**Seoul Bike Sharing Demand Prediction**

**Ajinkya Dakhale**

**Harshjyot Singh**

**Suvir Kapse**

**Data science trainees,**

**AlmaBetter, Bangalore**

**ABSTRACT:**

A bike-sharing system is a service in which bikes 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" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system. Rental Bike Sharing is the process by which bicycles are procured on several basis- hourly, weekly, membership-wise, etc.

One of the problems with commercial bike sharing programs is unequal riding patterns that result in unequal bicycle distribution at the end of the day. This means that unless the bikes are redistributed at night, there will be insufficient bikes at certain locations and too many bikes at other locations for the number of riders who wish to use them. With the data given to us of seoul we will analyse the demand hourly so that the docks have sufficient bikes to provide service.

**Introduction:**

Bike-sharing system is a shared micromobility service for short term bike rental. The service can be free of charge (e.g., paid by a city) or offered for a price. There are around 2000 bike-sharing services available around the world, mostly based in cities.The first bike-sharing system was introduced in 1965 in Amsterdam when a group named Provo left 50 bikes unlocked around the city for everyone to use. The second-generation arrangement in bike-sharing was a coin-deposit system. In this system, the users could unlock the bicycles with a coin that was refunded to them upon return of the bicycle.

Today, the most popular bike-sharing systems are automated docking stations (third generation) and dockless systems (fourth generation). In many cities, renting a bike has become an everyday digital service. For unlocking the bike, the users only need a subscription card or their smartphone. The users can leave the bikes to suitable docking stations or areas. This makes bike-sharing a convenient alternative to both public transportation and private cars.

Bike-sharing customers don’t depend on transportation routes and can use the service on demand. They can reach their destination without traffic jams or parking costs while contributing to a cleaner city environment as well as their health. For these benefits to have an effect, both the cities and service providers need to continually invest in infrastructure, system maintenance and innovation.

**Problem Statement**

Given dataset contains weather information   
(Temperature, Humidity, Wind speed,   
Visibility, Dewpoint, Solar radiation, Snowfall,   
Rainfall), the number of bikes rented per hour,   
and date information.

**The crucial goal of the machine learning project is to:**

Search factors and reasons which influence shortage of bike and time delay of availing bike on sharing service. This project aims to investigate the data to determine what variables are decisive in predicting the count of bikes required at each hour for stable supply of rental bikes. Hourly count of bikes for rent will also be predicted.

In short we aim to

Maximize: The availability of bikes to the customer.

Minimize: Minimise the time of waiting to get a bike on rent.

**DATASET PREPARATION:**

Bike sharing demand prediction dataset of a company from Seoul contains 14 features and 8760 observations of an entire year.

Date : year-month-day  
● Rented Bike count - Count of bikes  
rented at each hour  
● Hour - Hour of the day  
● Temperature-Temperature in Celsius  
● Humidity - %  
● Wind Speed - 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)

| **FEATURE NAME**  Date : year-month-day  Rented Bike Count  Hour  Temperature(**°C)**  Humidity (%)  Wind speed (m/s)  Visibility (10m)  Dew Point temperature (**°C)**  Solar Radiation (MJ/m2)  Rainfall (mm)  Snowfall(cm)  Seasons  Holiday  Functioning day | **TYPE**  Date  Int64  Int64  Float64  Int64  Float64  Int64  Float64  Float64  Float64  Float64  Object  Object  Object |
| --- | --- |

**DATA PREPROCESSING:**

A dataset may contain noise, missing values, and inconsistent data, thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

**DATA CLEANING:**

After loading the dataset, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after loading it into our system.

Irregularities are of different types of data.

* Missing Values: out dataset contains no missing values.
* Incorrect Format
* Incorrect Headers
* Anomalies/Outliers

**DATA DUPLICATION:**

It is very likely that a large dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset so it is essential to check for duplicate values in the dataset, EDA shows that our dataset has no duplicate values.

**HANDLING OUTLIERS:**

Data points known as outliers deviate from other observations for a variety of reasons. At the time of EDA phase, the common task is to find and work on these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

**FEATURE TRANSFORMATION:**

The distribution of the variables may also be improved by transforming the skewed variables. These could be square, square root, or logarithmic transformations. In our dataset Dependent variable i.e. Rented\_bike\_count having a moderate right skewed, to apply linear regression dependent features have to follow the normal distribution. Therefore, we use square root transformation.

