**ABSTRACT:**

A bike-sharing system is a service that provides bicycles for short-term, shared usage to individuals for a fee or for free. Numerous bike share systems permit someone to rent a bicycle from a "dock" that is typically computer-controlled; the user enters payment information, and the system unlocks the bicycle. This bicycle may then be returned to a different dock within the same system. Rental Bike Sharing is the process of acquiring bicycles on various terms, such as hourly, weekly, membership-based, etc. Climate change, catastrophic natural disasters, ozone layer depletion, and other environmental abnormalities have resulted from this phenomenon as a result of the global attempt to reduce the carbon footprint. In our study, we selected to analyse a dataset relating to Rental Bike Demand in the South Korean city of Seoul, which included climate variables such as Temperature, Humidity, Rainfall, Snowfall, and Dew Point Temperature, among others. For the provided raw data, a thorough pre-processing was performed initially, followed by a regression analysis. In part, our linear model was able to explain the variables influencing the hourly demand for bicycle rentals.

*Keywords- Bike sharing demand prediction, exploratory data analysis, feature engineering, Retention, Higher Subscriber Base, Telecommunication, Data mining*

**INTRODUCTION:**

Recent studies predict that more than 60 percent of the world's population will reside in cities by 2050, compared to 50 percent in the present. Some nations around the world are implementing virtuous scenarios, which provide transportation at a reasonable price and reduce carbon emissions. In contrast, other cities are well behind. Typically, urban mobility accounts for 64% of the total kilometres travelled around the globe. It should be modelled and supplanted by intermodality and networked self-driving vehicles, which also provide a sustainable way of transportation. Mobility on Demand systems play a crucial role in expanding the supply of automobiles by increasing their idle time and number.

Bike-sharing MOD systems already play a significant role in short commutes and as 'last mile' mobility resources for inter-modal journeys in a number of locations. The layout of the station design, the fleet size and capacity of the station, the ability to identify damaged, stolen, or lost bikes, pricing, the observation of traffic and customer activity in order to encourage positive customer behaviour, the use of marketing campaigns, and other difficulties are common in the maintenance, design, and management of bike-sharing systems. System balance is the most difficult task: During the day, some stations are likely to be packed with bike traffic while others are likely to be empty, impeding pick-up and drop-off, respectively. So, to restore the equilibrium, a variety of manual procedures, such as relocating bicycles via trucks, cars, and even volunteers, are implemented. Dynamic systems are the focus of data analysis techniques and studies, and optimising techniques are used to supplement the acquired knowledge.

Today, bike-sharing systems are proliferating in cities throughout the world. To accomplish a short journey more quickly, renting a bicycle is preferable to walking. In contrast to driving, it is also relaxing and environmentally friendly.

**PROBLEM STATEMENT:**

Maximize: The availability of bikes to the customer.

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

**The primary objective of the research is to identify the variables and causes that contribute to the lack of bicycles and the delay in their availability for rent. This article will analyse the provided data to find which characteristics, if any, are connected with customer attrition. In addition, an hourly estimate of bicycles for rent will be made.**

**DATA DESCRIPTION:**

The first data gathering phase is followed by actions to become familiar with the data during the data description phase. This process involves identifying issues with data quality, gaining preliminary understanding of the data, and spotting interesting subgroups to generate hypotheses from hidden information. Details about customers' usage are included in the data that is gathered from a Seoul-based company that rents out bicycles in order to be studied. The information was obtained from the provider of hired bikes. There are 8760 rows and 14 columns. The majority of pieces discussed rental bikes by the hour. The other column showed how the weather affected the number of bikes driven

**DATASET PREPARATION:**

The rental bike company from Seoul's bike sharing demand prediction dataset has 14 attributes and 8760 observations covering the entire year.

I.e. from 1.12.2017 to 31.11.2018. Below Table shows the data features.

**Data-set description**

|  |  |
| --- | --- |
| **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 |

**FEATURE BREAKDOWN:**

**Date**: The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, *we need to convert into date-time format.*

**Rented Bike Count**: The number of rental bikes per hour is our dependent variable, and we must predict it.

**Hour:** The time of day, ranging from 0 to 23. It has a digital time format.

**Temperature (°C):**  Temperature of the weather in Celsius and it varies from -17**°**C to 39.4**°**C.

**Humidity (%)**: Availability of Humidity in the air during the booking and ranges from 0 to 98%.

**Wind speed (m/s):** Speed of the wind while booking and ranges from 0 to 7.4m/s.

**Visibility (10m):** Visibility to the eyes during driving in “m” and ranges from 27m to 2000m.

**Dew point temperature (°C)**:Temperature

At the beginning of the dayand it ranges from -30.6**°**C to 27.2**°**C.

**Solar Radiation (MJ/m2):**  Sun contribution or solar radiation during ride booking which varies from 0 to 3.5 MJ/m2.

**Rainfall (mm):** The amount of rainfall during bike booking which ranges from 0 to 35mm.

**Snowfall (cm):** Amount of snowing in cm during the booking in cm and ranges from 0 to 8.8 cm.

**Seasons:** Seasons of the year and total there are 4 distinct seasons I.e. summer, autumn, spring and winter.

**Holiday:** If the day is holiday period or not and there are 2 types of data that is holiday and no holiday

**Functioning Day:** If the day is a Functioning Day or not and it contains object data type yes and no.

**EXPLORATORY DATA ANALYSIS:**

To put EDA into basic terms, it implies making an effort to comprehend the provided data much better so that we can make sense of it. To explain the essential elements of each feature, such as the lowest and maximum value, average, standard deviation, and others, univariate frequency analysis was used. Additionally, it was used to generate a value distribution to spot outliers and missing numbers.

EDA is the process of analysing the dataset that is available to find patterns, identify anomalies, test hypotheses, and validate presumptions using statistical metrics. We will examine the procedures for carrying out excellent exploratory data analysis in statistical terms.The main purpose of EDA is to examine what the data can tell us beyond the formal modelling or hypothesis testing tasks carried out in Python. To do this, data visualisation is employed. A statistical model can be used or not.

**DATA ANALYSIS:**

One of the most important steps involving descriptive statistics and data analysis is this one. The primary tasks include summarising the data, identifying hidden correlations and relationships between the data, developing predictive models, assessing the models, and determining the accuracies. Summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models are a few of the methods used for data summarization.

**DATA SOURCING**

Finding and adding data to our system is known as data sourcing. There are primarily two methods for finding data.

1. Personal Data

2. Public Data

Data gathered from various sources needs to be properly maintained and sent to the appropriate information technology staff inside an organisation. As was previously noted, data may be gathered from a variety of objects and events utilising a variety of sensors and storage technologies.

**DATA PREPROCESSING:**

Pre-processing is necessary to improve the quality of the data and reduce the time needed for data mining because a dataset may contain noise, missing values, and inconsistent data.

**DATA CLEANING**

Data Cleaning comes next in the EDA process once Data Sourcing is finished. After entering the data into our system, it is crucial to remove any errors and tidy it up.

Different forms of data might be irregular.

* Missing Values
* Incorrect Format
* Incorrect Headers
* Anomalies/Outliers

**DATA TRANSFORMATION:**

To increase the effectiveness and accuracy of data mining, the process of normalising and aggregating the data is known as data transformation.

**DATA DEDUPLICATION:**

It is quite possible that duplicate rows exist in your dataset. To improve the dataset's quality, they must be eliminated.

**MISSING VALUES:**

For each consumer, a depiction of each service and item is available. Because not all consumers have the same subscription, missing values might happen. Some of them could offer a variety of services, while others might offer something entirely different. Additionally, there are other fields linked to system configurations that might contain null values, but there are none in the orange telecom data set.

Prior to performing any statistical analysis, we must deal with any missing data in the Dataset.

Three different categories of missing values predominate.

1. MCAR (Missing entirely at random): These values are independent of all other aspects.

2. MAR (Missing at Random): These values could be influenced by additional characteristics.

Missing not at random (MNAR): These values are missing for a specific cause.

**DROPPING MISSING VALUES:**

Simply removing missing values from our dataset is one way to deal with them. We are aware that the pandas library's isnull() and notnull() functions can be used to identify null values.

**HANDLING OUTLIERS:**

Data points known as outliers deviate from other observations for a variety of reasons. Finding and filtering these outliers is one of our frequent tasks during the EDA phase. The presence of such outliers can seriously impair statistical analysis, which is the main driver behind their detection and filtering.

Two categories of outliers exist:

**UNIVARIATE OUTLIERS:**

Data points with values outside the expected range for one variable are referred to as univariate outliers.

**MULTIVARIATE OUTLIERS:**

When you plot data with one variable alone, some values of that variable might not deviate significantly from the expected range, but when you plot the data with another variable, these values might deviate significantly.

**MEASURES OF CENTRAL TENDENCY:**

The average or mean value of datasets, which is intended to provide an ideal summarization of the entire set of measurements, is typically described by the central tendency measure. This value is a numeric expression that has some significance to the set. The mean, median, and mode are the three most frequently used metrics for analysing data's distribution frequency.

