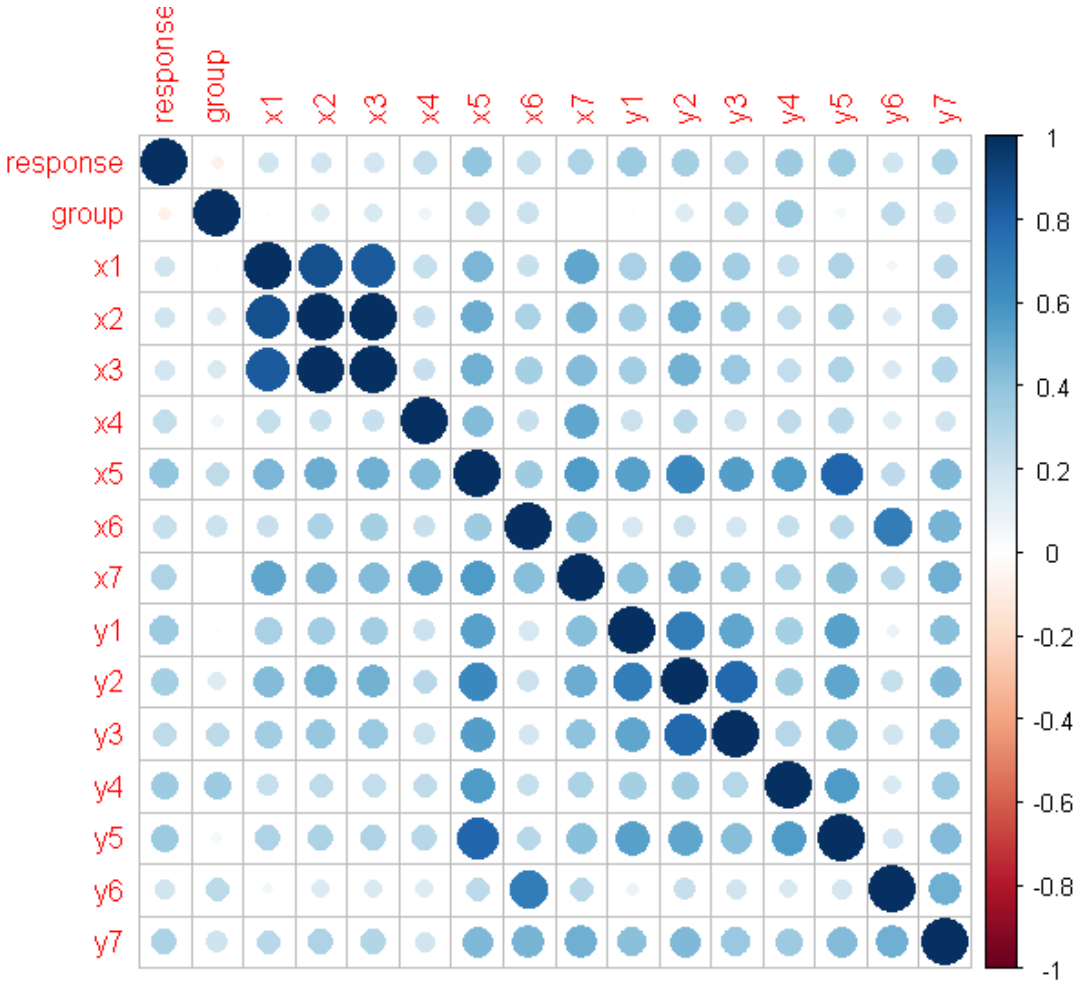

Figure 2: Correlation of the variable in the dataset

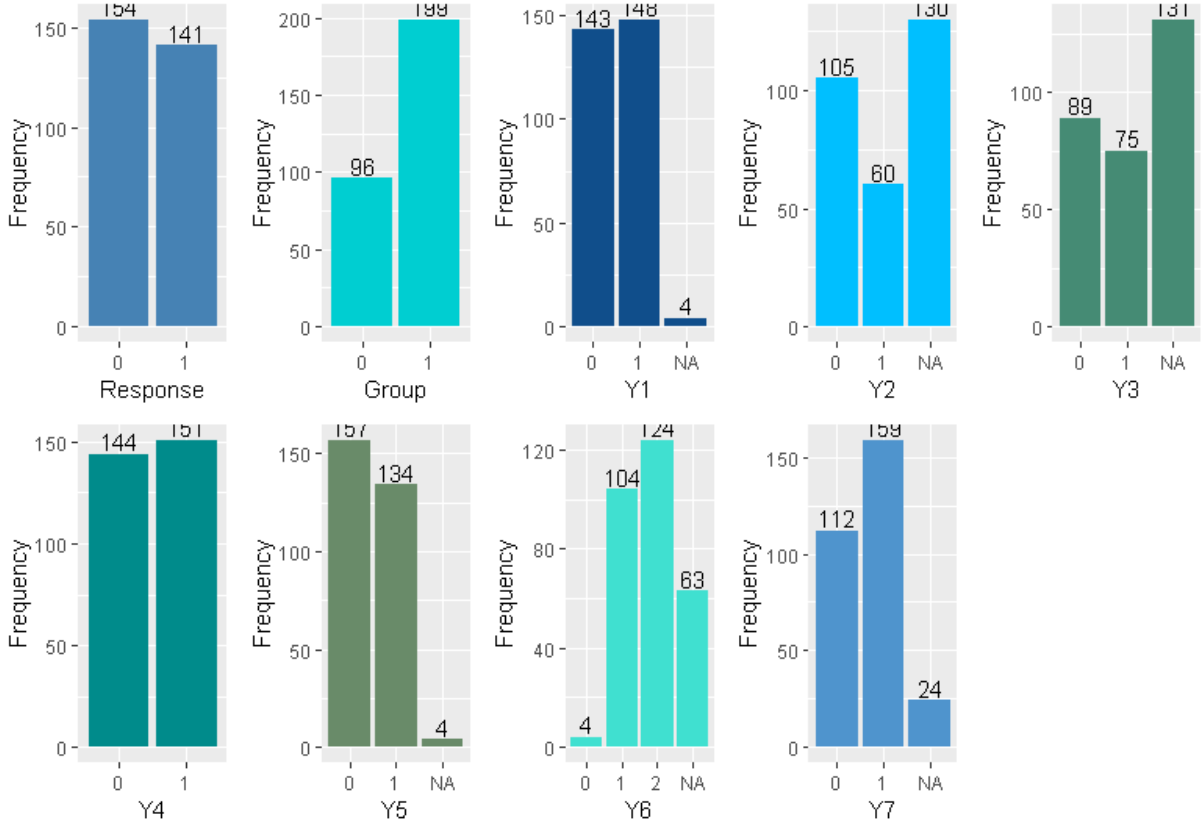

Figure 3: Frequency Bar Plot of Categorical Variables

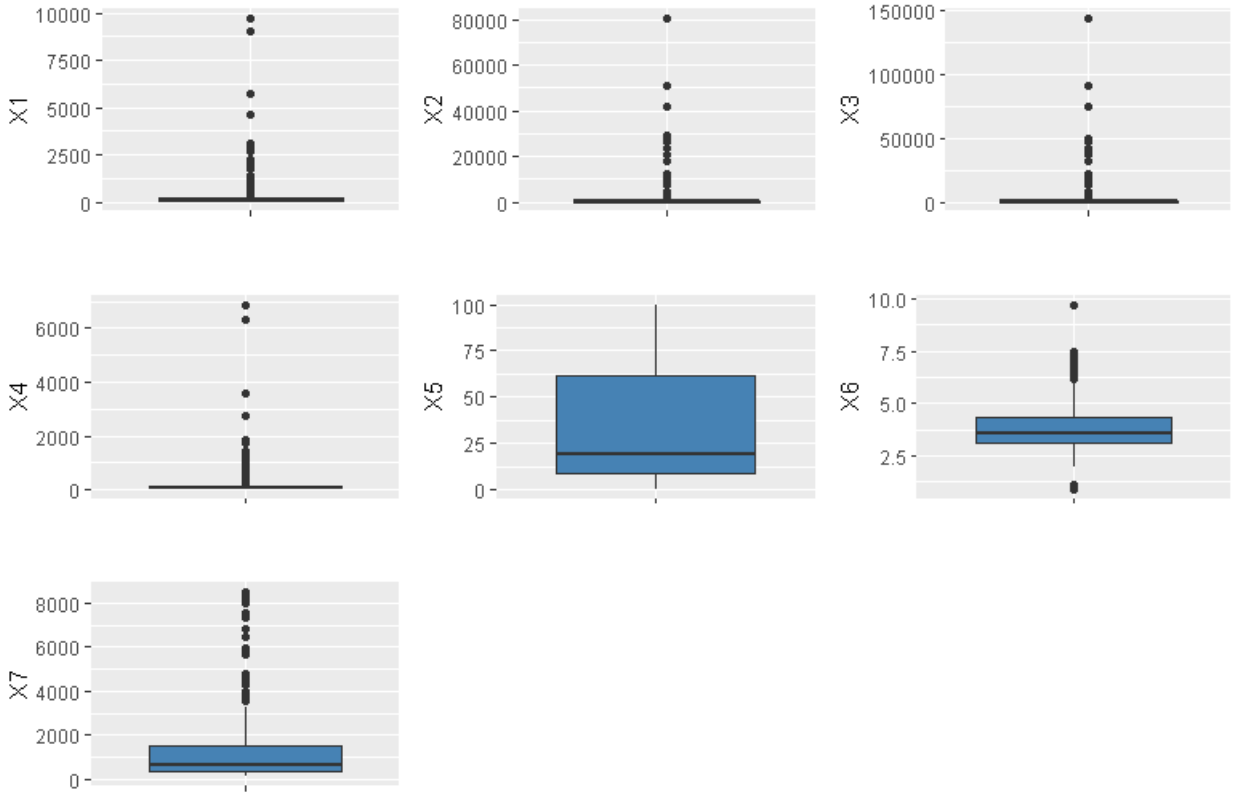

Figure 4: Box Plot of Continuous variables before Outlier treatment

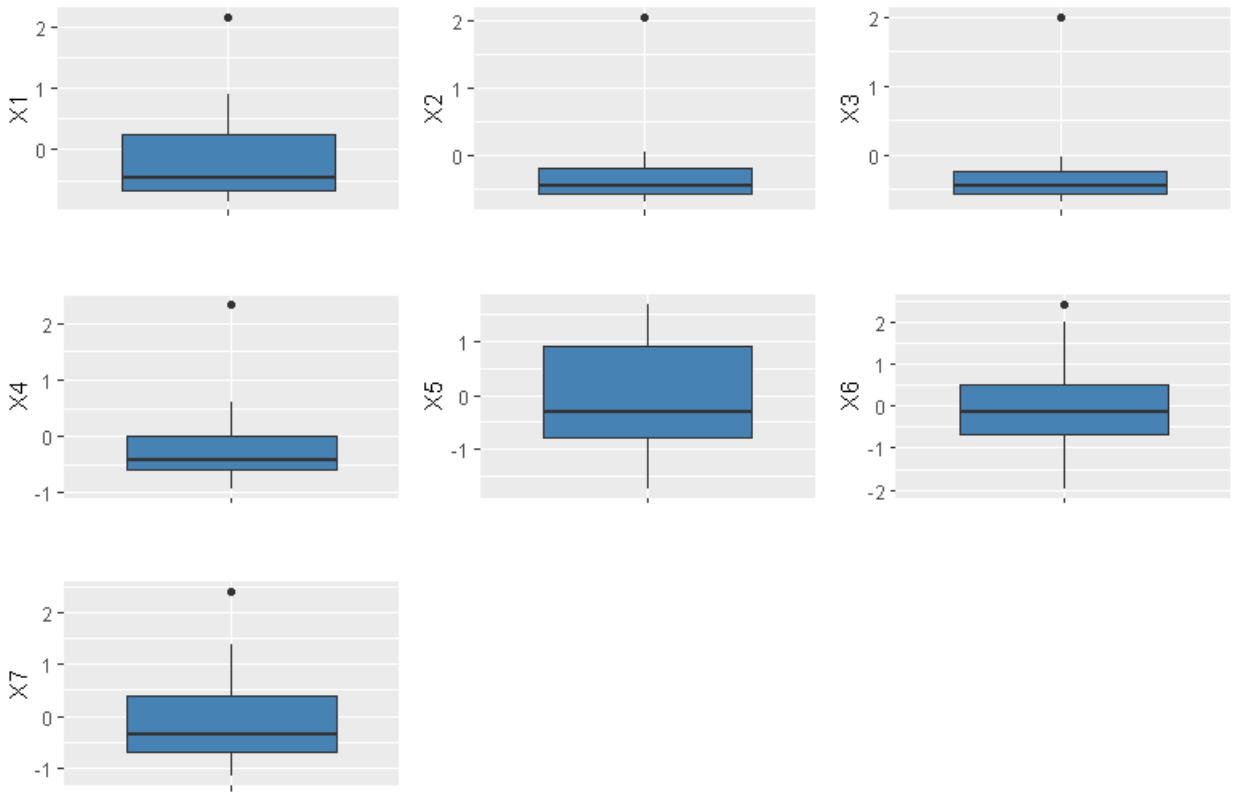

Figure 5: Data Distribution of Continuous variables after Outlier treatment

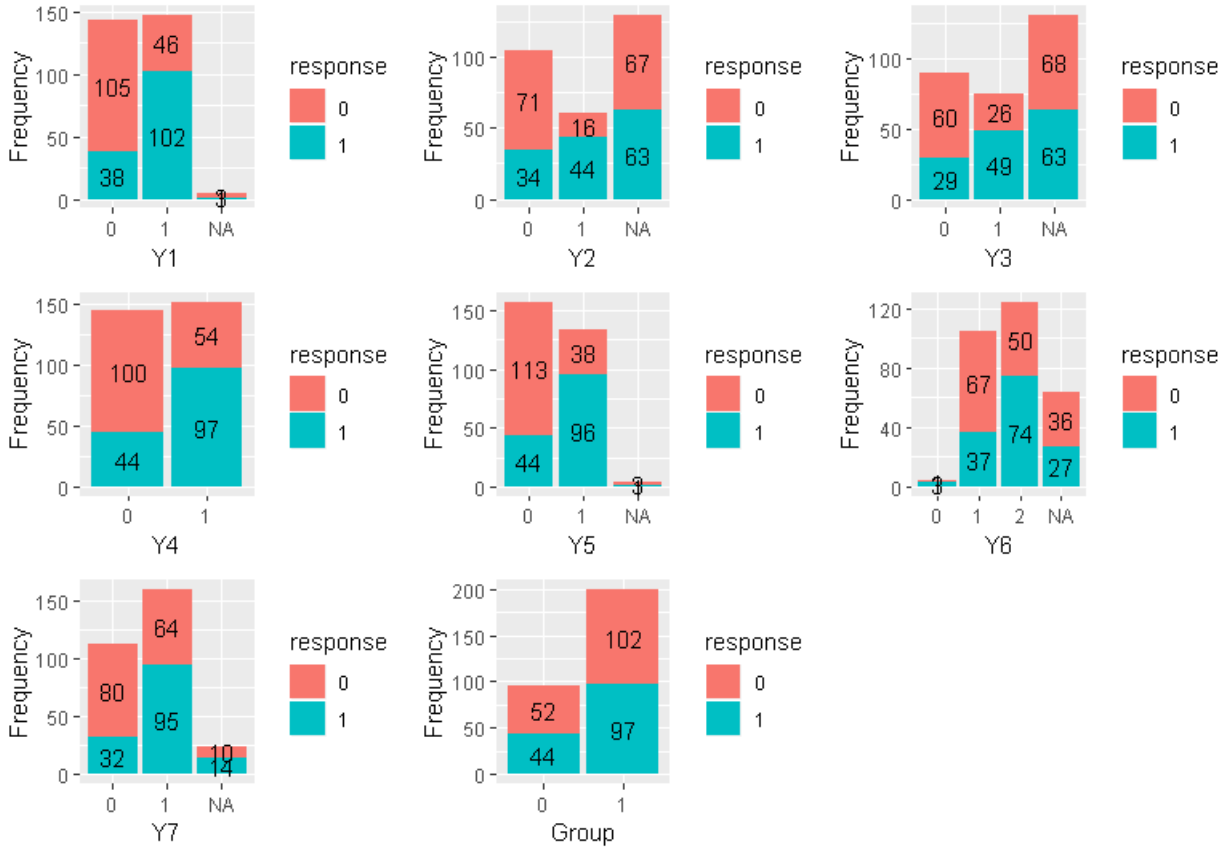

Figure 6: Frequency of Categorical variable with respect to the response

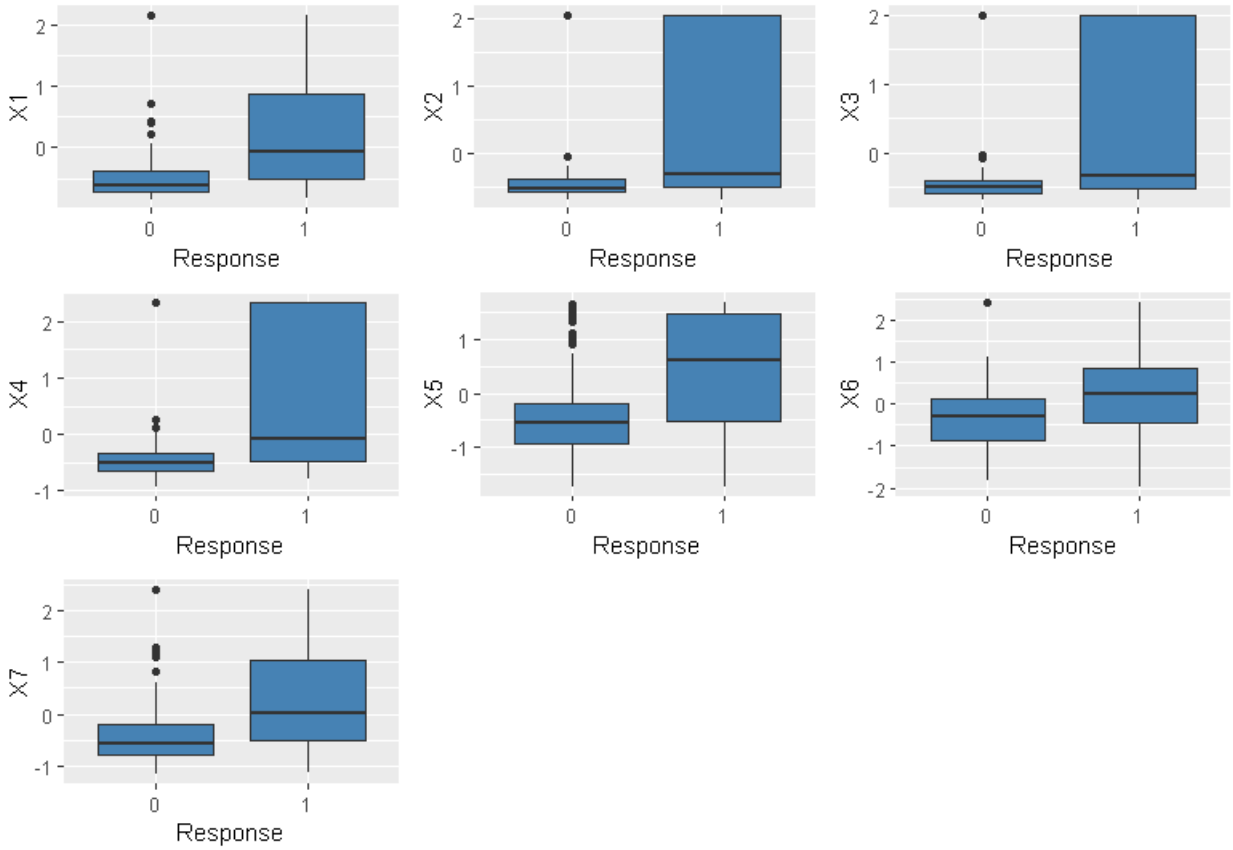

Figure 7: Box Plot of Continuous variable with respect to the response

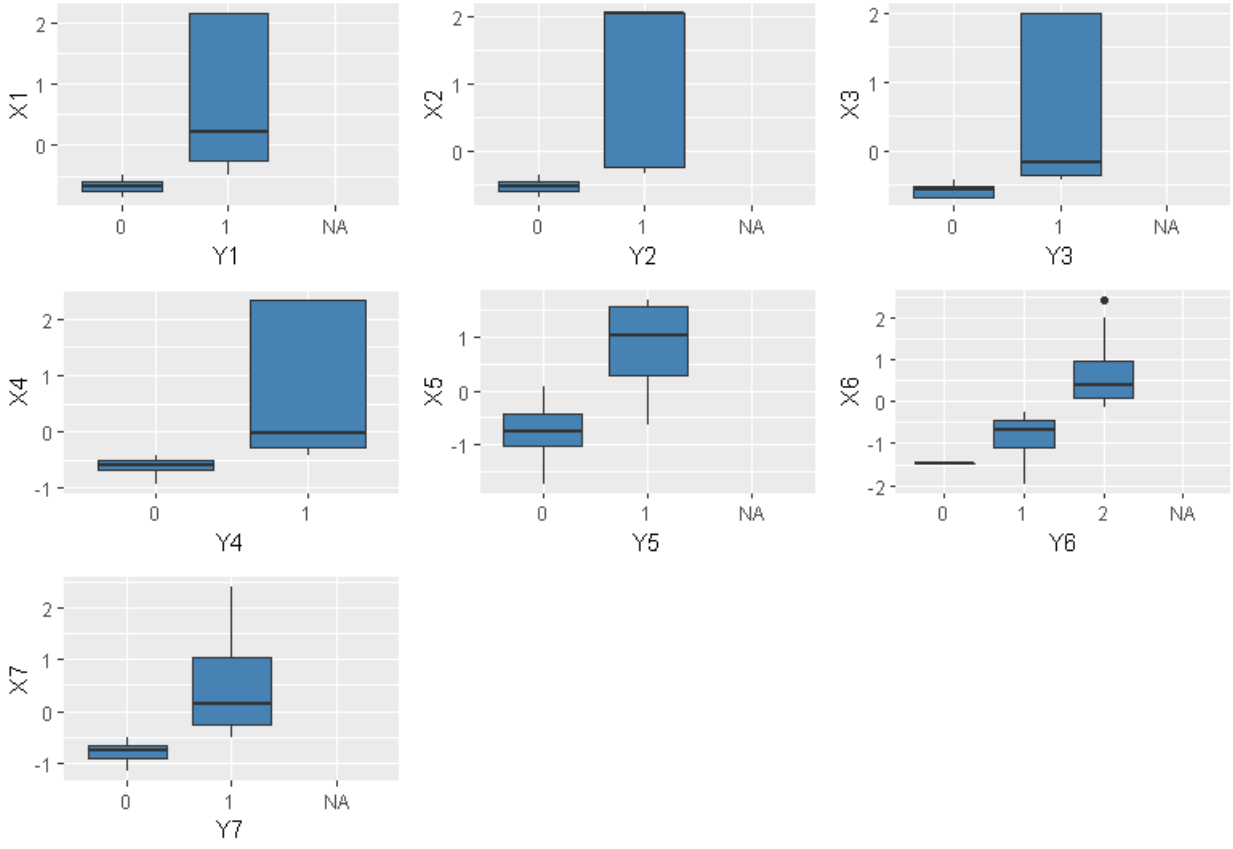

Figure 8: Box Plot of X vs Y

Data Preparation

As part of Data preparation duplicate records were removed. Dummy variables were created for Y6 as it had 3 categories resulting in Y6.1 and Y6.2. Then all the variable has been put into consistent data types, for example the categorical variable has been converted to factors and the continuous variable have been set to numerical data type. As discussed earlier the outlier treatment was done using 90% winsorization technique along with square root transformation and scaling the values between -1 and 1. The datasets were divided into Response,X's, and, Group and Response,Y's, and group for missing value imputation. Missing X's and Y's were separately imputed using MICE package in R using Bayesian linear regression(norm) for X's and Bayesian logistic regression(logreg) for Y's. Once data were imputed the datasets were combined into a single dataset on which learning algorithm models were built.

Model Building

The dataset was split into train and test in the ratio 75:25 ensuring that there is approximately equal distribution of responses. For model building, three algorithms were used Decision Trees, Random Forrest, and Logistic Regression as these algorithms can explain how much the independent features influence the accuracy of the model through variable importance (tree based) and coefficients (Logistic regression). The models were evaluated using Accuracy, Sensitivity, and Specificity evaluation metrics. All the three models were built on the training dataset and then this model is used to predict on the test dataset the Response probabilities. Using the probabilities, the right cut-off point is chosen so that the Accuracy, Sensitivity, and Specificity all remain close to each other. This was done by creating a matrix with cut-off point and their respective Accuracy, Sensitivity, and Specificity values noted. The cut-off was 1000 different points between 0 and 1 equally distributed. From the matrix we obtain the value with the almost near Accuracy, Sensitivity, and Specificity and their respective probability cut-off value. For each of these algorithm three of the below models were built and their respective evaluation metrics were tabulated along with the plot of Feature important(higher the importance higher is the predictive power of the variable)-

1. With X's and Group as independent variable and Response as dependent variable
2. With Y's and Group as independent variable and Response as dependent variable
3. With X's,Y's, and Group as independent variable and Response as dependent variable

Decision Trees

A base tree model with a very low cp of 0.0001 was built so that the tree over fits the dataset. Then we use the pruning technique to remove the unnecessary tree branches that does not yield in good reduction in error. This method is better than truncation technique as this lets the tree grow rather than following a pre-pruning without considering the future step. The cp value for pruning is picked from the decision tree cptable for the lowest xerror(CV error). Then for each model built we identify the test Accuracy, Sensitivity, Specificity, and feature importance which are tabulated in table 2 and the importance variable plot is in Figure 9. Decision trees are not performing very well for this dataset and all the three input sets results in an accuracy of about 70% and we can conclude that X's are good predictors for Decision tree based model.

Sl.no	Independent Variable list	Accuracy	Sensitivity	Specificity	Top 3 Important Variable
1	**X's and Group**	**0.6849**	**0.7143**	**0.6579**	**X1 > X7> X2**
2	Y's and Group	0.7123	0.6000	0.8158	Y2>Y1>Y3
3	X's,Y's and Group	0.6849	0.7143	0.6579	X1>X7>X5

Table 2: Decision tree model results

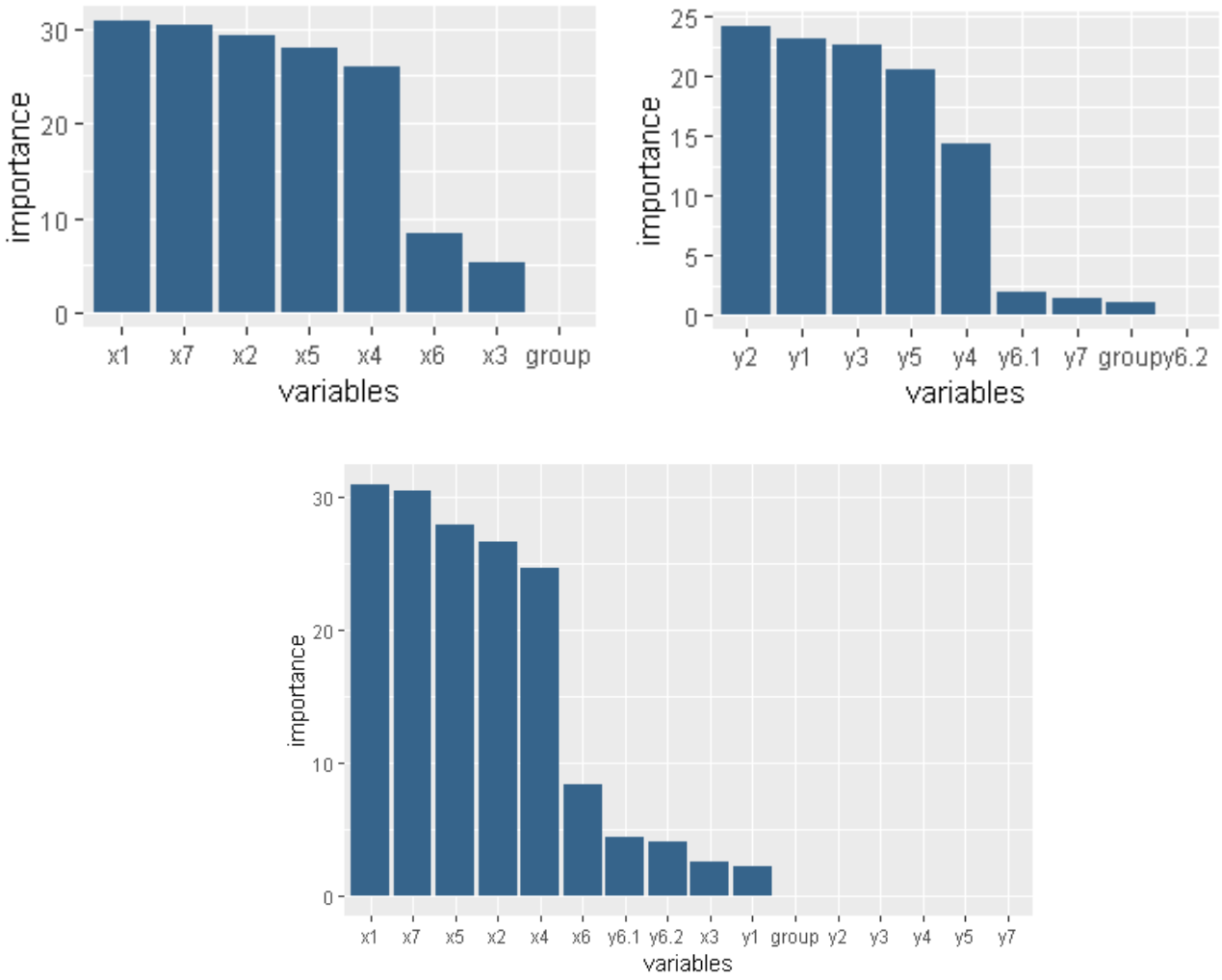

Figure 9: Decision tree Feature Importance for 3 different input sets

Random Forrest

The random Forrest model was built for the different set of input features. The number of ntree(number of trees_ and mtry(number of variable for each tree) was optimized using built-in cross validation tool. The ntree and mtry value which yielded the least out-of-the-bag error was used to build the final model on the training data set.The results and the feature importance are available in Table 3 and Figure 10.It is evident from the table that the X's have yielded the highest accuracy of about 80% for Random Forrest model.

Sl.no	Independent Variable list	Accuracy	Sensitivity	Specificity	Top 3 Variable
1	**X's and Group**	**0.8082**	**0.8286**	**0.7895**	**X4>x5>X1**
2	Y's and Group	0.7534	0.7714	0.7368	Y5>Y1>Y3
3	X's,Y's and Group	0.7534	0.7714	0.7368	X4>X5>X1

Table 3: Random Forrest Model Results

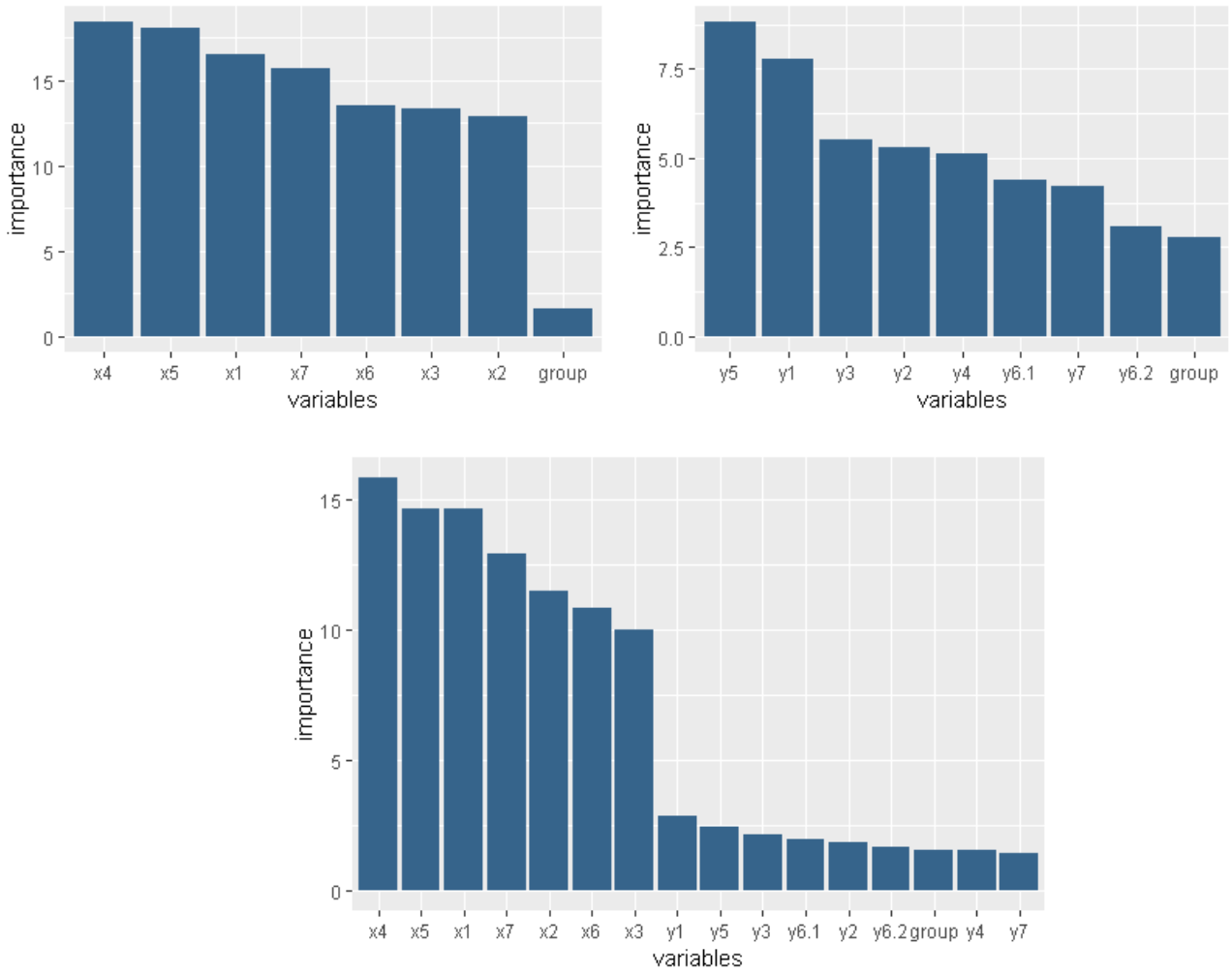

Figure 10: Random Forrest Feature Importance for 3 different input sets

Logistic Regression

A base logistic regression model is built with the different sets of input feature list. Then bi-directional Step AIC algorithm is used to identify the best set of inputs for the model. The STEP AIC algorithms adds and removes variable to find the perfect combination of input features and the best model is chosen based on the AIC value. The final model is used to predict response and the test Accuracy, Sensitivity, Specificity, and regression coefficient (Importance) is available in Table 4 and Figure 11. The Logisitic regression coefficients coefficients could be interpreted as the log-odds of the probability prediction. For logistic regression Y's have yielded the highest accuracy of nearly 80%.

Sl.no	Independent Variable list	Accuracy	Sensitivity	Specificity	Top 3 Variable
1	X's and Group	0.726	0.7429	0.7105	X4>x7
2	**Y's and Group**	**0.8082**	**0.8286**	**0.7895**	**Y4>y3>Y7**
3	X's,Y's and Group	0.7808	0.8000	0.7895	X4>y3>Y2

Table 4: Logistic Regression Model Results

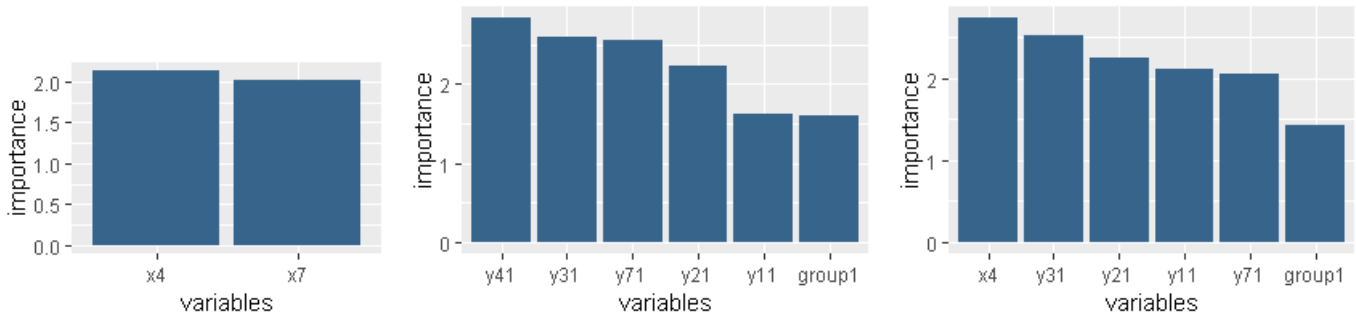

Figure 11: Logistic Regression Feature Importance for 3 different input sets

Conclusion

An accuracy of about 80% is achieved building predictive models. Different algorithms require different set of input features for yielding highest accuracy. The tree based algorithms performed well with the X's as input variable while the Logistic regression model performed the best with Y's as input. Combination of X's and Y's didn't not result in increase in the accuracy but we could see that X's have higher predictive power comparatively. From different models build we could confidently say that X4 is an important predictor for the response.