**Seoul Bike Sharing Demand Prediction**

**Aashruti Agrawal, Raneev K, Kunal Mahadik**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Rental Bike are slowly getting its acclaim in the urban cities for better mobility comfort to the public. In order to maintain the smooth operation, availability and accessibility of the rental bike with lesser waiting time period is the most crucial concern.

We were provided the Seoul bike sharing dataset containing various features, where we are supposed to predict the count of rental bike required each hour in order to maintain the constant supply of the rental bike.

***Keywords: machine learning, rental bike, regression analysis, demand prediction***

**1.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.

The main objective is to build a predictive model, which could help them in predicting the demand of rental bike sharing in order to maintain the bike count to make it available and accessible to the users. Also, this would help in reducing the waiting time which in turn potentially increases their market.

* Rented Bike Count - Number of rented bikes per hour. It is our target variable as well
* Date : The date of rented bike
* Hour - Hour of the day ranging from 0-23.
* Temperature (°C)-Temperature in Celsius
* Humidity(%) - Humidity in the air in percentage.
* Wind speed (m/s) - Speed of the wind in m/s.
* Visibility (10m) - Visibility in m.
* Dew point temperature(°C) - The temperature at which the water starts to condense out of the air.
* Solar Radiation (MJ/m2) - Electromagnetic radiation emitted by the Sun.
* Rainfall(mm) - Amount of rainfall in mm.
* Snowfall(cm) - Amount of snowfall in cm.
* Seasons - Season of the year. There are 4 seasons namely, Summer, Winter, Spring and Autumn.
* Holiday – Whether it is holiday of the user or not.
* Functioning Day - Whether the day is functional or not in terms of rental bike rented for that day.

**2. Introduction**

### In a span of few decade, the sharing of bicycle system has seen enamours growth (Fishman, [2016](https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789)). This system is a recently developed transportation system which provides people with bicycle for common use. Bicycle system provides user to rent a bike from one docking station, where user can ride and then return in another docking station. Amsterdam in Netherlands was the one where initial bicycle sharing system has started in 1965 (Shaheen, Guzman, & Zhang, [2010](https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789)). Main motive of the system was to focus on environmental and social welfare. With enormous advancement of Intelligent Transportation System and information technology after 2000s, this bicycle system employed globally. Situation over decade has changed in sharing bicycle. Today it is much easier for the public to rent bicycles. Global Positioning System enabled mobile application allows people to know the nearby bicycle station for renting the bicycle.

### Till today there are more than 50 countries having 712 which implemented bicycle sharing method (Shaheen, Martin, Cohen, Chan, & Pogodzinski, 2014). They have now found to be important face of transportation system in major cities due to several factors such as health problems, heavy traffic and environmental conditions. Bike sharing Systems namely OFO and Mo-bike in places like Beijing has enabled people to find position of unused bicycle and use them. Once the bicycle is used, it can be locked at any docking station across the city. For expanding availability of bicycle for public use, the operators running this service allocate a truck that collects bicycles parked in various station and relocate them to the original station gradually (Schuijbroek, Hampshire, & Van Hoeve, 2017). There are number of policy issues involved in the management of intelligent bicycle sharing system (Gast, Massonnet, Reijsbergen, & Tribastone, 2015). Many countries have bike sharing system, such as Ddareungi is a bike sharing system in South Korea, which started in the year 2015, known as Seoul bike in English. It was started to overcome issues like greater oil prices, congestion in traffic and pollution in the environment and to develop a healthy environment for citizen of Seoul to live. Han River is the initial place where Ddareungi was first started on October 2015 in Seoul, few months later, total number of bike sharing station touched 150 with as much as 1500 were there. In order to cover the entire people in Seoul, in 2016 there is a gradual incline in number of docking station. As large as 20,000 bikes were made available which was confirmed by Seoul Mayor Park won-soon. With the help of growing technologies, Seoul city is now equipped with 1500 bike renting station which are operational round the clock. With the help of internet-enabled device or mobile phone, people can know the number of bikes available for the people to rent. Bike are locked which can be unlocked with the help of password which people accessing to it will receive the password through mail. Users are allowed and can rent and leave the bike in any station. Seoul Rental Bikes are built to be utilized by all kinds of people including women, elderly persons and infirm. Seoul Bikes are manufactured using durable and light-weight materials. This giving user more stability in driving and convenience. These public bikes are made available to the persons who are 15 and older. Seoul bike docking stations are equipped in excessive traffic areas including subway entrances/exits, bus stops, residential complexes, public offices, schools, and banks. Docking stations are computerized stands for the purpose of pickup and dropoff of the rental bikes. Users of Seoul public bikes can rent and return rental bikes at any docking station. Users can verify their trip details (distance, duration) and measure of bodily activities (burnt calories). With this kind of smart technology and convenience, the use of Rental bike is increasing every day. So, there is a need to manage the bike rental demand and manage the continuous and convenient service for the users.

