**Bike Sharing Demand Prediction**

**Shubham Chandrakar**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

In Seoul city, South Korea, 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. The main challenge is to provide the city with a stable supply of rental bikes.

This project will help us understand what are the factors which affect demand for rental bikes. In this project we performed data wrangling, feature engineering, VIF analysis and exploratory data analysis (EDA) to understand the factors which affect demand for rental bikes in both negative and positive ways. After that, we built various supervised machine learning models to predict the rental bike demand.

***Keywords:machine learning, EDA, rental bike demand, VIF analysis.***

**1. Problem Statement**

In this project, the main objective is to predict the number of rental bikes required per hour to meet the demand. This would help to maintain the required amount of supply of rental bikes efficiently. Here, the dependent feature is rented\_bike\_count. We need to explore and analyze the data to understand various features to efficiently address the problem.

Various features provided in the dataset were as follows:

* Date : year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of he 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)

We need to ascertain the effect of these features on the determination of dependent variables.

**2. Introduction**

In Seoul ( a city in South Africa), rental bikes are currently 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. In this EDA project, following features are provided i.e. date, rented bike count, hour, temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, rainfall, seasons, holiday , functional day. Dependent feature here is rented bike count

### Our goal here is to build a predictive model, which could help us predict the estimated demand of rental bikes at any given hour.

## **3. Types of features**

* Numerical features
* Categorical features

### The numerical features are the features which are continuous in nature. In this database it is found that following are numerical features:

* rented bike count,
* temperature,
* humidity,
* wind speed,
* visibility,
* dew point temperature,
* solar radiation,
* rainfall

In Correlation analysis it is found that dew point temperature and temperature feature were ninety six percent correlated. Hence , the dew point feature was dropped.

The categorical features are the features they take values or levels. They are discrete in nature i.e. not continuous. Following are various categorical features in this database:

* date,
* hour,
* seasons,
* holiday
* functional day

It is found that the date feature itself is not relevant instead days, months and week days extracted as features from the date column were much relevant.

## **4. Factors affecting rental bike demand**

Following are the factors that affects rental bike demand:

1. Hour of the day (Hourly time period)
2. Time slot of the day (day,night,morning,evening)
3. Type of seasons.
4. Month of the Year
5. Temperature

## **Demand for rental bike is most at the peak hour**

It can be seen that bike demand rises after 5 AM and peaks at 8 AM, then again rises after 2 PM and peaks at 5PM then demand remain significantly above average demand 6PM and 11PM

That means in this 11 hours of a day bike demand is most.

## **Demand for rental bikes depends upon time slot**

We have divided 24 hours into four time slot: morning, afternoon, evening, night

Bikes are mostly rented at night time.Second most time slot at which bike is rented.At afternoon, least no of bikes are rented.

## **Demand for rental bikes depends on season**

It can be seen that demand of bike is in following order: Summer > Autumn > Spring > Winter

## **Demand for rental bikes is more on peak months**

It is found that Rental bike demand is low in January, February and December.

Rental bike demand is very high between May to August. So these months can be considered as peak months.

## **Demand for rental bikes is directly proportional to rise in temperature**

It can be seen that the demand for bikes increases with increase in temperature. When temperature is high demand for rental bikes is more during evening.When temperature is low demand for rental bikes ,during morning and night.

**6. Steps involved:**

* **Data Wrangling**

After loading the dataset we performed this method by comparing our dependent variable that is rented\_bike\_count with other independent variables. In these step we performed these steps:

1. Checking duplicate row
2. Checking Null/Nan Values
3. Outlier Removal
4. Feature engineering
5. Dependent variable distribution check.
6. Correlation analysis
7. VIF analysis

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is rented\_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.

* **Encoding of categorical columns**

We used pd.get\_dummies to encode our categorical features because categorical features that are in string format cannot be understood by the machine and need to be converted to numerical format.

* **Log transformation of our dependent variable**

Log transformation of dependent variable was done to normalize the positive skewed distribution of dependent variable.

* **Splitting our dataset into test and train set**

First we separated our data into dependent (y) and independent variables (X). After that,we split our dataset into a test and train set in ratio 4:1.

* **Standardization of features**

MinMaxScaler() is used to standardize the features. Our aim of this step was to scale our data into uniform values that 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 modeling we tried various regression algorithms like:

1. **Logistic Regression**
2. **Lasso Linear Regression**
3. **Random Forest Regressor**
4. **XGBoost Regressor**

* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.

Method used for hyper parameter tuning : GridSearchCV

It is found that hyperparameter tuning can identify the best parameter and it results in boosting accuracy over the baseline model.

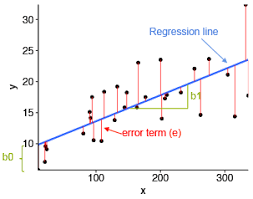
**7. Algorithms:**

**7.1 Linear Regression:**

In the most simple words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable..

**y = b0 +b1.x1 + b2.x2 + b3.x +......+bn.xn**

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

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

In the above diagram,

* x is our dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
* Black dots are the data points i.e the actual values.
* bo is the intercept which is 10 and b1 is the slope of the x variable.
* The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.

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.

.

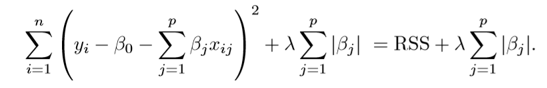
**7.2 Lasso Linear Regression:**

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique (will be discussed later in this article). It is used when we have more features because it automatically performs feature selection.

The word “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularization of data models and feature selection.