**MEASURES OF DISPERSION**:

The measure of dispersion, commonly referred to as a measure of variability, is the second category of descriptive statistics. The mean or average may not always be the best way to represent the data if we are closely examining the dataset because it will change when there are significant differences between the data. In this situation, a measure of dispersion will much more accurately depict the variability in a dataset.

In our dataset, the measurements of dispersion are provided through a variety of methodologies. The lowest and maximum values of the variables, range, kurtosis, and skewness are a few ways that are frequently utilised.

**STANDARDIZING VALUES:**

We must ensure that the numbers in the same column are on the same scale in order to execute data analysis on a collection of values. For instance, the entire column should be in meters/sec or miles/sec scale if the data comprises the peak speeds of several brands' automobiles.

**UNIVARIATE ANALYSIS:**

Univariate Analysis is the process of analysing data from a dataset over a single variable or column. In a univariate analysis, each attribute is examined separately. When we analyse a feature alone, we often pay little attention to other features in the dataset and are largely interested in the distribution of its values.

The simplest type of data analysis is called a univariate analysis. It implies that we undertake analysis on our data, which only contains one sort of variable. Univariate analysis is mostly used to acquire data, summarise that data, and identify trends among the results. It doesn't discuss the reasons for the ideals or their connections. Central tendency (that is, the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation) are two methodologies that are used to characterise the patterns discovered in univariate data.

**BIVARIATE ANALYSIS:**

Bivariate analysis is the process of analysing data by taking two variables or columns from a dataset into account.

**(a)Numeric-Numeric Analysis:**

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

* Scatter Plot
* Pair Plot
* Correlation Matrix

**( b) Numeric - Categorical Analysis:**

Numeric-category analysis involves examining one numerical variable and one categorical variable from a dataset. We mostly use mean, median, and box charts to analyse those.

**MULTIVARIATE ANALYSIS:**

Multivariate analysis is the 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 matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

**CORRELATION AMONG VARIABLES**:

Correlation is a statistical method that investigates the relationship between two variables and explains if and how strongly they are connected to one another. Correlation provides answers to issues like how one variable alters in connection to another. If so, how much or how strongly will it change? Furthermore, we may forecast future behaviour if the relationship between those factors is strong enough.

**GRAPHICAL REPRESENTATION OF THE RESULTS:**

In this stage, the dataset will be shown to the intended audience as graphs, summary tables, maps, and diagrams. This is a crucial stage since one of the main objectives of EDA is for the dataset's analysis to provide results that the business stakeholders can understand. Line charts, bar charts, scatter plots, area plots, and stacking plots are the majority of graphical analysis approaches. Chart types include pie, table, polar, histogram, and lollipops.

**ALGORITHMS:**

**1. LINEAR REGRESSION:**

Linear regression is a type of supervised machine learning model that is commonly used in predicting. Supervised machine learning models are ones in which we use training data to develop the model and then use the loss function to verify its correctness.

Linear regression is a well-known time series forecasting approach that is used in predictive modelling. It presupposes, as the name implies, a linear relationship between a collection of independent factors and the dependent variable (the variable of interest).

We're going to install a line

y = β0 + β1x

according to our data. In this case, x is referred to as the independent variable or predictor variable, while y is referred to as the dependent variable or response variable. Before we discuss how to perform the fit, let's take a closer look at the key quantities from the fit:

• β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis

• β0 is the intercept of the line.



**2. RIDGE REGRESSION:**

Ridge regression is a model tuning method used to analyse multicollinear data. This method performs L2 regularisation. The projected values are far from the real values when the multicollinearity problem is present, least-squares are impartial, and variances are significant.We have come to the conclusion that we want to reduce the model complexity, or the number of predictors. While we could do this using forward or backward selection, we would not be able to determine how the deleted variables affected the response. Removing predictors from the model can be seen as settings their coefficients to zero. Let's penalise them if they deviate too much from zero instead of forcing them to be exactly zero, forcing them to be small continuously. By doing this, we can maintain all of the model's variables while reducing model complexity. This, basically, is what Ridge Regression does.



**3. LASSO REGRESSION:**

We have come to the conclusion that we want to reduce the model complexity, or the number of predictors. While we could do this using forward or backward selection, we would not be able to determine how the deleted variables affected the response. Removing predictors from the model can be seen as settings their coefficients to zero. Let's penalise them if they deviate too much from zero instead of forcing them to be exactly zero, forcing them to be small continuously. By doing this, we can maintain all of the model's variables while reducing model complexity. This, basically, is what Ridge Regression does.