### reference: *https://www.tandfonline.com/doi/full/10.1080/22797254.2020.1725789*

# 

# **3. Challenge Faced:**

## Majorly, handling the distribution of data and handling outliers was the major challenge faced by us. Since, entire dataset was skewed including target variable, we have applied transformation to only target variable.

Feature engineering and feature selection was a tricky part as all the independent features were somehow correlated with each other.

**4. Approach:**

* **Data Preprocessing**

Firstly, after loading the dataset we preprocessed the raw data to have better understanding of the features, and make the data high quality four analysis ready. We performed the following steps in order to clean our dataset:

* + **Null values**: luckily, this dataset doesn’t have any null values
  + **Duplicated values**: No duplicated values were found
  + **Improper format**: updated datatype of date feature from object to datetime64. Also, we have renamed the features for our convenience.
  + **Outlier handling**: The entire dataset is very skewed and has many outliers. Square root transformation has been applied on target variable to treat the skewness and outlier.
  + **Feature engineering**: Some of the new features were engineered like temperature and dew point temperature, month, weekday, year. Some of the irrelevant or redundant columns were dropped
* **Exploratory Data Analysis**

We performed EDA trying to find out the patterns and behavior in the dataset. We compared our target variable that is Rental\_bike\_count with other independent variables to figure out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable. We addressed the following questions for EDA analysis:

* + Summer was the most season for bike renting, with June being the most active month in terms of bike renting.
  + Monday comes out to be most active day in terms of bike renting.
  + Most of the bikes were rented on working day.
  + Most of the bike were rented at morning 7 and evening 6, indicating users might use them to commute to their work place.
  + In terms of temperature, most of the bike were rented at room temperature (250C), and least during negative temperature.
  + In terms of humidity, wind speed, rainfall, snowfall we saw decremental trend
  + Visibility and Solar radiation do not show any impact on renting bike.

Also, relation amongst different independent variable was also looked for using correlation heatmap. This helps us finding the multi collinearity, which becomes pivotal when it comes to linear regression model.

* **Encoding of categorical columns**

We used One Hot Encoding to encode our categorical features to make them recognizable for building machine learning model, since categorical features in string format cannot be translated by the machine and needs to be converted to numerical format.

* **Detecting Multi collinearity**

Since, to implement linear regression model, knowing multicollinearity amongst independent variable is an important aspect. We have applied variance inflator factor (VIF) to detect multi-collinearity. VIF score more than 10 indicates the presence. All of our score were less than 10, indicates independent relationship.

* **Standardization of features**

Standardization of features was performed to bring out all the features to uniform scale which would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Fitting different models**

For modelling we tried various regression algorithms like:

1. **Linear regression**
2. **Lasso regularization**
3. **Ridge regularization**
4. **Elastic net regularization**
5. **Decision tree regressor**
6. **Random Forest regressor**
7. **Gradient Boosting regressor**

* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting especially in case of tree based models like Random Forest regressor and GBoost regression. We have used GridSearchCV for hyper tuning the parameters.

* **Model Building**

Different algorithms were implemented to build the predictive model. The parameters were hyper-tuned using GridSearchCV. MSE, MAE, RMSE, r2 score and adjusted r2 score were calculated for each of the model to assess the performance of the model.

