Capstone Project - 2

Project Title: Seoul Bike Sharing Demand Prediction

By

Sapana Pawar

Points for Discussion

- Problem Statement
- Introduction
- Data Summary
- EDA
- Data Visualization
- Feature Engineering
- Model Implementation and Evaluation
- Hyperparameter tuning
- Feature Importance
- Conclusion

Problem Statement

- 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.

Introduction

- Currently rental bikes are introduced in many urban cities for the enhancement of mobility and comfort. The purpose of this movement is to modernize cities and encourage people to head to a green world. Let's take the examples of Paris in 2007, where "velibs" were introduced and Amsterdam, where there are more bikes than cars. The goal is to facilitate the commute in the Seoul and reduce the amount of cars and pollution. Indeed, the development of the way to commute has reduced the use of cars to go to work and visit the city.
- It is important to make the rental bike available and accessible to the public, as it provides many alternatives to commuters in metropolises. There are a lot of advantages to bike rents, it is convenient because it permits people not to keep the bike all day long, whether it is at work or at school. Furthermore it is the healthiest way to travel and it has many environmental benefits.

- In this capstone project I have developed a model which could predict the bike count required at each hour for the stable supply.

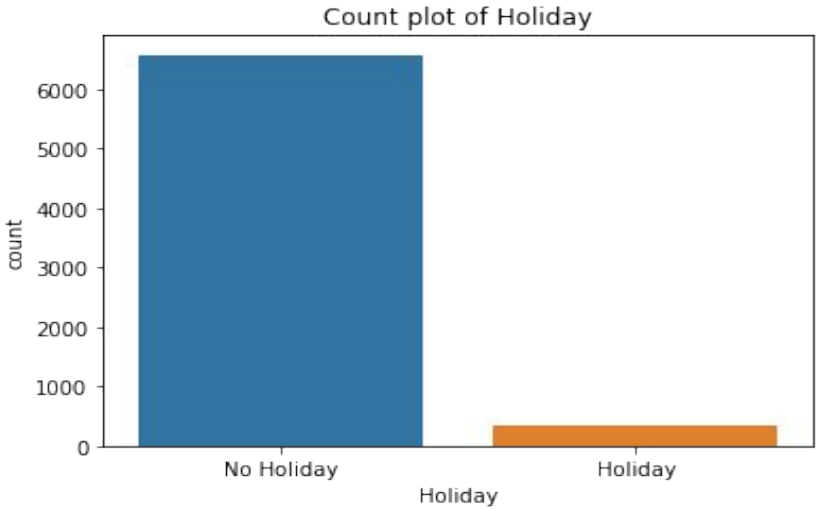
Data Summary

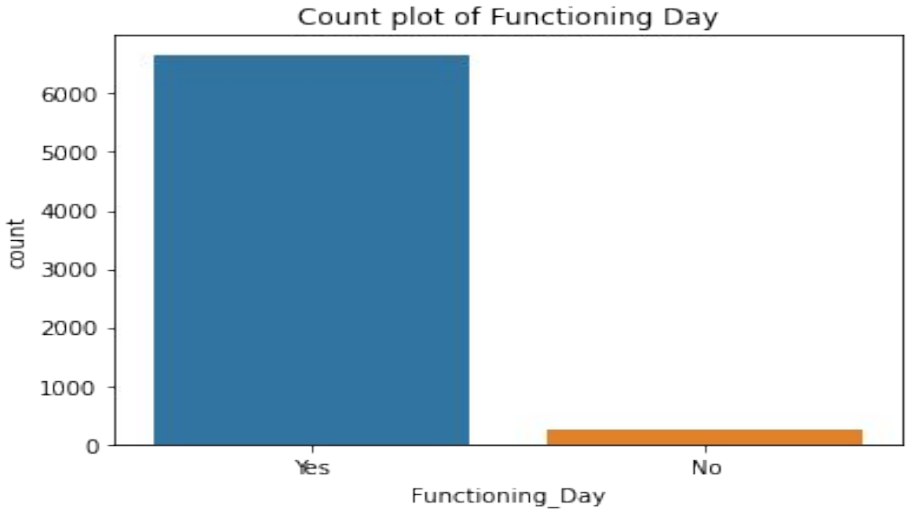
- **Date :** The day of the day, during 365 days, type : str
- **Rented Bike Count :** Number of rented bikes per hour which is the target, type : int
- **Hour:** The hour of the day, type : int
- **Temperature(°C):** Temperature per hour, type : Float
- **Humidity(%):** Humidity in the air in %, type : int
- **Wind speed (m/s) :** Speed of the wind in m/s, type : Float
- **Visibility (10m):** Visibility in m, type : int
- **Dew point temperature(°C):** Temperature at the beginning of the day, type : Float
- **Solar Radiation (MJ/m2):** Sun contribution, type : Float
- **Rainfall(mm):** Amount of rain in mm, type : Float
- **Snowfall (cm):** Amount of snow in cm, type : Float
- **Seasons:** Season of the year, type : str
- **Holiday:** If it is holiday period, type: str
- **Functioning Day:** If it is a Functioning Day, type : str

