**Seoul Bike Sharing Demand Prediction**

**By Gyanveer Meena**

**Abstract**

As a convenient, economical, and eco-friendly travel mode, bike-sharing greatly improved urban mobility. However, it is often very difficult to achieve a balanced utilization of shared bikes due to the asymmetric user demand distribution and the insufficient numbers of shared bikes, docks, or parking areas. If we can predict the short-run bike-sharing demand, it will help operating agencies rebalance bike-sharing systems in a timely and efficient way.

**Key words**

Machine learning, Data mining, Bike sharing demand prediction.

**1.Introduction**

According to recent studies, it is expected that more than 60% of the population in the world tends to dwell in cities, which is higher than 50% of the present scenario. Some countries around the world are practising righteous scenarios, renderings mobility at a fair cost and reduced carbon discharge. On the contrary other cities are far behind in the track. Urban mobility usually fills 64% of the entire kilometres travelled in the world. It ought to be modelled and taken over by inter-modality and networked self-driving vehicles which also provides a sustainable means of mobility. Systems called Mobility on Demand has a vital part in raising the vehicles’ supply, increasing its idle time and numbers.

**Problem Description**

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.

Today, bike-sharing systems are blooming across more cities around the world. To complete a short trip renting a bike is a faster way when compared to walking. Moreover, it is eco-friendly and comfortable too compared to driving.

**Problem Statement**

* Maximize: The availability of bikes to the customer.
* Minimize: Minimise the time of waiting to get a bike on rent.

**The main goal of the project is to** Finding factors and cause those influence shortages of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with bike demand prediction. Hourly count of bike for rent will also be predicted.

**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

**Attribute 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 – No Func(Non Functional Hours), Fun(Functional hours)

**3. Libraries used**

* **NumPy**: I have used the numpy library for the numerical analysis and better calculations.
* **Pandas**:  I have used the pandas for the analysis and manipulation and of better representation
* **Matplotlib**: I have used matplotlib for data visualization and graph plotations.
* **Seaborn**: I have used seaborn for the better visualizations and to use more colourful graphs.
* **Scipy**: Used for solving mathematical, scientific, engineering, and technical problems.
* **Warnings:** is used to show warning massages, but here we choose “ignore” for warnings.
* **Datetime:** this library use for manipulating date and time.
* **Sklearn:** the sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.  Sklearn is used to build machine learning models.

**Steps Involved**

**1. MISSING VALUES**

One of the ways to handle missing values is to simply remove them from our dataset. We have known that we can use the null() and not null() functions from the pandas library to determine null values. Since there are no missing values in this data set, we are not going to do anything now.

**2. DATA DUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset. Since there is no duplicate value in these datasets.

**3. EXPLORATORY DATA ANALYSIS**

After loading the dataset, we performed this method by comparing our target variable that is bike\_count with other independent variables. This process helped us figuring 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.

**3.1 Categorical comparison**

* **Summer** season have highest bike counts followed by Autumn and Spring. Lowest counts were in Winters.
* **June** has the highest bike counts and least in January.
* Most of the bikes rented in the morning and evening, indicating users might use them from workplace to home.
* All the bikes were rented in **Functioning Day.**
* **Weekdays** and **Weekends** have not much difference in bike counts
* **No Holidays** have the high bike counts than on Holidays.

**3.2 Numerical Comparison by regression plotting**

  Regression plots as the name suggests creates a regression line between 2 parameters and helps to visualize their linear relationships.

* **Temperature** shows positive relation with dependent variable. 20 **° to 30 ° have highest bike counts.**
* **Humidity shows negative relation means more humid less user.**
* **Wind speed and Visibility both does not affect that much but have slightly positive relation.**
* **Solar radiation shows positive relation, means high radiation less bike counts.**
* **Snowfall and Rain both shows highly negative relation, means people don’t rent bike in rain and snow.**

**4. Normalizing the Data**

**We are performing the normalization on our dependent variable because our dependent variable is highly right skewed and have outliers. There is assumption of regression that data is following normalised distribution. Performing square root method to normalize the data.**

**5. Correlation Heatmap**

Correlation heatmaps are a type of plot that visualize the strength of relationships between numerical variables. Correlation plots are used to understand which variables are related to each other and the strength of this relationship.

In our project, heatmap showing that some features are positively correlated and some are negatively.

**6. Feature Selection**

In these steps we using information gaining method finding correlation coefficients between variables mostly effects that are removed from our data set i.e. **temperature** and **dew-point temperature** are most effects for this we taken weighted average and removed this items removed from the dataset

**7. 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.

**8. Fitting the Models**

Here these are the models I have used for analysis:

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

**9. Tuning the hyperparameters for better accuracy**

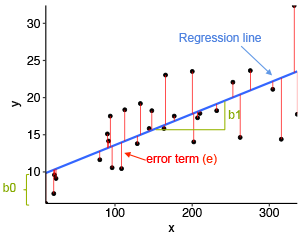
Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting especially in case of tree-based models **like Random Forest regressor** and **GBoost** regression. We have used **GridSearchCV** for hyper tuning the parameters.

**10. 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.

**10.1 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 **0.77,** **0.77** same for train and test dataset. This model does not show any over fitting or underfitting.

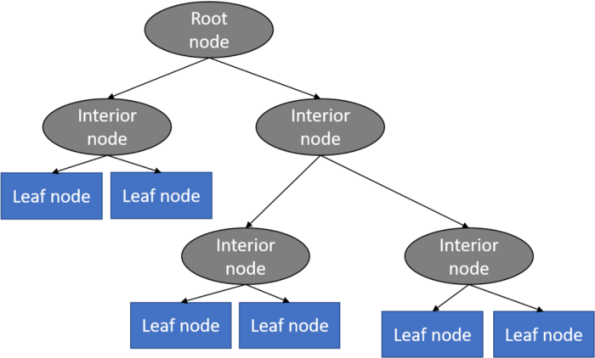
To get better accuracy, we tried regularization technique like **L1 (Lasso)** and **L2 (Ridge)** regularization. However, we found the similar results in all the models, for **L1** **r2 score** is **0.77** and **0.78** for train and test dataset respectively, for **L2** **r2 score** is **0.77** and **0.78** same as L1 for train and test dataset respectively which is aligning with the fact that we use regularization to overcome overfitting. Since, this model doesn’t have overfitting, model perform almost same.

**10.2 Elastic Net Regression**

Elastic net linear regression uses the penalties from both the **lasso** and **ridge** techniques to regularize regression models. The technique combines both the [lasso](https://corporatefinanceinstitute.com/resources/knowledge/other/lasso/) and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models. I have also tried Elastic net regression and the results were same with the **r2 score** of **0.77** and **0.78** for train and test data respectively, that means our model is not overfitted.

**10.3 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.

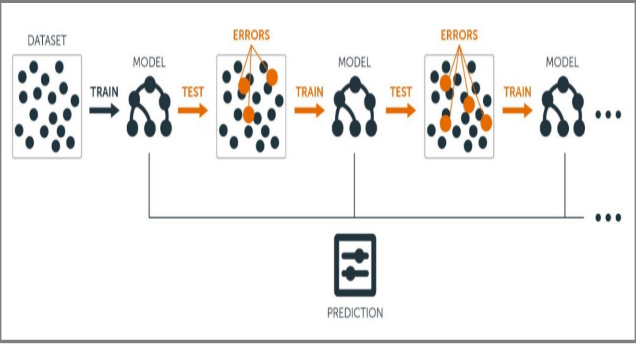
We found that our decision tree algorithms did not show any overfitting but has lesser r2 score with **0.65** and **0.62** for train and test dataset respectively.

**10.4 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. I have tuned our random forest regressor with **GridSearchCV** for best parameters for our algorithm and the result came with the r**2 score** of **0.78** and **0.76** for train and test data respectively.

**10.5 Gradient Boosting Regression**



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** and **0.92** 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.

**11. 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, **Gradient Boosting** with hypertuned via GridSearchCV model performed really well for this dataset with r2 score of 95% and 92%. Although, Linear, Lasso, ridge and Decision Tree has comparatively less r2\_score of round 77% 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 , functioning day yes and humidity.