Seoul Bike Sharing Demand Prediction

## Technical Document

## **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.

## **Business Objective**

Predicting the optimal number of bikes needed at any specific moment and day is a crucial and intricate business challenge. Striking the right balance is of paramount importance: too few bikes can lead to resource inefficiencies, including maintenance costs and the need for parking and security infrastructure, while an excess of bikes can result in financial losses. These losses encompass immediate revenue reduction due to a limited customer base and potential long-term impacts, such as the erosion of customer trust and loyalty.

Predicting bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers

Hence, it is imperative for bike rental enterprises to possess a robust demand estimation mechanism. This mechanism equips them to operate efficiently, maximize resource utilization, and ensure a seamless experience for their customers. Accurate demand forecasts are essential for informed decision-making and the overall success of bike-sharing businesses.

**Data Overview**

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

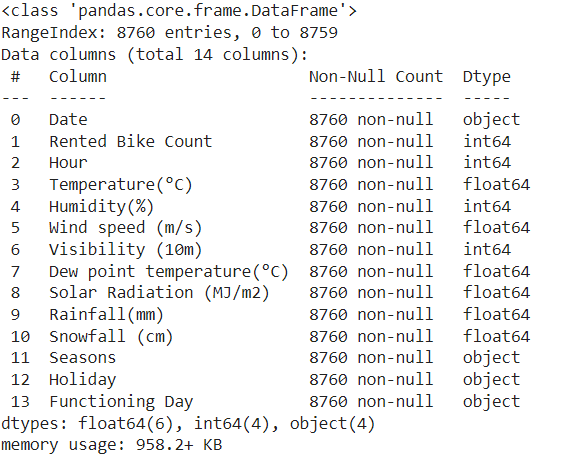
Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

CodeText

**Steps Involved :**

1. **Data Exploration:**

At the very first we started with loading the dataset and importing the required libraries. Next step was inspecting the data by looking at the top and bottom rows, checking its shape, columns and basic info of the dataset. Our data consisted of both continuous variables as well as categorical variables.



We checked for duplicate values and found that there were no duplicate values present in the dataset. We also checked for missing values and no null values were present.

We found out that Date column has datatype called as Object. So to transform Date feature we parsed it first and stored it in variable and then this variable is converted into %y-%m-%d format.

Then we extracted Date , Month and Day of the week from that Date column and dropped that column.

1. **Exploratory Data Analysis:**

After preparing the data we started exploratory data analysis by plotting graphs to understand the trends. Firstly we analysed Dependent Variable in which , we plotted distribution plot for ‘Rented Bike Count’ and we found out that this feature is Positively Skewed . So we Applied square-root transformation method to remove skewness and transform distribution into normal distribution.

Then we performed Univariate analysis on Numerical Columns And categorical columns where we plotted histogram plot for numerical columns and checked skewness for each feature. After that we checked outliers present in each column using boxplot. And Identified outliers present in columns.

Then we plotted countplot for each categorical column and noted the observation.

Then for Bivariate Analysis of Numerical Variables Firstly All Numerical Variables are plotted Against ‘Rented Bike Count’ with scatter plot and noted the observation. After that we plotted regression plot against all the numerical columns and observed linear relationship between ‘Rented Bike Count’ and all other numerical variable. Then all numerical variables are plotted against the ‘Hour’ column to check the effect of numerical columns on ‘Rented Bike column’ on hour ly basis.

Then we did Bi variate Analysis on Categorical Variables by plotting barplot of ‘Categorical variables;’ against ‘Rented Bike Count’ .

And then each Categorical variable is plotted against Rented Bike Count for hourly bases and found the relation between them.

1. **Data Preparation:**

Data Preparation consist of Feature Engineering and feature Selection. Here we first plotted heatmap where we found out the correlation between different columns.There was case of Multicollinearity between which was solved by VIF method.

We Found out that ‘Temperature’ ,‘Hour’, ‘dew point temperature ’ and ‘solar radiation are highly co related.

Then in feature creation we created a dummy column called weekends and fill the values with either 1s or 0s.

In LabelEncoding we converted categorical values like Seasons , Functional Day and Holiday into numerical form.

1. **Fitting Models:**

**We Employed 5 ML models named:**

1. **Simple Linear Regression:**

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line

Linear regression uses a linear approach to model the relationship between independent and dependent variables. In simple words its a best fit line drawn over the values of independent variables and dependent variable. In case of single variable, the formula is same as straight line equation having an intercept and slope.

Gradient descent is the process by which the algorithm tries to update the parameters using a loss function . Loss function is nothing but the diffence between the actual values and predicted values(aka error or residuals). There are different types of loss function but this is the simplest one. Loss function summed over all observation gives the cost functions. The role of gradient descent is to update the parameters till the cost function is minimized i.e., a global minima is reached. It uses a hyperparameter 'alpha' that gives a weightage to the cost function and decides on how big the steps to take. Alpha is called as the learning rate. It is always necesarry to keep an optimal value of alpha as high and low values of alpha might make the gradient descent overshoot or get stuck at a local minima. There are also some basic assumptions that must be fulfilled before implementing this algorithm. They are:

* No multicollinearity in the dataset.
* Independent variables should show linear relationship with dv.
* Residual mean should be 0 or close to 0.
* There should be no heteroscedasticity i.e., variance should be constant along the line of best fit.