EDA

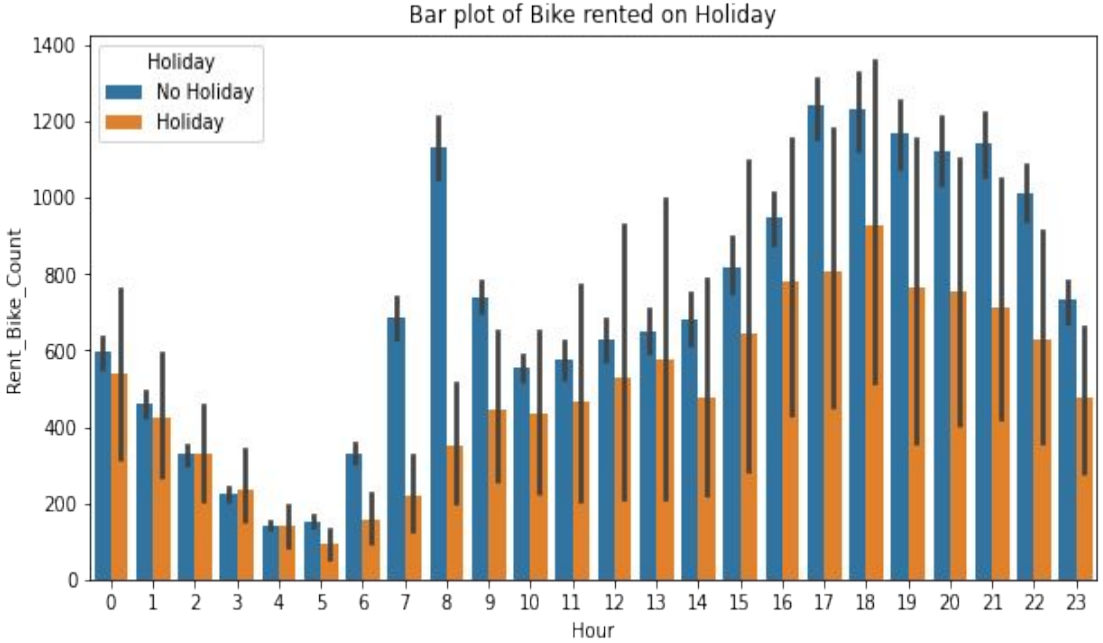
In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