**1. Linear regression:**

For linear regression, the model performed good with r2 score of 0.75, 0.766 for train and test dataset respectively. This model does not show any over fitting or underfitting.

To get better accuracy, we tried regularization technique like L1, L2, elastic net regularization. However, we found the similar results in all the models which is aligning with the fact that we use regularization to overcome overfitting. Since, this model doesn’t has overfitting, same is depicting in the model performance. In a nutshell, elastic net performed best out of all linear regression models.

1. **Decision tree:**

Decision tree are usually very helpful when our dataset has many outliers. To hyper tune the parameters we have used GridSearchCV and again trained the model using the best estimated parameter.

We found that our decision tree algorithms did not show any overfitting but has lesser r2 score with 0.689, 0.680 for train and test dataset respectively. Feature engineering may help in getting r2 score for this case. We also tried searching for the feature of importance as decided in decision tree based on information gain.

1. **Random forest:**

Random forest regression is a robust algorithm used for classification as well as regression problem. It builds decision trees on different samples and takes their majority vote and takes average in case of regression. We found that with default parameter our model is performing well with r2 scores of 0.98 for train dataset and 0.90 for test dataset. Surprisingly, with hypertuned parameter our performance dropped to 0.72 and 0.69 for train, test dataset which is indicative that we were unsuccessful in hyper tuning the parameters, and we need to tune the parameter for better model performance.

1. **Gradient Boosting:**

Gradient boosting builds a model in a stage-wise fashion and generalizes the model by allowing optimization of an arbitrary differentiable loss function. It combines weak learners into a single strong learner in an iterative fashion. As each weak learner is added, a new model is fitted to provide a more accurate estimate of the response variable. We found r2 score of 0.95, 0.90 for train and test dataset respectively. The performed good but there is a scope for further feature engineering or further optimize the parameter to minimize the overfitting

1. **Summary & Conclusion:**

Lets summarise this entire project from beginning to end and at last we will write our conclusions.

At first we loaded our dataset and did data some data wrangling & EDA to clean our dataset and to understand the features thoroughly. Thankfully there was no missing and duplicate values were present in the dataset. Some of the feature names were quite lengthy so we renamed some features and checked for any wrong data types. “Date” was given as object datatype so we converted it into “date time” format. “Hour” was given as “int64” datatype but we don’t need it as continuous variable so we converted it into categorical variable.

Then in EDA part first of all we plot the distribution of target variable, unfortunately the distribution was slightly skewed towards right side and there was some outliers also. We did squire root transformation to normalize the distribution. We plot charts of different categorical variables and numerical variables against our target variable and found that in summer season rented bike count is maximum also around 8 am at morning and 6pm at evening there is a peak in rented bike count. We can say that most of the working professionals depends on rented bike for commute. When temperature is extremely low the rented bike count is also less and it follows a linear relationship upto 25-28 degree celcius.

After plotting correlation heatmap we found that “temperature” and “dewpoint temperature” are highly correlated each other, so we merge them together and create a new feature to remove multicollinearity.

After all these data preprocessing we implemented different machine learning algorithms like “linear regression”, ”lasso”, ”ridge”, “elastic net”, ”decision Tree”, “random forest”, “Random forest regression with GridsearchCV” and “Gradient Boosting gridsearchcv”. We found that

1. Almost all algorithms performed really well on both training dataset and testing dataset so we can say that variance is less and no issues of overfittings are present.
2. Both "Random forest regression" and "Gradient Boosting regression (gridsearch cv) has highest R2 score.
3. Performance on "Decision Tree" algorithm is comparatively less with an R2 score of 68%
4. For decision tree we got highest feature importance for "temperature\_and\_dewpoint temperature",incase of Random Forest "season\_winter" is the most important feature and for Gradient boost "temperature\_and\_dewpoint\_temperature" has highest feature importance value.

We know that this data is time dependent, the values for variables like temperature, solar\_radiation, wind\_speed 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.