**Bike Sharing Demand Prediction**

**Vikash Kumar**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

The attached dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information. Use the Regression to predict the count of bike count required at each hour for the stable supply of rental bikes.

I'm going to use this dataset for some EDA. And I'll delete the unnecessary columns and create some new ones. After all of this, I'll make future predictions using linear regression, decision trees, and random forests.

***Keywords: machine learning, linear regression, decision trees, random forests***

**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 rented bike count proactively. This would in turn help them in prediction of bike count required at each hour for the stable supply of rental bikes quickly and efficiently.

* 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 – speed of wind in m/s
* Visibility – visibility of road in 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)

**2. Introduction**

Currently, the bike-sharing scheme is well-received throughout the world. It is a shared bike service to individuals, which is free of charge and for a short-term basis at a minimal rate. Most bike-sharing systems permit people to borrow and return a bike from a bike station to another station that belongs to the same network. Bike-sharing gains a vast range of attention in recent years as part of initiatives to boost the use of cycle, improve the first mile/last mile link to other modes of transportation, and to minimize the negative effect of transport activities on the environment. Bike-sharing has significant impacts on establishing a larger cycling community, increasing the use of transportation, minimizing greenhouse gas emissions, enhancing public health and also traffic troubles.

Bike-sharing program progress is slow initially but after 1960 effective tracking tactics for bikes with improved technology are developed. During this decade, the development gave rise to the rapid spread of bike-sharing systems across several continents. South Korea is turning into a land of two wheels, as cycling facilities provide transit flexibility vehicle emission reductions, health advantages, low congestion fuel efficiency, and financial benefit for individuals and also the paths are extended in cities across the nation as well as in rural areas.

## **3. Number of rental bikes overall in relation to various characteristic values**

1. During the summer, a lot of bikes are rented as compared to other seasons.
2. The average number of rental bikes is high outside of holidays. That suggests a large proportion of people ride rental bikes to work.
3. The number of rented bikes grows as the temperature rises, as seen by the line graph. However, the number of rental bikes starts to decline beyond 30 degrees Celsius.
4. People travel to or from work primarily around morning 9 AM and evening 7 PM. The increased rental bike count during those hours may therefore be due to this. Additionally, more people prefer to hire bikes when they are leaving work than when they are going.
5. The most bikes are rented in the month of June. Additionally, the lowest number of bikes are rented in January and February.
6. The most bikes are rented in the month of June. Additionally, the lowest number of bikes are rented in January and February.
7. For Sundays the rented bikes are used less as compared to other days

## **4. Dropping of columns**

From correlation heatmap it is clear that only Temperature, Hour, Dew point Temperature and solar radiation are highly correlated with Rented bike. All other features are less correlated. Temperature has a very high correlation with dew point temperature and hence both has the same variation so dew point temperature can be eliminated.

High correlation between:

1. Temperature and Seasons
2. Temperature and Dew Temperature
3. Dewpoint Temperature and seasons
4. Humidity and dew point temperature

Dropped the columns from data frame with high correlation

**5. Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is Surge\_Pricing\_Type 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.

* **Null values Treatment**

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project inorder to get a better result.

* **Encoding of categorical columns**

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

* **Fitting different models**

For modelling we tried various classification algorithms like:

1. Linear Regression
2. Decision Tree
3. Random Forest

**6. Algorithms:**

1. **Linear Regression:**

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



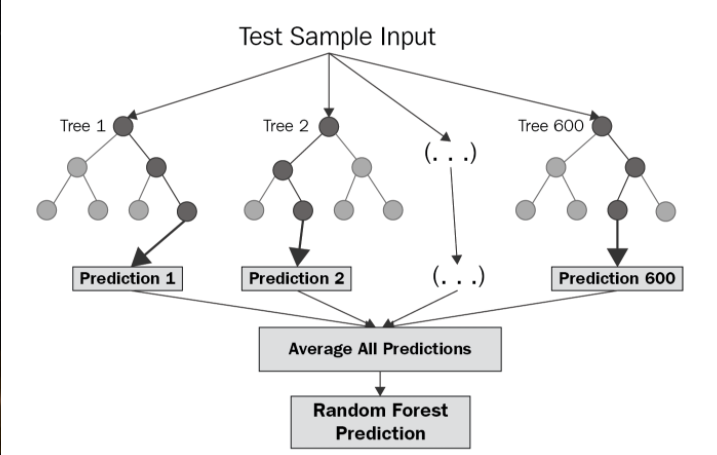
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.  
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

1. **Decision Tree**:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label

# **Random Forest Regression:**

**Random Forest Regression** is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

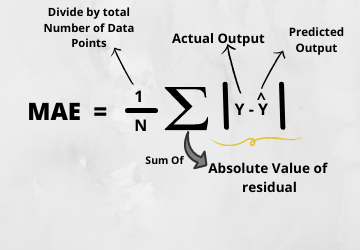


**7. Model performance:**

Model can be evaluated by various metrics such as:

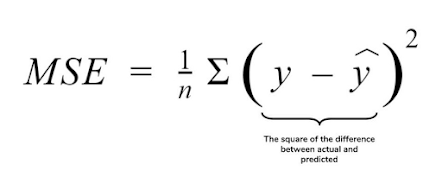
## **1. Mean Absolute Error(MAE)**

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.



## **2. Mean Squared Error(MSE)**

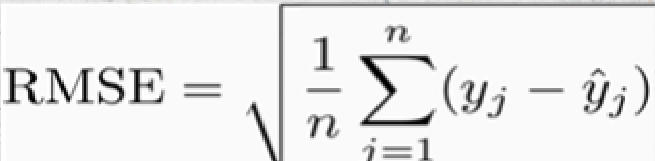
Mean squared error states that finding the squared difference between actual and predicted value.



## 

## **3. Root Mean Squared Error(RMSE)**

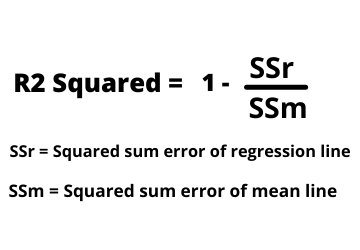
As RMSE is clear by the name itself, that it is a simple square root of mean squared error.



## **4. R Squared (R2)**

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.

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 a mean line.

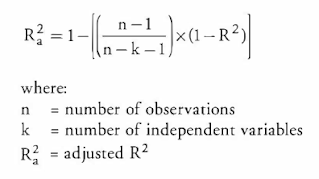


**5. Adjusted R Squared**

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.



**8. Conclusion:**

We have analysed Seoul city bike sharing dataset. Through analysis, we saw that in general the number of bike rents in 2018 was more than in 2017. The highest number of bike rents occur in summer while the least bike rents occur in winter. In daily basis, the trend of bike rents is almost similar with slight peaking demands on Thursday while drops on Sunday. In hourly basis, the bike counts peak in the afternoon (from 15.00 to 20.00). There are two peak occurrences, at 7.00 and at 17.00, which is most likely to be caused by workers going to office in the morning and going back home in the afternoon.

The hourly movement of bike counts seems to correlate with temperature, visibility, windspeed, and humidity. The bike counts peak in the afternoon (from 15.00 to 20.00) where temperature is the highest, with the most visibility, windspeed, and least humidity. There were days in a weekday when the bike sharing facility was not functioning. However, during public holidays, the facility was still operating.

Based on this analysis,

Future events can be predicted with 49.73% accuracy using linear regression.

With an accuracy rate of 85.69%, decision trees can forecast the future.

Random Forest has a 92.18% accuracy rate for future prediction.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya