Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1)**ArunTeja Lonka:**

(arunteja.lonka@gmail.com)

1. Analyzing the problem statement
2. Exploring new python packages
3. Future Engineering
    a) Introduced new features
    b) Introducing Dummy Variables
4. Data Visualization:
    a) Heatmap
    b) Distribution of Dependent Variable
    c) Joint Plot
    d) Scatter plot
5. Regression Analysis:
    a) Linear Regression
    b) Lasso
    c) Ridge
    d) Elastic Net
    e) Random forest
    f) Gradient boosting
6. PPT

2)Upasana Kumari:

(toupasana@gmail.com)

1. Analyzing the problem statement
2. Collecting information about Seoul Public Bike
3. Exploring new python packages
4. Correlation Analysis
    a) Between Independent Variables
    b) Between Dependent and Independent Variables
5. Data Visualization:
    a) Bar graphs
    b) Distpolt
6. Regression Analysis:
    a) Linear Regression
    b) Lasso
    c) Ridge
    d) Random forest

3) Zunaid

(qureshizunaid686@gmail.com)

1. Analyzing the problem statement
2. Collecting information about Seoul Public Bike
3. Exploring new python packages

4. Correlation Analysis
   a) Between Independent Variables
   b) Between Dependent and Independent Variables
5. Data Visualization:
   a) Bar graphs
   b) Distplot
6. Regression Analysis:
   a) Linear Regression
   b) Lasso
   c) Ridge
   d)XG boosting

4)Sukesh Shetty

(Sukesh3112@yahoo.com)

1. Analyzing the problem statement
2. Collecting information about Seoul Public Bike
3. Exploring new python packages
4. Correlation Analysis
   a) Between Independent Variables
   b) Between Dependent and Independent Variables
5. Data Visualization:
   a) Bar graphs
   b) Distplot
6. Regression Analysis:
   a) Linear Regression
   b) Lasso
   c) Ridge
   d)XG boosting

Please paste the GitHub Repo link.https://github.com/AruntejaLonka/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)

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 crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

So, in our dataset there were no null values and after handling outliers we observed that Dew Point Temperature was correlated so after removing that we performed Label Encoding, Feature Engineering and splitting them for training & testing datasets (80-20), we applied 7 models (, i.e., Linear Regression, Rigid Regression, Lasso Regression, Random Forest Regressor, Gradient boosting, Elastic Net Regressor and XGboostig Regressor). After using these models, we selected Random Forest Regressor and performed Tuning in it as it was giving the best results, we observed that Rental Bike Count was dependent on Hour, Temperature, Humidity, Functioning Day and if it is Raining outside or not.

- **Bike rental count is mostly correlated with the time of the day as it is peak at 10am morning and 8pm at evening**
- **We observed that bike rental count is high during working days than non-working days.**
- **We see that people generally prefer to rent bike at moderate to high temperatures, and when little windy**
- **It is observed that highest number bike rentals count in Autumn & Summer seasons and the lowest in winter season.**
- **We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.**
-

We observed that with increasing humidity, the number of bike rental counts decreases

Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set.

Feature Importance value for Random Forest and Gradient Boost are different.

We can deploy this model.
However, this is not the ultimate end. As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well.

As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time.

Therefore, having a quality knowledge and keeping pace with the ever-evolving ML field would surely help one to stay a step ahead in future.