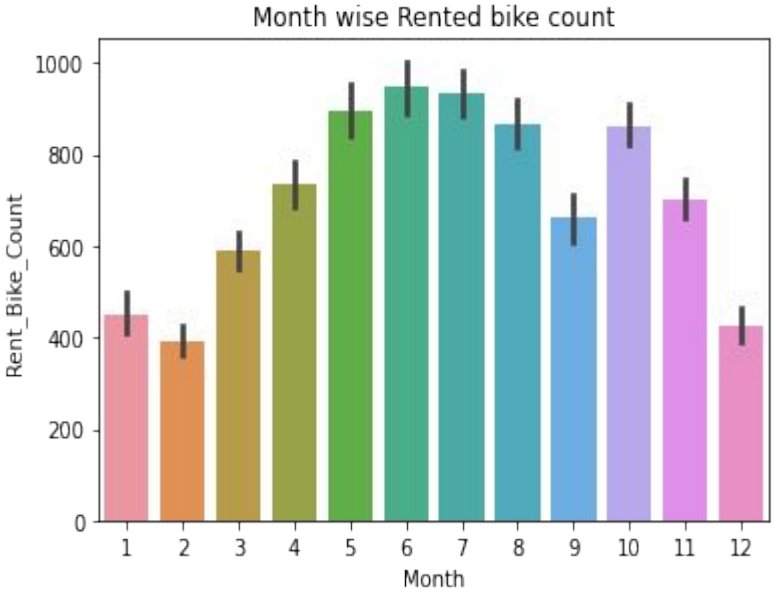
- The count of the day when the day was not holiday is more than the day when the day was holiday.
- The count of functioning day was more than that day's where there were no functioning day.





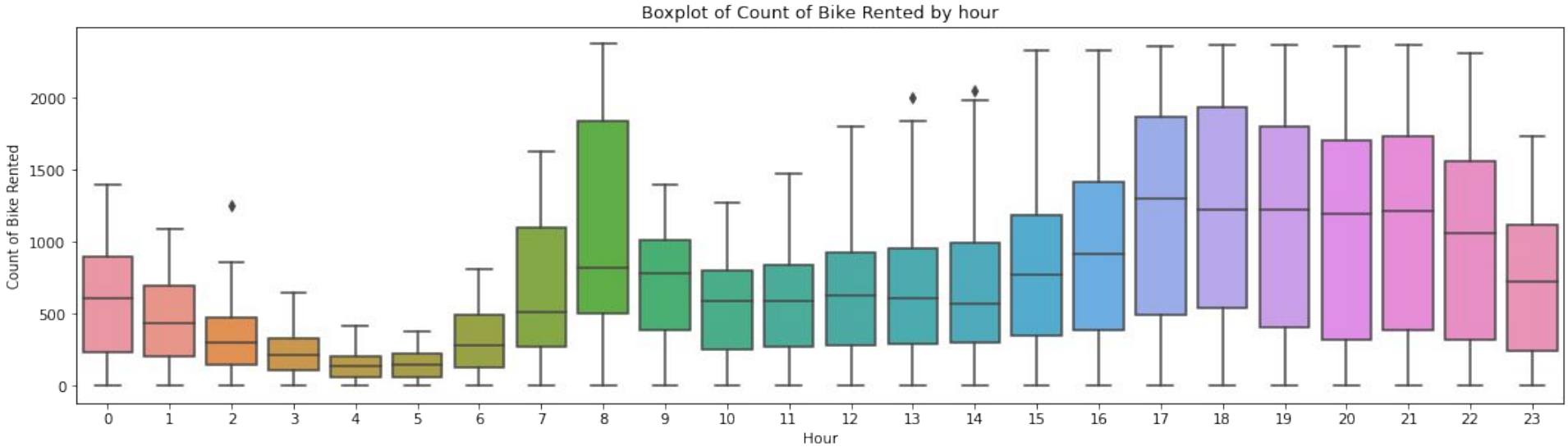
EDA continued…





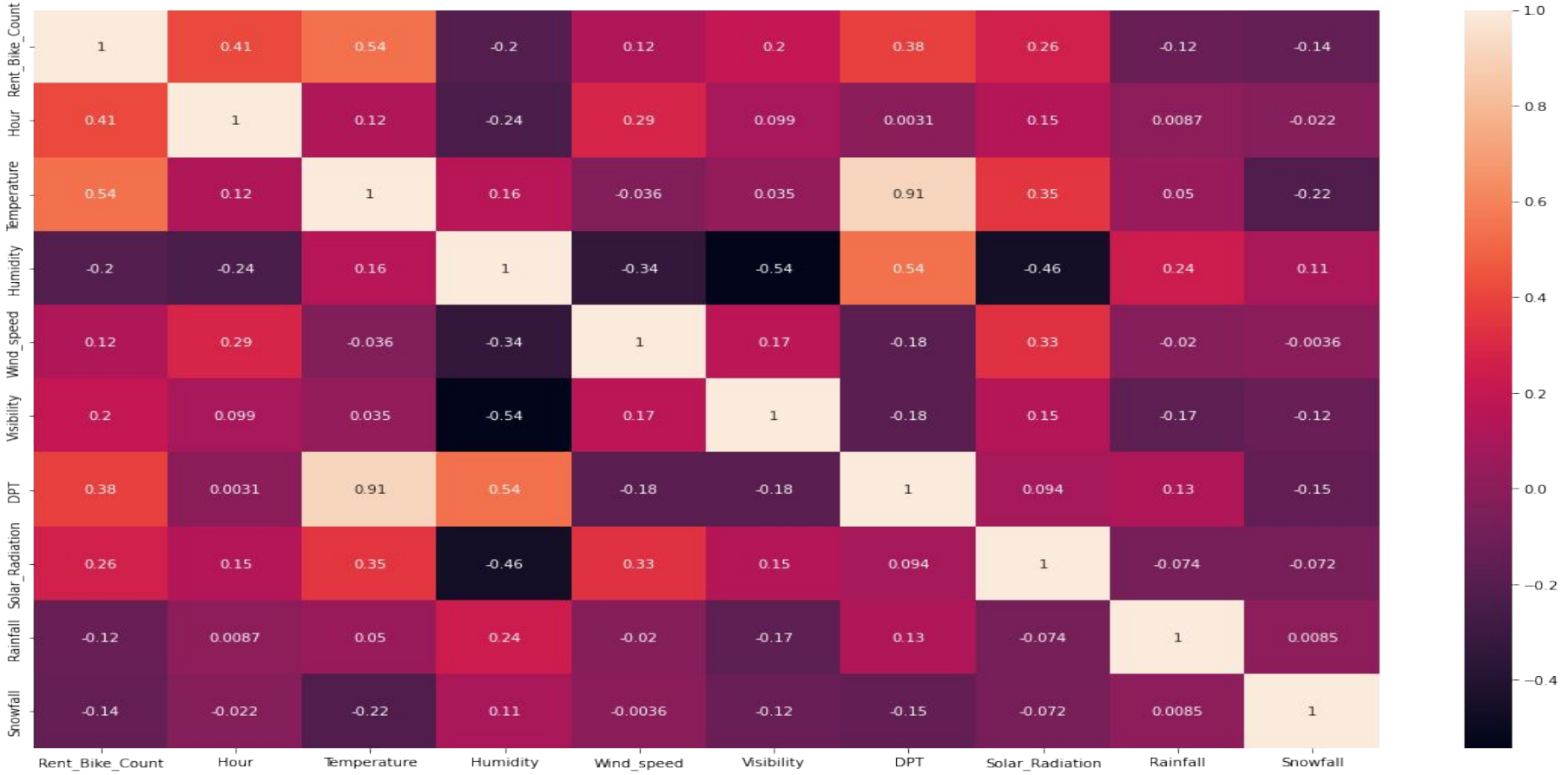
- In this plot we can say that when the day was not holiday, rented bike count is maximum than when the day was holiday.
- Most of the bike rented in the month of June.

EDA continued…



- Above, we can see the trend of bike rent over hours. Quickly, we'll segregate the bike rent in three categories:
- High : 7-9 and 16-22 hours
- Average : 10-15 hours
- Low : 3-5 hours Here we have analyzed the distribution of total bike rent.

Correlation Heatmap



Feature Engineering

Label Encoding:

- Implemented Label Encoding on the columns' holiday' and 'functioning day'.

One Hot Encoding:

- One-Hot Encoding is the process of creating dummy variables.
- Implemented One hot encoding on the column 'Season' and created dummy variables.

Feature Selection:

- Dropped column 'Date', 'Day' and 'Year' which are not important.
- 'Snowfall' and 'Rainfall' are highly skewed towards zero so we dropped them.

Model Implementation

For modeling we tried various algorithms like:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net Regression
5. Decision Tree Regressor
6. XGBoost Regressor
7. Random Forest Regressor

Model Evaluation

Evaluation Metrics:

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at MAE, MSE, R-squared and RMSE.

