Capstone Project

Bike Sharing Demand Prediction Project

(Supervised Machine Learning- Regression)

Piyush S Kutemate, Prince Chauhan

1. Problem Statement-

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. The client is Seoul Bike, which participates in a bike share program in Seoul, South Korea. An accurate prediction of bike count is critical to the success of the Seoul bike share program. 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.

2. Data Description-

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours

3. Steps involved-

1) Importing important libraries Our main motive through this step was to import all the important libraries to help us explore the problem statement and perform EDA to draw conclusion on the basis of the data set.
2) Understanding the data set Next, we worked on checking the data set. How big the data set is? How many rows and columns are available? What could be the important columns to solve the problem statement? How many null values we have in the data set? We imported the important libraries along with our data set.
3) Null values Treatment Our dataset contains a large number of null values which might tend to disturb our insights. Hence, we replaced them with '0' for numerical data and 'undefined for 'categorical data' to get a better result.
4) Exploratory Data Analysis After treating the null values, we started with the EDA. We performed EDA.

4. Exploratory Data Analysis

5. Preparation of data for model building-

a. With the heat map we dropped highly correlated variables.
b. Later by using variation inflation factor we dropped 'Visibility' and 'Humidity' features as they had VIF value more than 5.
c. Next, we created dummy variables for categorical Seasons column and did mapping with 0 and 1 for holiday and functioning column.

Thus, we prepared our data for model building.

6. Model Selection and Evaluation-

As this is the regression problem, we are trying to predict continuous value. For this we used following regression models.

1) Linear Regression
2) Lasso regression (regularized regression)
3) Ridge Regression(regularized regression)
4) Decision Tree regression.
5) Random forest regression
6) Gradient Boosting regression.

7. Observation-

1) **Linear, Lasso, Ridge and Elastic Net:**
   Linear, Lasso, Ridge and Elastic regression models have almost similar R2 scores (61%) on both training and test data. (Even after using GridserachCV we have got similar results as of base models).

2) **Decision Tree Regression:**
   On Decision tree regressor model, without hyper-parameter tuning, we got r2 score as 100% on training data and on test data it was very less. Thus our model memorized the data. So it was a over fitted model.After hyper-parameter tuning we got r2 score as 88% on training data and 83% on test data which is quite good for us.

3) **Random Forest:**
   On Random Forest regressor model, without hyper-parameter tuning we got r2 score as 98% on training data and 90% on test data. Thus, our model memorized the data. So, it was a over fitted model, as per our assumption
   After hyper-parameter tuning we got r2 score as 90% on training data and 87% on test data which is very good for us.

4) **Gradient Boosting Regression (Gradient Boosting Machine):**
   On Random Forest regressor model, without hyper-parameter tuning we got r2 score as 86% on training data and 85% on test data. Our model performed well without hyper-parameter tuning. After hyper-parameter tuning we got r2 score as 96% on training data and 91% on test data, thus we improved the model performance by hyper-parameter tuning.

9. Conclusion-

According to calculations, Gradient Boosting Regression (GridSearchCV) and Random forest (GridSearchCv) gives good r2 scores. We can deploy these models.