Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name: Soma Pavan Kumar
Email: spkumar1998@gmail.com
Contribution:
 	
1. Data Importing and Cleaning
2. Data Visualizations and getting insights
3. Getting the Statistics of Data
4. Data Modeling with various machine Learning algorithms
    1. Multiple Linear Regression
    2. Decision tree Regressor
    3. Random Forest Regressor
    4. XGBoost Regressor.
5.Conclusion

Please paste the GitHub Repo link.
Github Link:- https://github.com/PavanKumar181098/Bike-Sharing-Demand-Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)
Write here the short summary

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern.
The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

For the First step, the data is imported from a csv file and converted to a panda's DataFrame . Pandas is an inbuilt library in python that is used in handling and manipulating Dataframes. Dataframe is basically a collection of Rows and Columns.

Once the data is clean, we check for the statistics of the cleaned data. The statistics of the data tells us the mean, median and the distribution of the data and some more info.

After the data is cleaned, Visualizations can be done on the data and inferences can be noted.

One such visualization is to find the correlation between the variables and find out that Temperature and Dew point Temperature are highly correlated and dew point has to be deleted from the DataFrame as it is lower in importance compared to Temperature.

Another such visualization is to find out what are the months in which the people in seoul like to drive the rented bicycles. Plotting that we found that people like driving more in summers than in any other weather.

Similarly, we can plot the frequency of bikes rented to the hour of the day and found out that there is a significant increase in the number of rentals at 08:00 in the morning and 18:00 in the evening. We also see that there is a huge demand in the evening from 16:00 to 21:00, peaking at 18:00. So, we can infer that people using the rented bikes are mostly office going people.

Also, we plotted the number of rented bikes according to the weekdays and found out that most of the traffic is on weekdays and least on Sunday.

Similarly, by plotting the temperatures in which people like to ride the bikes to find out that the sweet spot of temperature for riding is around 25 degrees Celsius.

Also, by plotting the time of day (day or night) in which the riders prefer riding, to find out that more than 75% rides happen in daytime and the rest in night time.

Another such Visualization is plotting the amount of rainfall and snowfall in which riders are found renting the bikes most and found out that people are found riding in very less to nil rainfall and snowfalls.

Coming to the second part, which is Data Modeling in which we try out different Supervised Machine Learning Regression algorithms fits the best to our dataset and gives us the best accuracy (R2 scores).

Starting with a primitive Regressor Algorithm as Multiple Linear regression and fitting it to our dataset to find out that, this algorithm doesn't give us satisfactory results as the R2 score for testing and training datasets is around 55%. The RMSE value is around 420. So, we end up discarding this model and look into new models.

Let us take a look at the next Algorithm "Decision Tree Regression". Decision Tree Regressor is a fairly complex algorithm when compared to Linear Regression models. Coming to the R2 scores, fairly appealing with R2 scores around 84% and 77% respectively on training and testing datasets. The RMSE value is around 300.

Last but not the least, Random Forest Regressor which is more complex than Decision Tree and hence, we could expect more accurate results. As per our expectation the accuracy results are better than Decision tree Regressor, with training R2 scores of 98% and testing scores of 89%. The RMSE value is around 200.

Coming to our last regressor, the XGBoost regressor. As it is the most complex of all our previous models and hence has the best results in the lot. The R2 scores for Training dataset are 90% and the score for the testing dataset is 88.5%. The RMSE value is around 200, the same for the Decision Trees algorithm. The XGBoost Regressor algorithm is also used to find out the best/important features of the dataset.

Concluding, We choose to go ahead with Random Forest Regressor as it had the best performance and R2 scores. The Random Forest algorithm is used to predict the bike count and plot the graph in order to ensure the availability of bikes in order to reduce the waiting time.