**Predicting Rental Bike Counts to Improve Environmental and Social Welfare**

**1. Introduction**

The city of Ourra has developed a bike sharing system called Drpia in order to improve environmental and social wellbeing. The secretary of transportation is attempting to understand the drivers of rental bike demand in order to effectively balance the available supply at any given time. This project is to predict the rental bike counts in demand to improve environmental and social welfare.

**2. Exploratory Data Analysis**

Exploratory data analysis (EDA) is the process of studying and visualizing data in order to gain a better knowledge of it and gain insight. In this project, we used some tools for EDA including importing data, descriptive statistics, cleaning data, preprocessing data and visualizing data. In this project, the train dataset comprises 15 variables, 5 of which are character variables and 10 of which are numerical variables. For this dataset, we used descriptive statistics to get the minimum, median, mean, standard deviation, and maximum values for each numeric variable. The correlation matrix allows us to determine whether or not there are strong relationships between the variables. Temperature is substantially positively connected with Dew point temperature, with a value of 0.92, according to the generated correlation matrix for our dataset. Furthermore, Humidity is inversely connected with Visibility, with a correlation value of -0.56. The correlation plot depicts the same conclusion as the correlation matrix.

Chart, bubble chart

Description automatically generated

In the data cleaning section, we mostly check for missing, null, NA, NaN, or empty entries in our dataset. Missing values are often handled in two ways: removal and imputation. There are no missing or null values in our dataset. The duplicate rows add nothing to the model's or algorithm's learning process, but they do add storage and processing cost. We tested our dataset using the ID column and discovered that there are no duplicate rows. In some circumstances, models, such as linear regression and support vector machine, may be particularly sensitive to outliers, and outliers may impair model performance. A decision tree, on the other hand, would be unaffected because it deals with each observation individually. In our project, we utilized boxplots to see if there were any outliers in the columns, and we discovered some in the Wind, Solar, Rainfall, Snowfall, and Count variables.

Data visualization is extremely useful for observing data distributions and trends. To analyze the distribution of a classed variable, a bar chart can be utilized. We generated three bar plots to show the number of rental bikes in different seasons, whether they were holiday or non-holiday, and whether they were operating or not. They claim that the number of rental bikes in the summer is higher than in other seasons. The number of rental bikes was substantially higher on non-holiday days than on holidays. The demand for operable rental bikes in functional hours was far greater than that for non-functioning hours. Categorical data is classified into groups, classes, and text. Seasons, Holidays, and Functioning are categorical variables in our dataset. One hot encoding is a common approach where we can create new columns, one for each unique classes in categories. As a result, we converted the category variables into numerical values. Next, we extracted the month and day from the Date column and created three bar plots across the Month, Day, and Hour columns, revealing that the rental bike counts in June were significantly higher than in other months, that the rental bike counts on Saturday were significantly higher than those on other weekdays, and that the rental bike counts at 18:00 were significantly higher than those at other hours. Finally, we removed the ID column as well as the original Seasons, Holiday, Functioning, and Date columns. After cleaning and preprocessing, our new dataset has 15 columns.

Chart, pie chart

Description automatically generatedChart

Description automatically generated

**3. Methods Overview**

Before fitting a simple model to each segment, basic regression trees split a data set into smaller units. Regression trees offer several advantages, including the fact that they are simple to read. Predictions are created fast and without the need of sophisticated computations. It's easy to determine which aspects are important in forecasting. The mean squared error is the model selection criteria in our research (MSE). Furthermore, these trees may be trained utilizing efficient, trustworthy methods. However, there are several severe flaws: single regression tree variance is high, leading in unstable results. Because of their great variety, single regression trees have low prediction accuracy. We investigate regression tree, bagging, random forest, and generalized boosted regression models in our project.

Bagging regression trees have the potential to be extremely powerful and effective. Furthermore, this lays the groundwork for more complex tree-based models like random forests and gradient boosting machines. The practice of mixing and averaging several models is known as bagging. Averaging over numerous trees lowers overfitting and limits the variability of any one tree, both of which improve forecast accuracy. Bagging is accomplished in three simple steps: Generate m bootstrap samples using the training data; bootstrapped samples allow us to generate a large number of different datasets that all have the same distribution as the entire training set; train a single, unpruned regression tree for each bootstrap sample; individual forecasts from each tree are averaged to provide an overall average expected value.

Random Forests is a flexible machine learning approach that can handle both regression and classification tasks. Random Forests outperforms bagged trees due to a little adjustment that decorrelates the trees. When creating these decision trees, each time a split in a tree is examined, a random sample of m predictors from the whole set of p predictors is picked as split candidates. Only one of the m predictors may be used in the split. When a high proportion of the data is missing, Random Forests is particularly successful in estimating missing data and maintaining accuracy. It can also balance errors in datasets with imbalanced classes. Most importantly, it is capable of handling big datasets with high dimensionality. However, one problem of employing Random Forests is that they are prone to overfitting noisy datasets, especially when used for regression.

Boosting is another method for improving the predictions produced by a decision tree. It is a generalized approach that, like bagging and random forests, may be used to various statistical learning methods for regression or classification. Boosting works in a similar way compared to bagging, except that the trees are produced sequentially: each tree uses information from previously built trees. Instead of using bootstrap sampling, each tree is fitted on a modified version of the original dataset. Boosting is quite useful when you have a large amount of data and expect the decision trees to be extremely complicated.

**4. Summary of Results**

For the regression tree method, we used cross-validation to select a good pruning of the tree. It shows that the tree size of 10 could be the most appropriate number for the tree construction in our case. The MSE is 125440.5 as the same as the value without pruning. Next, we applied Random Forest model to predict the rental bike counts and calculated the MSE is 36000.21 after tuning the hyperparameter (number of variable randomly sampled) in the model. The model with 11 random variables samples at each split has the lowest MSE. Calculating the variable importance after tuning the model, we noticed that variables Hour, Temperature, Humidity, Wind, Visibility play vital role in predictions of rental bike counts.

In the Bagging method, we calculated the MSE with value 104743.8 after tuning the hyperparameter that is the number of bootstrap replications. The bagging model with 11 bootstrap replications has the lowest MSE. Meanwhile, the most important variables in the tunned model are Hour, Humidity, Temperature, Solar, Rainfall. Finally, we came to the generalized boosting regression model and found out the most important variables are Temperature, Hour and Humidity in the model before tuning the hyperparameters.

Next, we tuned the hyperparameter the total number of trees to fit and it took long time when we set the large number of tree size, for example, 10,000. So, we decided to set the number of trees to fit from 100 to 5000. The tuned boosting model gives us the lowest MSE with value 34462.01 and the most important variables Temperature, Hour and Humidity.

Table

Description automatically generated

**5. Conclusion**

Based on the results above, we found out the generalized boosting regression model has the lowest MSE with 5000 tree size. If we increase the number of total trees to fit, the MSE would decrease slightly in the boosting model. However, the computation cost is high because the trees become much more complicated. Looking at the table above, it is easy to notice that variables Temperature, Hour and Humidity were selected in all estimated models due to the importance. In other words. those variables as the most important consistent were across the performing models. Compared the generalized boosting regression model, the MSE calculated from estimated Random Forest model is slightly higher than that of boosting model. But the computation time is faster than that in boosting model.

The project is to predict the rental bike counts to improve the environmental and social welfare. According to the model performance and predictive analysis, we strongly recommend that the rental bikes can be distributed more in Summer, especially in June and 18:00 pm can be the high volume of a day in rental bike counts.