Seoul Bike Sharing Prediction

Rahul Jha
(Data Science Trainee)
Almabetter,Banglore

1.Abstract

Public rental bike sharing has recently gained popularity due to its increased mobility, comfort, and environmental sustainability. People also like this idea because it avoids the need to invest a lot of money on a bike before riding it. People can rent bikes for a convenient, good, and affordable mode of transportation.

The data from Seoul Bike Sharing was used in the analysis. Temperature, precipitation, snowfall, visibility, solar radiation, and other weather factors are linked to the data for each hour.

In this study, we chose to analyse a dataset relevant to Rental Bike Demand from Seoul, South Korea, which contained climate variables such as Temperature, Humidity, Rainfall, Snowfall, and others.

2. Problem Statement

Apparently many big communities have introduced rental bikes to improve transportation comfort. It is critical to have the rental bike ready and accessible to the public at the appropriate moment, as this reduces waiting time. Eventually, delivering a steady supply of rental bikes to the city becomes a serious challenge. The critical component is predicting the number of bikes needed at each hour to ensure a steady supply of rental bikes.

3. Introduction

Bike sharing systems are a type of bike rental system in which the process of getting membership, renting, and returning the bike is automated through a network of sites around the city. People can hire a bike from one place and return it to another on an as-needed basis using these systems.

A typical bike sharing system includes various distinguishing features and characteristics, such as station-based bikes, the nearest station to the desired location, payment mechanism, membership, per-hour usage rates, and so on.

To restore equilibrium, many manual procedures, such as transferring motorcycles through trucks, automobiles, and even volunteers, are used. Data analysis approaches and research focusing on dynamic systems are used to supplement the knowledge foundation of using optimal rebalancing procedures.

Variables used in Analysis:
The columns used in the analysis are as follows:

- **Date: Date on which bike was rented (Format: year-month-day)**
- **Rented Bike Count: Count of bikes rented at each hour**
- **Hour: Hour of the day**

- Temperature: Temperature of the hour in Celsius
- Humidity: Humidity% of the hour
- Windspeed: Wind Speed of the hour in m/s
- Visibility: Visibility of the hour in units of 10m
- Dew Point Temperature: Dew Point Temperature of the hour in Celsius
- Solar Radiation: Solar Radiation of the hour in MJ/m2
- Rainfall: Rainfall of the hour in mm
- Snowfall: Snowfall of the hour in cm
- Seasons: 4 seasons (Winter, Spring, Summer and Autumn)
- Holiday: Whether the day is a Holiday or a Working Day
- Functional Day: Whether the day is functional for renting bikes or not.

4. Data Description:

The data description phase begins with initial data gathering and continues with efforts to familiarise the data. This step's actions include identifying data quality issues, discovering early insights into the data, and locating intriguing subsets to develop hypotheses from hidden information.

Data acquired from a hired bike provider firm in Seoul to be analysed includes consumer usage information. The information was obtained from a rental bike firm. It contains a total of 8760 rows and 14 columns. The majority of columns were about the hourly bike count for rent. The other column indicated how weather conditions affected bike count per hour.

- **Numerical Features:**

1.Rented Bike Count (Dependent Variable))
2. Temperature
3. Humidity
4. Windspeed
5. Humidity
6. Dew Point Temperature
7. Solar Radiation
8. Rainfall
9. Snowfall

- **Categorical Feature**

.1. Hour
2. Seasons
3. Holiday
4. Functioning Day

5.Data Preprocessing

- Initially, the shape of the dataset where the bikes rented count was 0 was checked, and then the shape of the dataset where the bike renting store was not functional was checked, and it was obvious that 0 bikes would be rented in the hour where the store was not functional, so the rows from the dataset where 0 bikes were rented in the hour were removed.

- Converted the ‘Date’ column to date-time datatype and extracted the year,month name and the day from it.

6.Exploratory Data Anallysis

- **The average number of bikes hired every weekday:-**

The least number of bikes were rented on Sunday, indicating that individuals tend to rent bikes on work days or for office purposes

- **Average number of bikes leased in each season:**

  Average number of bikes leased in each season:
  The highest number of bikes were hired during the summer season, while the lowest number of bikes were rented during the winter season, indicating that people prefer riding bikes during the summer season over the frigid winter season.

- **Average number of bikes rented in different years:**

  It was observed that less bikes were rented in 2017 and many more bikes were rented in 2018, so it can be said that the bike renting business was not much popular in 2017 but as the days went people developed the interest in bike renting and the business idea boomed.

- The most bikes are leased in the morning at 8 a.m. and in the evening at 6 p.m., implying that the bikes rented in the morning are for office and job purposes, while the bikes rented in the evening are for leisure and free time purposes. It is crucial to note that there is not a single hour when no bike is rented, indicating that the demand for renting a bike is enormous

- As the humidity rises, so does the quantity of rental bikes, resulting in a negative association.

- As the temperature rises, so does the number of bikes leased, resulting in a positive association.

- Snowfall does not establish a good correlation in determining the number of bikes getting rented.

- As the dew point temperature increases the number of bikes being rented also increases thus a positive correlation.

7. Data Prepration

- **HANDLING OUTLIERS:**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter 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.
There are two types of outliers:

• UNIVARIATE OUTLIERS:
Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

• MULTIVARIATE OUTLIERS:
While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

• MEASURES OF CENTRAL TENDENCY:
The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

- **MEASURES OF DISPERSION:**

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability.

- **CORRELATION AMONG VARIABLES:**

Correlation is the statistical approach that evaluates the link and explains if and how strongly two variables are related to one another. Correlation provides solutions to queries such as how one variable varies in relation to another. If it does alter, how much and how quickly?
Furthermore, if the relationship between those factors is strong enough, we may make predictions about future behaviour.

8.Algorithm

• Linear Regression

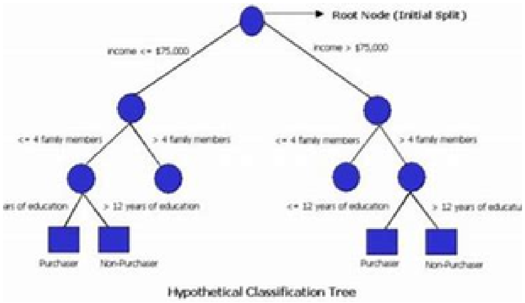
Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning models are those where we use the training data to build the model and then test the accuracy of the model using the loss function.Linear regression is one of the most widely known time series forecasting techniques which is used for predictive modelling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

• SVM

Support Vector Regression uses the same principle of Support Vector Machines. In other words, the approach of using SVMs to solve regression problems is called Support Vector Regression or SVR.

• Decision Tree Regressor

The decision tree is the most powerful and widely used classification and prediction tool. A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label. A tree can be "learned" by subdividing the source set depending on an attribute value test. This method is performed recursively on each derived subset, which is known as recursive partitioning.
As illustrated in the diagram above, testing the attribute defined by this node and then proceeding along the tree branch according to the value of the property.
This method is then repeated for the new node-rooted subtree.



• Random Forest Regressor

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects
the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

- ## Gradient Boosting Regressor

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term ‘gradient’ is related here.
Gradient boosting re-defines boosting as a numerical optimisation problem where the objective is to minimise the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimisation algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

9.CONCLUSIONS:

Bicycle sharing systems can be the new boom in India, with use of various prediction models the ease of operations will be increased. The four algorithms are applied on the bike share dataset for predicting the count of bicycles that will be rented per hour.
We got some good results and accuracy with random forest. The accuracy and performance has been compared between the models using Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), R2 and Adjusted R2. If these systems include the use of analytics the probability of building a successful system will increase

- **People tend to rent and ride more bikes on weekdays (working days) as compared to weekends (holidays).**
- **People prefer riding bikes in hot weather and in hot seasons.**
- **The bike renting business is being loved by the people and the business is prospering and growing by the time.**
- **Need of maximum bikes for renting can be seen at the morning time of 8am and the evening time of 6pm.**
- **As the temperature, solar radiation, dew point temperature and visibility increases the demand for renting bikes also increases.As the humidity decreases the demand for renting bikes increases.**
- **Demand for renting bikes is more when there is no rainfall and snowfall.**
- **The training and testing R2 Score obtained for Random Forest Regressor after hyperparameter tuning is 0.96 and 0.89 respectively.**
- **The training and testing R2 Score obtained for Gradient Boosting Regressor after hyperparameter tuning is 0.99 and 0.91 respectively.**
- **The training and testing R2 Score obtained for XGBoost Regressor after hyperparameter tuning is 0.96 and 0.84 respectively.**
- **The most important features for our model training were found out to be Temperature and Hour.**