1.Mean Absolute Error (MAE):

2.Mean Squared Error (MSE):

3.Root Mean Squared Error (RMSE):

4.R – Squared (R2):

Model Evaluation Continued…

Model	MAE	MSE	RMSE	R- Squared
1. Linear Regression	308.7282	159536.57	399.42	56.31%
2. Lasso Regression	308.7100	159532.38	399.42	56.31%
3. Ridge Regression	308.7253	159535.61	399.41	56.31%
4. Elastic Net Regression	315.4553	171410.07	414.01	53.05%
5. Decision Tree Regressor	247.3007	118143.64	343.72	67.64%
6. XGBoost Regressor	165.2677	60720.73	246.41	83.37%
7. Random Forest Regressor	139.8953	50467.60	224.64	86.18%

Hyperparameter Tuning

- Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm.
- Random Forest Regressor model has given better results as compared to other algorithms that I have tried.
- So to increase R2 score and reduce error I have done hyperparameter tuning on RF model.

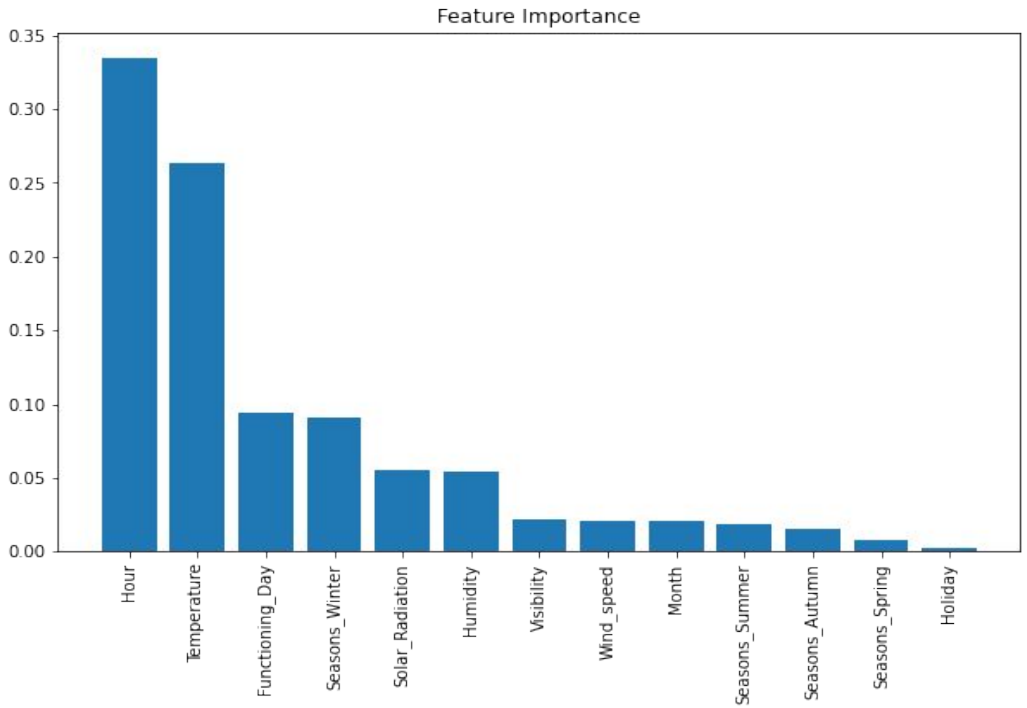
Types of Hyperparameter Tuning

1. **Random Search**
2. **Grid Search**

- After applying hyperparameter tuning on RF model R2 score is increased and errors are reduced by some amount.

Feature Importance

- Hour, Temperature and Functioning day are the most important feature to predict rentend bike count.
- While Holiday, Seasons_spring and Seasons_Autumn are less important to predict the target variable.



Conclusions

- Random forest model has given better result as compared to other algorithms that I have tried.

After Hyperparameter Tuning:

- R2 score is increased from **0.8617816523757432** to **0.8626419132455122.**
- MAE is reduced from **139.89526721232545** to **142.2532542757118**.
- MSE is reduced from **50467.600547279726** to **50153.49389871826**.
- RMSE is reduced from **224.64995114016767** to **223.9497575321712**.

Thank You