Cost function for LASSO regression:



In statistics, it is known as the L-1 norm.Lasso Regression tends to make coefficients to absolute zero whereas Ridge regression never sets the value of coefficient to absolute zero.

**7.3 Random Forest Regressor:**

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique. Design a specific question or data and get the source to determine the required data.

* Make sure the data is in an accessible format else convert it to the required format.
* Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
* Create a machine learning model
* Set the baseline model that you want to achieve
* Train the data machine learning model.
* Provide an insight into the model with test data
* Now compare the performance metrics of both the test data and the predicted data from the model.
* If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.
* At this stage, you interpret the data you have gained and report accordingly.

**7.4 XGBoost Regressor:**

Its is also known as *“Extreme Gradient Boosting”.*

It is called “an enhanced gradient boosting library” that makes use of a gradient boosting framework. Neural Network performs well when it comes to prediction problems that involve unstructured data like images and text.

But, decision tree-based algorithms are considered to be good performers when it comes to small to medium structured data or tabular data. XGboost is commonly used for supervised learning in machine learning. It was created by *PhD student Tianqi Chen, University of Washington.*

It carries out the gradient boosting decision tree algorithm. It has several different names like gradient boosting, gradient boosting machine, etc.

Boosting is nothing but ensemble techniques where previous model errors are resolved in the new models. These models are added straight until no other improvement is seen. One of the best examples of such an algorithm is the AdaBoost algorithm.

Gradient boosting is a method where the new models are created that computes the error in the previous model and then leftovers are added to make the final prediction.

It uses a gradient descent algorithm that is the reason it is called a “Gradient Boosting Algorithm”. Weather classification or regression methods are supported for both types of predictive modeling problems.

**8. Model performance:**

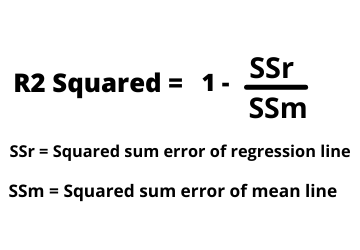
Model can be evaluated by various metrics such as:

**1. R2 Score**

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides. The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically R2 squared calculates how must regression line is better than the mean line.

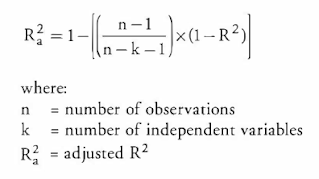
Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.

**2. Adjusted R2 Score**

The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

Hence, To control this situation Adjusted R Squared came into existence.

****

Now as K increases by adding some features so the denominator will decrease, n-1 will remain constant. R2 score will remain constant or will increase slightly so the complete answer will increase and when we subtract this from one then the resultant score will decrease. so this is the case when we add an irrelevant feature in the dataset.

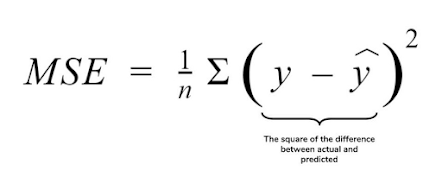
And if we add a relevant feature then the R2 score will increase and 1-R2 will decrease heavily and the denominator will also decrease so the complete term decreases, and on subtracting from one the score increases.

**3. MSE**

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

So, above we are finding the absolute difference and here we are finding the squared difference.

What actually the MSE represents? It represents the squared distance between actual and predicted values. We perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.



**3. RMSE**

As RMSE is clear by the name itself, that it is a simple square root of mean squared error.

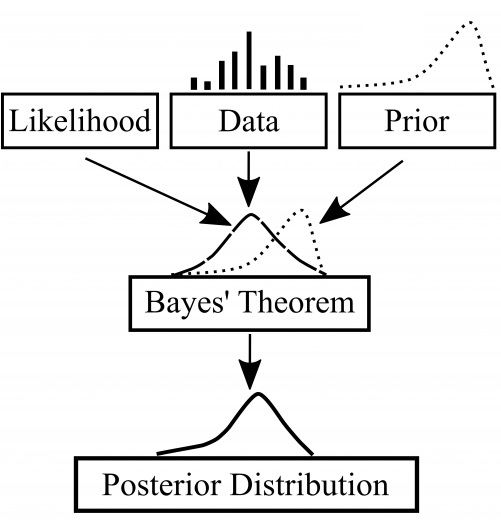
**9. Hyper parameter tuning:**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

1. **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.
2. **Randomized Search CV-** In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist’s control

# **Bayesian Optimization-** Bayesian Hyperparameter optimization is a very efficient and interesting way to find good hyperparameters. In this approach, in naive interpretation way is to use a support model to find the best hyperparameters.A hyperparameter optimization process based on a probabilistic model, often Gaussian Process, will be used to find data from data observed in the later distribution of the performance of the given models or set of tested hyperparameters.



As it is a Bayesian process at each iteration, the distribution of the model’s performance in relation to the hyperparameters used is evaluated and a new probability distribution is generated. With this distribution it is possible to make a more appropriate choice of the set of values that we will use so that our algorithm learns in the best possible way.

**8. Conclusion:**

It is found that demand for bike rises with rise in temperature.At night demand for rental bike is most.In summer season the demand for rental bike is most.In monthly period it is seen that rental bike demand is low on January, February and December and high between may to august.It can be seen that bike demand rises after 5 AM and peaks at 8 AM, then again rises after 2 PM and peaks at 5PM then demand remain significantly above average demand 6PM and 11PM.That means in this 11 hours of a day bike demand is most. *XG boost regression model* can predict rental bike demand with best accuracy of 94.14% accuracy.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya