**MANALI YASHWANT**

**GODSE**

STAT 745 – PREDICTIVE ANALYTICS

Individual project

(Spring 2020)

*Professor – Dr. Peter Bajorski*

**INTRODUCTION**

The basic objective of this project was to build a predictive model based on the training data to predict the response variable (Y) for the test data. It also requires us to guess the MSE value of our predictions based on the MSE of the training data.

Here, we are provided with two datasets to work with, i.e. training data set containing both the response and predictor variables while a test data set just containing the predictor variables. Both the datasets have 5200 observations.

Further, I am using some feature engineering techniques and machine learning algorithms to build the model with a lower MSE to predict the response variable for the test data. The programming language R and its packages in RStudio are used to perform the analysis and implement the ML algorithms.

**METHODOLOGY**

1. **INITIAL EXPLORATORY DATA ANALYSIS –** The very first step in the process of building a prediction model is to understand the structure of the data, look for null values, establish the relationship of predictor variables with that of the response variable and then process the data further as needed.

So, the data was studied by looking for its structure, dimensions, the classes of variables, missing or redundant values, etc. For this purpose, different functions are libraries were used.

Preliminary EDA resulted in the following conclusions about the data:

* Understanding of the basic structure of the dataset.
* No missing or null values are present in the data.
* The data does not include any categorical variables. All the variables are numeric, and the data is continuous in nature.
* Incomplete data or data duplication was not observed. Also, there was no presence of outliers.

In addition to this, EDA was conducted on the training data to establish which predictor variables are significant in affecting the response variable. Histograms for all the variables were plotted to check the normality of the data. The Pearson Correlation Coefficient was computed for each variable to measure the correlation between the response and predictor variables. It was observed here that the variables are positively correlated, correlation ranging between 0.35 to 0.68. Also, the conclusion that all the variables are significant in predicting the response variable was reached.

The original train data was then divided into 80:20 ratio to build a model and test its accuracy by computing the Mean Square Error of the predictions. The model with the lowest MSE is considered as the final model for the predictions.

1. **IMPLEMENTING DIFFERENT ML ALGORITHMS TO BUILD A PREDICTION MODEL**
2. **Linear Models –** Initially implemented the basic linear regression model on data, which gave the MSE of around 326000. Since it is an extremely high number, performed Lasso and Ridge regression on Generalized Linear Models to obtain the best value of lambda for the GLM model. Also, performed a repeated Cross-validation to get the best tuning parameters to the model. Then, the linear model was trained using these new parameters but the MSE reported was still on the higher side.
3. **Feature Selection for Linear Models -** Using ‘*stepAIC’* function in both directions performed feature selection to obtain the best features for the LM model. Extracted variable importance measure from Random Forest to select the top 20 attributes for the linear model. Also, different techniques like Boruta, Relative Importance for Linear Regression, etc. were performed to identify the significant variables. It was observed that most of the features selection efforts yielded similar results. However, feature selection did not help a lot in decreasing the MSE for linear models.
4. **Generalized Boosted Regression Modeling** **(GBM) -** GBM model was implemented with different parameters like the number of trees, the depth of each tree, and the number of observations at each node. Cross-Validation was performed to estimate the generalization error. Although, even after tuning a lot of parameters, this model was not able to achieve a low MSE value.
5. **Tuning Random Forest and Xgboost –** Implementing Random Forest with default parameters gave an extremely high MSE. Therefore, parameter tuning was carried out for various combinations to get better results. The least MSE of *17,769* was observed for the following parameters: ntree = 500, mtry = 100 and node size = 20.

Since, even for this model, the MSE was on the higher side, **Gradient Boosting (xgboost)** technique was implemented as theoretically it is said to give better results compared to Random Forest. However, Xgboost suffered from data overfitting, giving larger MSE values and difficulty in tuning of the parameters.

The linear models failed to improve with all the efforts while Random Forest was able to give MSE as low as 17769 indicating that the relationship between the predictor and the response relationship does not seems to be linear. The next step was analyzing the data and focus on establishing the relationship between the variables.

Thus the next step was to determine Spearman’s correlation as it determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines.

In addition to this, plotted single predictor variables against the response variable *Y* to study the relationship between them. Through these plots, it was established that the variables have a non-linear relationship. Therefore, the following non-linear regression models were implemented:

1. **Generalized Additive Model –** Since GAM does not assume a linear relationship between the independent and dependent variable, tried implementing it by executing various combinations of smoothing parameters on all the variables. The MSE resulted was still high and the training time was also very high for using all the parameters.
2. **Log Transformation –** Since the relationship was non-linear, a logarithm transformation on all the predictor variables was used for training the model which did not perform any better than the simple linear model.
3. **Polynomial Regression –** Polynomial regression using all predictor variables with varying degrees like 2, 3, 4, and 5 were used to train the GLM model. It was observed that the GLM model with the polynomial degree of 5 resulted in the lowest MSE of approximately 5.5.

**CONCLUSION**

It is seen from the above analysis that the GLM model with the polynomial degree of 5 is the best fit for the given train data. Thus, the said model is selected as the Final model for our predictions. Further, cross-validation was done on the model which gave MSE of 5.97.

Finally, this model was trained using the entire ‘Data.train’ dataset given and predictions of the response variable for ‘Data.test’ were done. Since average RMSE for 10 folds of CV in our final model is around 2.70 which results in MSE to be 7.30, we can say that the MSE of our predictions to be in the same range. Thus, my guess for the MSE on my predictions of response variable for ‘Data.test’ is to be approximately near **‘7.305’**.