Capstone Project Submission

Team Member's Name, Email and Contribution:

Shubham Kishorrao Joshi (shubhkjoshi5@gmail.com)

- Data Wrangling

➢ SeoulBikeData

➢ Loading and Pre-processing

➢ Structuring Data

➢ Enriching Data

➢ Data Validation

- Data Mining
- Data Analysis
- Model Development

➢ Decision Tree Regression

➢ Random Forest Regression

➢ XGBoost Regression

➢ Random Forest Regression with gridSearchCV

- Visualizations

➢ Box Plot

- Segmentation
- Summarization
- Observations
- Conclusions

Kanika Kakra (kostubikakra11@gmail.com)

- Data Wrangling

➢ SeoulBikeData

➢ Loading and Pre-processing

➢ Structuring Data

➢ Enriching Data

➢ Data Validation

- Data Mining
- Data Analysis
- Model Development

➢ Linear Regression

Please paste the GitHub Repo link.
Github Link:- https://github.com/11-Kani/BikeSharing_DemandPredction_CapstoneProject_II Drive Link:- https://drive.google.com/drive/folders/1v-_Ax5E90WvbJWLaOmuneY1VuG-4lFtJ

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

A bike rental or bike hire business rents out bicycles for short periods of time, usually for a few hours. 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. Therefore we have implemented such models which predicts the bike count required at each hour for the stable supply of rental bikes.

In this project, we explored several types of information that influence bike rental count. Below is a quick summary of exploratory data analysis. These are the independent variables having an impact on while predicting values for bike count.

We initially did EDA on all the features of our dataset. We first analyzed our dependent variable, 'Rented Bike Count' and also transformed it. Next we analyzed categorical variable and dropped some variables which were not optimum for model building. We also analyzed numerical variable, found out the correlation, distribution and their relationship with the dependent variable.

The insights that independent variables shows.

Working Day: for a Working Day where the rental count high at peak hours (Most demand for bike is in between 7 to 9 AM and 5 to 8 PM .)

**Non-working day* *:where rental count is more or less uniform across the day with a peak at around noon.

Hour of the day: Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on a working day and is mostly uniform during the day on a non-working day

Season: We see highest number bike rentals in Fall (July to September) and Summer (April to June) Seasons and the lowest in Spring (January to March) season

Weather: As one would expect, we see highest number of bike rentals on a clear day and the lowest on a snowy or rainy day

Humidity: With increasing humidity, we see decrease in the number of bike rental count.

After this we move to model building through implementing Supervised machine Learning regression algorithms. We have distributed the features into categorical and numerical values .To handle the skewness of data we applied different methods of transformation. We split our data into train and test in proportion of 75% and 25% respectively. Upon implementation of various algorithms No overfitting is seen, as we can see the models are performing well with the test data with good results. Random forest Regressor, Random forest Regressor with gridsearchCV and XGB Regressor gives the highest R2 score of 92%, 89% and 86% respectively for test dataset. Feature Importance value for Random Forest and Gradient Boost are different.

Therefore the Random Forest Regression has given the maximum value for R-sq.Thriving to conclusion that our model predicts the value with utmost accuracy but other factors are time dependent the features like[Snowfall, Rainfall Temperature,etc] are changing in every span of time so we will always keep improvising our model such that it rigorous changes in values should not affect the performance of model.