**Seoul Bike Sharing Demand Prediction**

**Sujit Musale**

**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.

I 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, regression analysis, demand prediction, rental bike***

**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**

### A bicycle-sharing system, bike share program, public bicycle scheme, or public bike share (PBS) scheme, is a [shared transport](https://en.wikipedia.org/wiki/Shared_transport) service in which [bicycles](https://en.wikipedia.org/wiki/Bicycles) are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Docks are special [bike racks](https://en.wikipedia.org/wiki/Bicycle_parking_rack) that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place. Other systems are dock less.

### We have been given a dataset of rented bikes which are used in Seoul. The dataset consists of numerous records and features among which the target variable is rented bike count.

### The main objective here is to build a ML model which is optimal and which is able to predict the number of rented bikes per hour. This would help various bike sharing companies to maintain a stable supply through the city and meet the demand appropriately.

**3. Approach:**

* **Data Loading and cleaning**

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

* + **Null values**: luckily, this dataset doesn’t have any null values, so just checked.
  + **Duplicated values**: same like null value No duplicated values were found.
  + **Improper format**: updated data type of date feature from object to datetime64. Also, we have renamed the features for our convenience.
  + **Outlier handling**: The dataset is mostly right skewed and has many outliers. Square root transformation has been applied on Out put 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 like temp and dew point temp.
* **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.
  + Most of the bikes were rented on working day.
  + Tuesday comes out to be most active day in terms of bike renting.
  + Most of the bike were rented at evening 6, indicating users might use them to commute from their work place.
  + In terms of Months, most of the bike were rented at Jun, and least during Jan.
  + In terms of humidity, wind speed, rainfall, snowfall we saw decremental trend
  + Temp and Visibility show positive impact on renting bike.

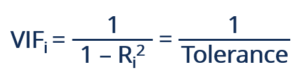
Also, relation amongst different independent variable was also looked for using correlation heatmap. This helps us finding the multi collinearity, which becomes pivotal when it comes to linear regression model.

* **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**

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in [regression analysis](https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/). It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.. We have applied variance inflator factor (VIF) to detect multi-collinearity.



**Ri2**represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

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**

Below are the some model that we used for regression :

1. **Linear regression**
2. **Lasso regularization**
3. **Ridge regularization**
4. **Decision Tree Regressor**
5. **Random Forest Regressor**
6. **Gradient Boosting Regressor**

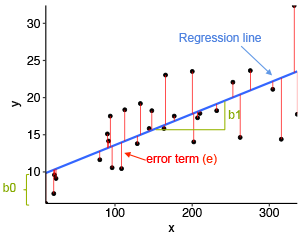
* **Tuning the hyperparameters for better accuracy**

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 and r2 score were calculated for each of the model to assess the performance of the model.

**1. Linear regression:** A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.



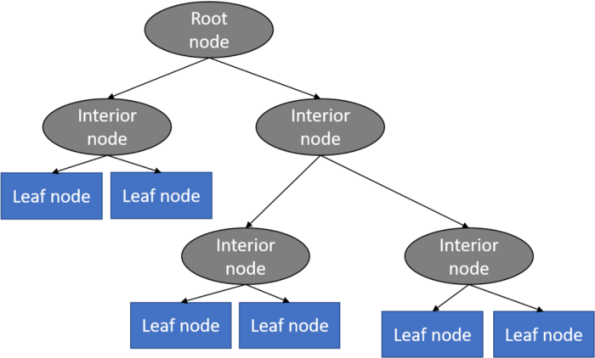
The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors. Here we try to minimize error for better fitting of model.

For linear regression, the model performed good with r2 score of 0.777, 0.784 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, 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, model perform almost same.

1. **Decision tree:**

Wecan usedecision tree for both classification and regression . here we use as a Regressor. Its work as the branches of tree so use to say Tree. As shown in below fig.



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.

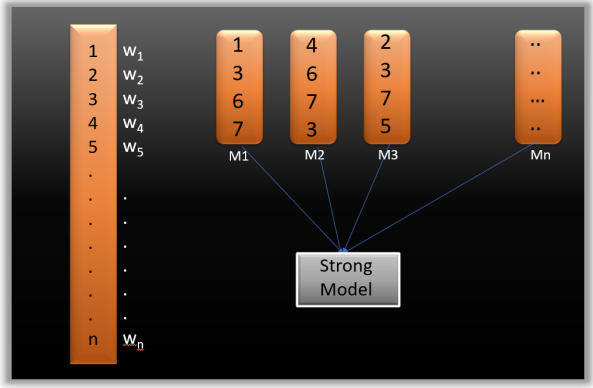
We found that our decision tree algorithms did not show any overfitting but has lesser r2 score with 0.795, 0.745 for train and test dataset respectively. Feature engineering may help in getting r2 score for this case. We also tried searching for the feature of importance as decided in decision tree 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 0.988 for train dataset and 0.913 for test dataset.

1. **Gradient Boosting:**

Its work as below



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 0.947, 0.91 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

**5. Conclusion:**

In conclusion, we can say that we implemented various models like Linear Regression, Lasso Regression, Ridge Regression,Decision Tree, Random Forest, Gradient Boosting on the dataset. . Out of all the built model, random forest and gradient boosting hypertuned via GridSearchCV model performed really well for this dataset with r2 score of 98.85% and 91.3% for random forest train and test dataset, for gradient boosting we achieved r2 score of 94.72%, 91.01% respectively. Although, Linear, lasso ,ridge and Decision Tree has comparatively less r2\_score of round 74% for all There were many outliers in the dependent variable which were fixed using square root transformation. The most important features which effect the rented bike count is temperature and dew point temperature.