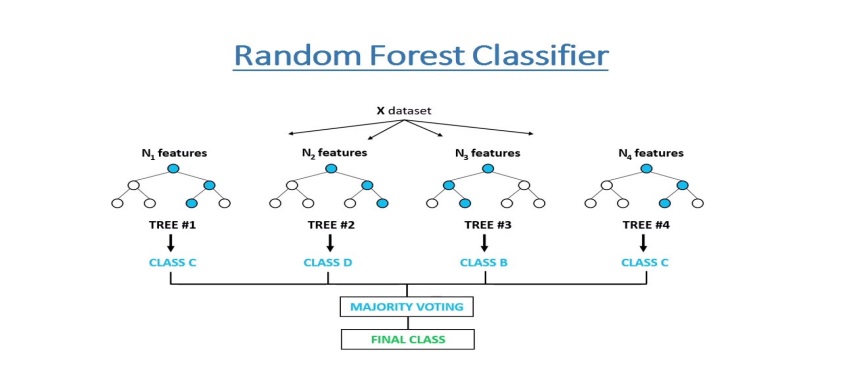


**4.DECISION TREE:**

Decision tree is the most powerful and popular technique for categorization and prediction. A decision tree is a flowchart-like tree structure where each internal node indicates a test on an attribute, each branch reflects an outcome of the test, and each leaf node (terminal node) has a class label. A tree can be “learned” by separating the source set into subsets based on an attribute value test. Using a technique known as recursive partitioning, this procedure is repeated on each derived subset. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, checking the attribute indicated by this node, and then progressing along the tree branch corresponding to the value of the attribute as illustrated in the above diagram. The subtree rooted at the new node is then subjected to the same procedure as before.



**5. RANDOM FOREST:**

 A decision tree bagging algorithm known as Random Forest builds several decision trees from a randomly chosen subset of the training set, gathers the labels from these subsets, and then averages the final prediction based on how often a given label has been correctly predicted across all of the decision trees.

**6. GRADIENT BOOSTING:**

Gradient and boosting are the two concepts that make up the umbrella phrase "gradient boosting." Gradient boosting is a boosting method, as we well know. Let's look at how this relates to the word "gradient."

By utilising gradient descent to add weak learners, gradient boosting redefines boosting as a numerical optimisation issue with the goal of minimising the loss function of the model. A local minimum of a differentiable function can be found using the first-order iterative optimization process known as gradient descent. Gradient boosting is a versatile approach that may be used for regression, multi-class classification, etc. because it is based on minimising a loss function.

**7. XGBoost**

XGBoost is an implementation of Gradient Boosted decision trees. Many Kaggle Competitions are dominated by XGBoost models. Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. Variables that the tree incorrectly predicted are given more weight before being placed into the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It can be used to solve problems including regression, classification, ranking, and custom prediction.

**8. Hyperparameter Tunning**

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. We can fit the model parameters by using existing data to train a model.

Hyperparameters, on the other hand, are a different class of parameter that cannot be directly learned through routine training. Usually, they are fixed before to the start of the programme itself. These parameters describe crucial model characteristics including complexity and learning rate.

Model hyperparameters include, for instance:

L1 or L2 regularisation, which serves as the penalty in the Logistic Regression Classifier the speed at which a neural network learns.

The C and sigma support vector machine hyperparameters.

The letter K as in k-nearest neighbours.

This article's goal is to investigate several methods for fine-tuning hyperparameters for machine learning models. Models may have a large number of hyperparameters, and determining the ideal set of parameters can be approached as a search issue. The following are the top two methods for hyperparameter tuning:

* GridSearchCV
* RandomizedSearchCV

**CONCLUSIONS:**

With the use of various prediction models, the simplicity of operations will be enhanced, bicycle sharing systems might become India's next big thing. The four methods are used to anticipate the number of bicycles that will be rented every hour using the bike sharing dataset. With random forest, we had some success in terms of accuracy and outcomes. Using R2, Adjusted R2, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), the accuracy and performance of the various models have been compared. The likelihood of developing a successful system rises if these systems incorporate analytics.

When we compare the root mean squared error and mean absolute error of all the models, Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set. So, ﬁnally this model is best for predicting the bike rental count on daily basis.

**REFERENCES:**

* Data science for business : what you think about data mining
* https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/
* <https://www.geeksforgeeks.org/ml-linear-regression/>
* <https://intellipaat.com/blog/gradient-boosting-in-machine-learning/>
* <https://www.youtube.com/watch?v=bPrmA1SEN2k&list=PLZoTAELRMXVPBTrWtJkn3wWQxZkmTXGwe>
* https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide