**SUMMARY**

1. **EXPLORATORY DATA ANALYSIS**

Dataset “Daily.csv” is selected for performing the analysis as it has data ordered by date. Dataset consists of 16 columns. Columns like “instant” and “dteday” indicates serial number and date. “Casual”, “Registered” and “cnt” represent the total number of bike shares on a given day. Every other column can be called as independent variables.

1. **FEATURE SELECTION**

There is an assumption that “Casual” and “Registered” can have different influencing factors. So an approach called “Backward elimination” is used to extract features. At first all the columns are selected for the model. One by one features are removed considering the p values of the columns from the current model.

1. **MODEL FINALIZATION**

At the end for “Casual”, the model with these features are finalized “season + yr + mnth + holiday + weekday + workingday + weathersit + windspeed”.

Similarly for “Registered ”, the model with these features “season + yr + weekday + workingday + weathersit + atemp + hum + windspeed” is finalized.

1. **INDEPENDENTLY IMPLEMENTING MODEL FOR CASUAL AND REGISTERED**

Now that there are two models ready to perform analysis. The models are used to predict the respective dependent variables. To obtain the final “Cnt”, predictions from both the models are added together at the end.

A dataset is splitted into two parts. Training and testing, where training consists of 75% data and testing consists of 25% of the total data. Once splitting is completed. Two models are created using the extracted features. First model predict for the “Casual” following the prediction by “Registered’s” model. The predictions are added at the end. As “Cnt” is the summition of the two columns. The final values of the predicted “Cnt” are ready to be compared against the actual “Cnt” values.

1. **REGRESSION METHODOLOGY**

Linear Regression model is used for the first instance and Random Forest Regressor is used for the second instance for prediction.

1. **CONCLUSION AND RESULT EXPLANATION**

The “Cnt” consists of continuous values. The predicted values are categorized based on the percentage of errors. The result shows the percentage of data points that has less than 10% of error. Similarly the percentage of the data point having error between 10% - 20% is shown. Which helps to understand the accuracy of each method used. There are other methodologies to predict the values. Here Random Forest Regressor and Linear Regression is used. On a comparison note. Both produces almost similar results. Although Random Forest Regressor was better on maximum number of the times. The final result is the average taken after perfroming the regression for 15 times. The Random Forest Regressor has overall less percentage of datapoints with errors below 20% of acutal values.

|  |  |
| --- | --- |
| Actual vs predicted graph for Random Forest Regression | Actual vs predicted graph for Linear Regression |
|  |  |