**EXPLORATORY DATA ANALYSIS:**

Investigating a dataset to find patterns and anomalies (outliers) and developing hypotheses based on our knowledge of the dataset is the process of exploratory data analysis (EDA).

EDA entails producing summary statistics for the dataset's numerical data and developing various graphical representations to aid with data comprehension.

We use univariate Bivariate and multivariate analysis to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is the process of analysing the dataset that is available to find patterns, identify anomalies, test hypotheses, and validate presumptions using statistical metrics.

**UNIVARIATE ANALYSIS:**

If we analyse data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis examine one feature at a time. When we analyse a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the database.

**BIVARIATE ANALYSIS:**

We analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis. We have two types of data numerical as well as categorical type data. We analyse our dependent variable Rented\_Bike\_Data vs. numerical as well as categorical variable. We can analyse it in two different ways.

* Regression Plot
* Bar Plot

**MULTIVARIATE ANALYSIS:**

Multivariate analysis is used for analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a Heatmap, from a heatmap we can conclude that dew point temperature is highly correlated with temp, thus we decided to drop dew point temperature.

Chart, Polar chart, Histogram, Lollipop chart etc.

**ENCODING OF CATEGORICAL COLUMNS:**

We used Dummy Encoding to produce binary integers of 0 and 1 to encode our categorical features (Hour, Holiday, Seasons, Functioning, and Month) because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**ALGORITHMS:**

**1.  LINEAR REGRESSION:**

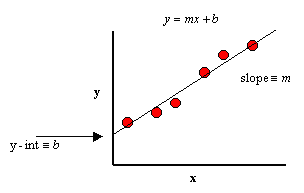
The main use of the supervised machine learning model known as linear regression is forecasting. Supervised machine learning models are ones that are built using training data, and their accuracy is then tested using a loss function.

When the dependent variable is continuous (e.g., bike count), linear regression is a good option. Linear regression is one of the simplest and most widely-used models. Linear regression assumes that the bike counts are linearly correlated to the features in the dataset such as temperature. It also assumes that attributes are independent of one another.

Linear regression fits a linear model with coefficients for each feature to minimize the mean square error in the linear regression approach; outliers can have a significant impact on the regression. Furthermore, linear regression may be prone to overfitting which will give low bias and high variance to overcome this issue we can use the regularisation technique (lasso ne ridge)

The function used in Linear Regression is given by:

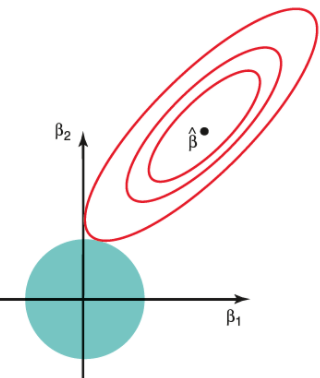
***y = mx + b***



Linear Regression

**2.  RIDGE REGRESSION:**

Ridge is used for Regularization when the coefficients are very high or overfitting. We use ridge regularization technique and this is used for better accuracy and there is no variable selection and both axes must have some value shrinking coefficients towards zero but rarely reach zero.



Ridge Regression

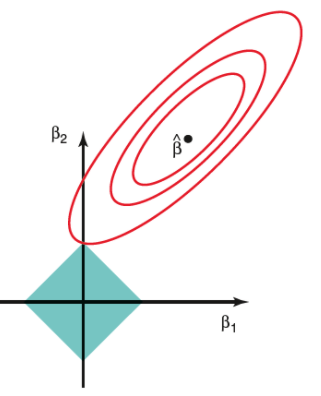
Any data that exhibits multicollinearity can be analysed using the model tuning technique known as ridge regression. This technique carries out L2 regularisation. Predicted values are far from the real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are substantial.

We decrease the model complexity that is the number of predictors. We can use forward or backward selection for this, but that way we won’t not be able to tell anything about the removed variables' effect on the response. It is possible to think of removing predictors as setting their coefficients to zero. Instead of forcing them to be exactly zero, it penalizes them if they are too far from zero, thus enforcing them to be small in a continuous way. In this manner, we maintain all of the model's variables while reducing model complexity.

**3.  LASSO REGRESSION:**

Lasso, or Least Absolute Shrinkage and Selection Operator it adds penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients, lasso penalizes the sum of their absolute values (L1 penalty). Because of this, many coefficients are precisely zeroes under lasso for high values of, which is never the case with ridge regression. The penalty terms are the only distinction between the ridge and lasso loss functions.

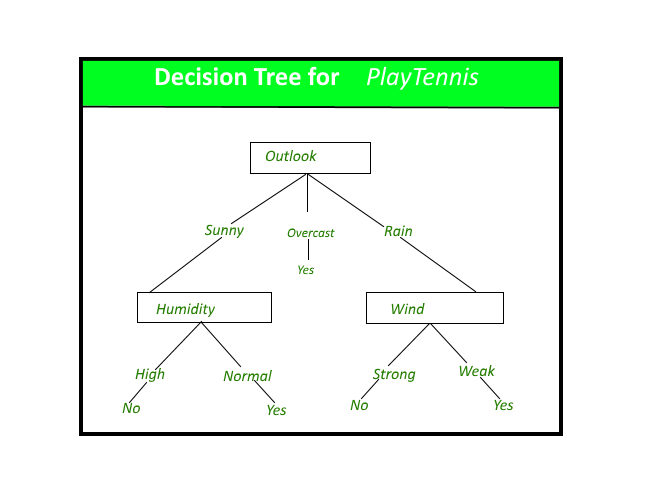
Under lasso, the loss is defined as:

https://lh4.googleusercontent.com/mP5jIfJDYJET0p84TmT5cJ5H5g7v6p3wTMsnOy63A9Lci3-VUmj03q-fUJgQ3zh2R4W2FTpeC9DkthpOzPQc4M-TRvOHUTcA2PR2D1g29tT74pdeZxe6PTLtEhzFdpG3z_e5L0YBDZc3HbxCKQ Lasso Regression

**4. DECISION TREE:**

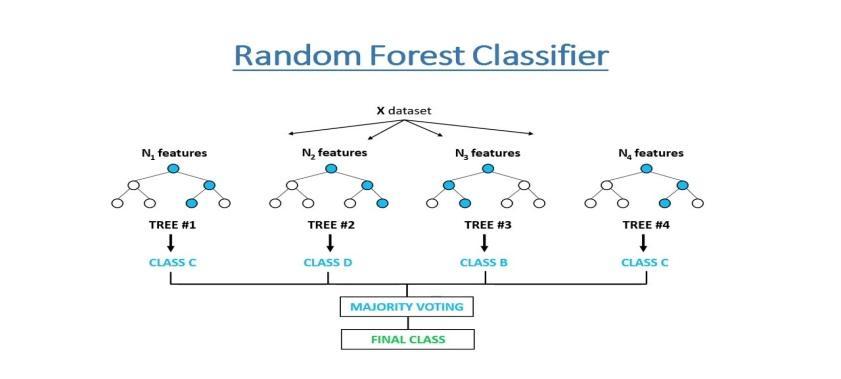
Decision trees are statistical models that measure a target value using a collection of binary rules. To make a decision between 2 nodes, decision trees use Attribute selection measure techniques to decide to split a node into more sub-nodes.

It is a graphical representation of all the possible solutions to a decision based on certain conditions. Classification trees and regression trees are two types of tree models where the objective variable can either accept a discrete set of values or continuous values (numbers).



**5. RANDOM FOREST:**

Random forest works by training a large number of decision trees and then calculating the mean prediction of the individual trees. The notion of random implies randomly created decision trees. Random decision trees are created on different subsets of the features and data points. For accurate predictions, random forest regressors can be optimized by hyperparameter tuning to ensure that the model does not depend too heavily on any single feature and that all potentially predictive features are considered equally. Also due to the previously mentioned random creation of decision trees, adding randomness prevents overfitting. Random forest regression provides a feature importance estimate. Using feature importance, the effort can aid in a deeper understanding of the solved problem and, in some cases, contribute to model improvements.



**CONCLUSIONS:**

Multiple regressors were compared and evaluated with Adj\_R2 and RMSE scores. With the best model being Random forest, achieved an accuracy of 92% and RMSE of 3.604, which indicates that machine learning indeed can predict bike counts required for a stable supply of rental bikes

The feature temperature accounts for the highest importance therefore people of Seoul are more likely to stay home during colder days than on warmer ones. Also feature hour has high importance, office hour is when huge demand of rented bike was observed.

**REFERENCES:**

1. [*https://towardsdatascience.com/*](https://towardsdatascience.com/)
2. GeeksforGeeks
3. Analytics Vidhya
4. [*https://stackoverflow.com/*](https://stackoverflow.com/)