Capstone Project-2
Bike Sharing Demand

Name, Email and Github link:

Name: - Prachi Jadhav
Email ID: - prachisj12@gmail.com
Github Link: -
https://github.com/PrachiJadhav12/Seoul_Bike_Sharing_Prediction

Name: - Dr. Raj Kumar
Email ID: - rjk.kaushik86@gmail.com
Github Link: - https://github.com/rajkumarpec/Supervised-ML

Summary of Project

The goal of data science is to construct the means for extracting business-focused insights from data. This requires an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities.

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.

In the initial phase, we have focused more on the data cleaning and analyzed data in various categories. In later part we have tried to come out with conclusion for give problem statement. From this we have tried to bring out best results out of our analysis.

We have faced major challenge in data cleaning on which feature to be dropped or kept. After data exploration we found that there was no duplicate and null values. We have converted columns to appropriate data types and also added appropriate columns.

From Exploratory Data Analysis we got to know that few columns which have categorical data having numerical type. Also, we have added columns from date.

With the cleaned data, we have performed Exploratory Data Analysis to understand the behaviour of our target variable. Our target variable was positively skewed. We did square-root transformation to make it normally distributed.

We drew some conclusion form EDA like on a regular day, there is a huge demand for rental bikes on morning 8 AM and Evening 6 PM. On holidays and weekends, the demand for rental bikes increases gradually throughout the day. The data contains outliers, but we didn’t handle them since by doing so, we may eliminate the patterns in the data we discovered.

Since the data contains outliers, and many categorical attributes, It won’t be wise to fit linear models, as they will give high errors. We will use tree models instead, since they can handle outliers and categorical attributes better than linear models.

We will use decision tree as a baseline model. Subsequently, to get better predictions, we will use ensemble models: Random forests, GBM, XG Boost. Final choice of model will depend on whether interpretability or accuracy is important to the stakeholders.

We fit 7 different type of models namely Linear regression, lasso Regression, ridge regression, decision trees, random forest, gradient boosting and extreme Gradient Boosting.

By observing Evaluation matrices for all the models-

- Linear Regression, Lasso and Ridge are not at its best.
- Decision Trees, Random Forest are quite good with linear models, but they are not giving optimum prediction.
- Gradient boosting type models are giving better results.

Thus, we have successfully built predictive models that can predict the demand for rental bikes based on different weather conditions and other. The XG Boost prediction model had the lowest RMSE.

The final choice of model for deployment depends on the business need; Out of all above models Random forest Regressor gives the highest R2 score of 98% for Train Set and XG Boost Gridsearch CV gives the highest R2 score of 91% for Test set.