DATA 602 Final Project Summary :Prediction for Bike Sharing Analysis

-Saivenkata Sravan Anne(FQ45331)

Business problem :

It is critical for bike sharing firms to guarantee that enough bikes are available at stations, but not too many, so that stations do not become overcrowded.Avoiding excess and bike shortages results in happier consumers, thus forecasting future demand is critical.

About the project :

The main aim or the goal of this project was to use data from Capital Bikeshare to forecast bike sharing demand in the Washington, DC metro region. I intended to contribute to the topic of demand prediction by comparing tree-based ensemble algorithms AdaBoost, random forests, and XGBoost and looking for ways to eliminate look-ahead bias. As a result, I forecasted the number of bike sharing clients for the following day around the city. I used data from the Capital Bikeshare website to get the bikes dataset, historical weather data from National Climatic Data Center, and information from the DC Department of Human Resources on which days are holidays.

Process:

The initial step is to merge the datasets which have been downloaded from their respective source. After combining the datasets the features that remained are the date feature, categorical feature (19 of them which involves columns related to the holidays dataset as well as the weather dataset which has been obtained from National Climatic Data Center website) and few numerical features as well as the main variable which is the target variable which is the total number of customers. Once this is done the Exploratory Data Analysis has been performed where I had plotted various box plots for various categorical variables which were initially part of the weather dataset but now it has been

obtained from the combined dataset.Plotting such distribution has shown that the number of customers who has used the service is also dependent on the time of the year. From the observations of the various plots it has also been found that the number of customers who will be using the bike service is very much dependent on the weather type as well. During this process the encoding of the temperature ranges has been done to the weather types which are categorical variables having different categories such as very cold,cold,warm and hot based on the temperature ranges. Based on the category of a weekday or a weekend there has not been much impact in the total number of customers who used the service during the weekday.So after the EDA has been performed, about 25 features have been used to find out the demand .

Data Modeling:

Tree based machine learning models have been used for prediction rather than the other models as these models have the ability to handle mixed data types and they are also easy to understand.As a result, I have utilized AdaBoost, Random forests, and XGBoost to forecast bike sharing demand and figure out which of these three models is best for this type of data. Last value technique has been used to predict the demand for the vehicles which are part of the Capital bikes firm.For finding which one is a better model there are two different metrics which have been considered which are the Mean average error(MAE) as well as Root mean squared log error (RMSLE).The MAE represents the prediction's average deviation from the true value.The penalty for underestimating the Actual variable in RMSLE is greater than the penalty for overestimating it.Thus these metrics are used to find out which of the model is better.As a reference model the Naives Bayes Baseline model has been also executed as it it can act as a benchmark to the other models.After executing the various models and comparing the machine learning models with the benchmark it has been found out that the XGBoost Model has performed well as compared to the other models.Because of the additional regularization hyperparameter that prevents overfitting, XGBoost is expected to perform better. While employing

bagging/bootstrapping the training sample reduces overfitting, this does not help with time series data when we wish to avoid unpredictability.

Conclusion:

Out of the three tree-based ensemble algorithms, XGBoost performs the best and is more resistant to overfitting and noise. It also allows us to ignore the fact that this data collection is stationary. However, the outcomes are still unsatisfactory. More work should be spent on feature selection and engineering in order to improve forecasts of bike sharing demand. Because the model is underfitting, new and/or additional features will almost certainly be required.