1. **Lasso Regression**

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. It uses the Linear regression model with L1 regularization.

1. **Ridge Regression**

Ridge regression is a method of estimating the coefficients of regression models in scenarios where the independent variables are highly correlated. It uses the linear regression model with the L2 regularization method.

1. **Random Forest Regression**

Random Forest Algorithm widespread popularity stems from its user-friendly nature and adaptability, enabling it to tackle both classification and regression problems effectively. The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

1. **Gradient Boosting Regression**

Gradient Boosting is a popular boosting technique used in various machine learning libraries, such as XGBoost and LightGBM. It is known for its high accuracy and robustness, and it is particularly effective in both classification and regression tasks.

**Hyperparameter Tuning**

In Hyperparameter tuning we choose a set of optimal hyperparameters for a learning algorithm.A value of Hyperparameter is chosen before the learning phase starts.Hyperparameter tuning is the one of the most important and key part of machine learning which enhances our model. So first we see business problem and them we customize our hyperparameter grid according to that.We use GridSearchCV and RandomSearchCV for this.

RandomSearchCV is similar to GridSearchCV and both are being used for optimizing Model.Practically they are similar in implementation. The main difference between the practical implementation of the two methods is that we can use n\_iter to specify how many pattern we want for sample and test.

**Model Performance :**

**Evaluation Metrics:**

**R2** shows how well terms (data points) fit a curve or line. **Adjusted R2** also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.

**Adjusted R2** will always be less than or equal to R2.

**R2** assumes that every single variable explains the variation in the dependent variable.

The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

**MSE** is a risk function that allows us to calculate the average squared difference between a feature’s or variable’s predicted and actual value.

**RMSE** is an abbreviation for Root Mean Square Error, which is the square root of the value obtained from the Mean Square Error function.

**Feature Importance:**

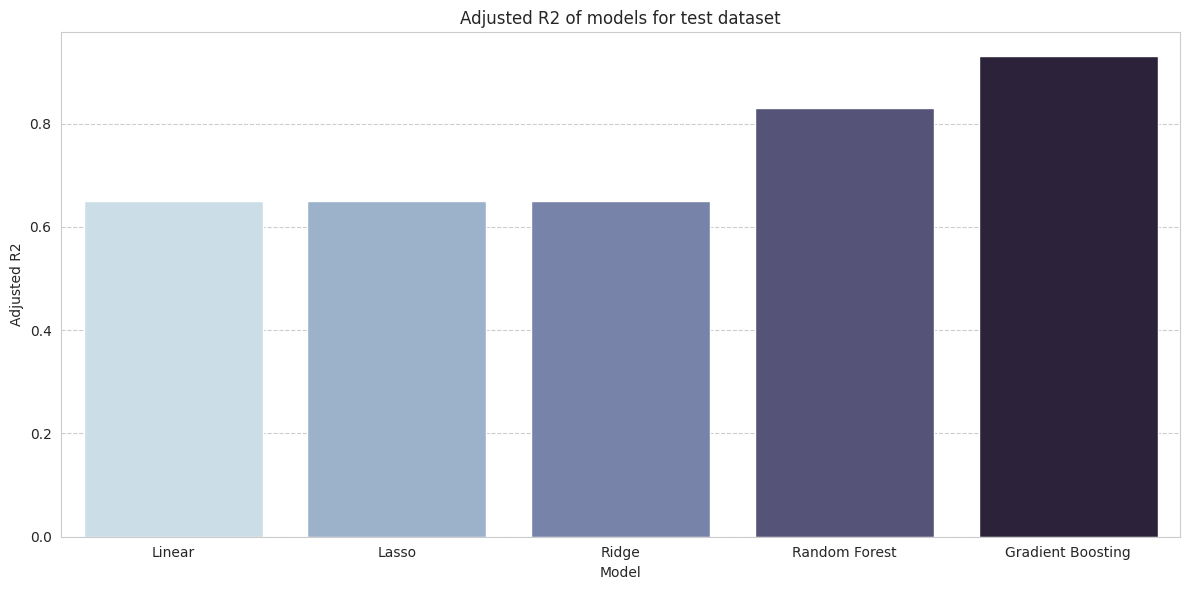
Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Feature importance scores play an important role in a predictive modeling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.

**Result:**

A screenshot of a table

Description automatically generated

**Graphical Representation:**



A graph showing different colored squares

Description automatically generated

**As R2 Square and Adjusted R2 Square values of Gradient Boosting are the highest and MSE ,MAE,RMSE values are lowest , this algorithm is good to go to solve this problem.**

**That’s why Gradient Boosting is the best algorithm between all tested Algorithms to solve this problem.**