**Technical Documentation**

**CAPSTONE EDA PROJECT**

**ON SEOUL BIKE SHARING DEMAND PREDICTION**

**ABSTRACT:**

A bike-sharing system is a service that makes bikes available to people for short-term, shared use that can be paid for or provided for free. Several bike share programmes enable users to pick up a bike from a "dock," which is typically computer-controlled and where they enter their payment details to have the bike unlocked. Thereafter, you can return this bike to a different system-affiliated dock. Bicycles are obtained through the rental bike sharing method on a variety of bases, including hourly, weekly, membership-based, etc. Due to efforts made worldwide to reduce carbon emissions, which have resulted in climate change, unheard-of natural disasters, ozone layer depletion, and other environmental oddities, the prevalence of this phenomenon has increased to significant proportions.In our study, we decided to analyse a dataset on the demand for rental bikes in Seoul, South Korea, which included environmental variables such as temperature, humidity, rain, snow, dew point temperature, and others. The hourly rental bike count is the regress and was used for the available raw data after a thorough pre-processing. Our linear model was able to partially explain the variables influencing the hourly demand for rental bikes.

Keywords: Bike sharing demand forecasting, data exploration, feature engineering, retention, increased subscriber base,Telecommunication,and data mining

**INTRODUCTION:**

Several studies predict that, compared to the current situation, more than 60% of the world's population would live in cities by the year 2050. Several nations all across the world are putting virtuous scenarios into practise, providing transportation at a reasonable cost and reducing carbon emissions. Some cities, on the other hand, are well behind schedule. 64% of all miles travelled worldwide are often in urban areas. It should be imitated and replaced by intermodal, networked self-driving cars that also offer environmentally friendly transportation. The supply of automobiles, as well as their numbers and idle time, are significantly increased via systems referred to as Mobility on Demand.

In many cities, bike-sharing MOD systems are already firmly occupying the effective role in quick commutes and as "last mile" mobility resources on multi-modal trips. Maintenance, design, and management of bike-sharing systems are dominated by a few problems, including station layout, fleet size, station capacity, finding broken, stolen, or lost bikes, pricing, traffic and customer activity monitoring to encourage moral customer behaviour, marketing through campaigns, etc. The most challenging task is system balancing: When it comes to pick-up and drop-off during the day, certain stations are probably overrun with bike traffic while leaving others empty.So, a number of manual approaches, like moving bikes past automobiles, trucks, and even by volunteers, are used to restore the equilibrium. Studies and data analysis approaches that concentrate on dynamic systems are used to supplement the body of information regarding the use of the best rebalancing policies.

Many cities throughout the world are now implementing bike-sharing programmes. Renting a bike is a faster option than walking for short distances. Moreover, it is eco-friendly and comfortable too compared to driving.

**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.

**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. It contains 14 columns and 8760 rows. The majority of pieces discussed rental bikes by the hour. The other column showed how the weather affected the number of bikes driven each hour.

**DATASET PREPARATION:**

The bike sharing demand prediction dataset from a rented bike provider company from Seoul contains 14 features and 8760 observations of a complete year I.e. from 1.12.2017 to 31.11.2018. Below Table shows the data features.

**DATA PREPARATION:**

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

Attribute Information:

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)

**FEATURE BREAKDOWN:**

**Date**: The day's date, formatted as DD/MM/YYYY for the 365 days from 01/12/2017 to 30/11/2018, needs to be changed to date-time format.

**Rented Bike Count**:We must forecast the number of rental bikes every hour, which is our dependent variable.

**Temperature (°C):** The range of the weather's temperature in Celsius is -17 to 39.4 degrees.

**Humidity (%)**:During the booking, the air's humidity is available and ranges from 0 to 98%.

**Wind speed (m/s):**The wind's velocity at the time of booking is between 0 and 7.4 m/s.

**Visibility (10m):**Driving visibility is measured in "m" and ranges from 27 to 2000 m.

**Dew point temperature (°C)**:The temperature is between -30.6°C and 27.2°C at the start of the day.

**Solar Radiation (MJ/m2):** When reserving a ride, the sun's contribution, or solar radiation, ranges from 0 to 3.5 MJ/m2.

**Rainfall (mm):**Rainfall during bike reservations, which ranged from 0 to 35mm.

**Snowfall (cm):**Snowfall during the booking, measured in centimetres, ranging from 0 to 8.8 cm.

**Seasons:**There are four distinct seasons that make up the year's seasons: summer, autumn, spring, and winter.

**Holiday:**There are two forms of statistics that indicate whether a given day is a holiday period: holiday and non-holiday.

**Functioning Day:**If the day has object data types and whether it is a functioning day, the answers are 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. Also, 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 this chapter.

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 duties include summarising the data, identifying hidden correlations and linkages between the data, constructing predictive models, assessing the models, and determining the accuracies. Summary tables, graphics, 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.

* Private Data
* Public Data

Data gathered from various sources needs to be properly maintained and distributed 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 Cleansing comes next in the EDA process after Data Sourcing is finished. After entering the data into our system, it is crucial to remove any errors and tidy it up.

Many 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 representation of each service and item is available. Because not all consumers have the same subscription, missing values could happen. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

Before to performing any statistical analysis, we must deal with any missing values in the Dataset.

Three different categories of missing values predominate.

* MCAR (Missing completely at random): These values do not depend on any other features.
* MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

* **DROPPING MISSING VALUES:**

Just removing missing values from our dataset is one method 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 jobs during the EDA process. The presence of such outliers can seriously impair statistical analysis, which is the fundamental driver for their detection and filtering.

Two categories of outliers exist:

* **UNIVARIATE OUTLIERS:**

Data points with values beyond the anticipated 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 predicted 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 summation of the complete 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 ideal way to describe 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 far more properly depict the variability in a dataset.

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

* **STANDARDISING 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 metres/sec or miles/sec scale if the data contains the top 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 independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

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 behind 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:**

Numerical-numeric analysis examines the two numerical variables in a dataset. There are three different ways we can analyse it.

* 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 primarily use mean, median, and box charts to analyse those.

* **MULTIVARIATE ANALYSIS:**

The analysis of three or more variables is referred to as multivariate analysis. As opposed to bivariate analysis, this enables us to examine correlations (i.e., how one variable changes in relation to another) and attempt to predict future behaviour more correctly.

Making a matrix scatter plot, also referred to as a pair plot, is a typical method of visualising multivariate data. Each pair of variables is plotted against one another in a matrix plot or pair plot. 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 related 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? Additionally, if there is a strong enough correlation between those characteristics, we can forecast how people will act in the future.

* **GRAPHICAL REPRESENTATION OF THE RESULTS:**

At this step, the dataset will be shown to the intended audience as graphs, summary tables, maps, and diagrams. This is also a crucial stage because one of the main objectives of EDA is that the dataset analysis results can be understood by the business stakeholders. Line charts, bar charts, scatter plots, area plots, and stacked plots make up the majority of graphical analysis approaches. Chart types include pie, table, polar, histogram, and lollipops.

**ALGORITHMS:**

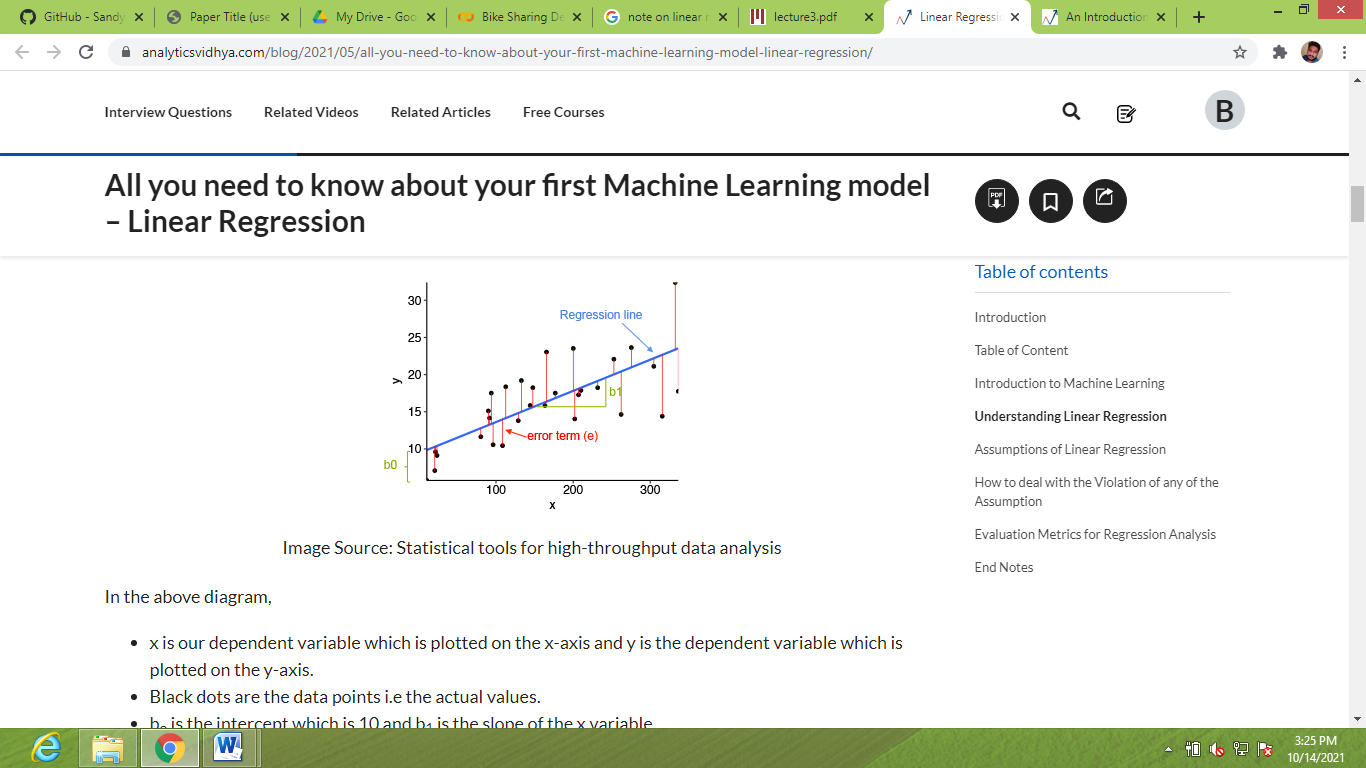
**LINEAR REGRESSION:**

A supervised machine learning model that is primarily utilised in predicting is linear regression. Supervised machine learning models are ones that are constructed using training data and then tested for accuracy using a loss function.

One of the most well-known time series forecasting methods used in predictive modelling is linear regression. As implied by the name, it presumes that a group of independent variables have a linear connection with the dependent variable (the variable of interest).

We’re going to fit a line

y = β0 + β1x



to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable. Before we talk about how to do the fit, let’s take a closer look at the important 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.

**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 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. Predictors' coefficients can be thought of as being set to zero when they are removed from the model. Let's penalise them if they deviate too much from zero instead of forcing them to be exactly zero, forcing them to be little continuously. In this manner, we maintain all of the model's variables while reducing model complexity. In essence, this is what Ridge Regression accomplishes.



**LASSO REGRESSION:**

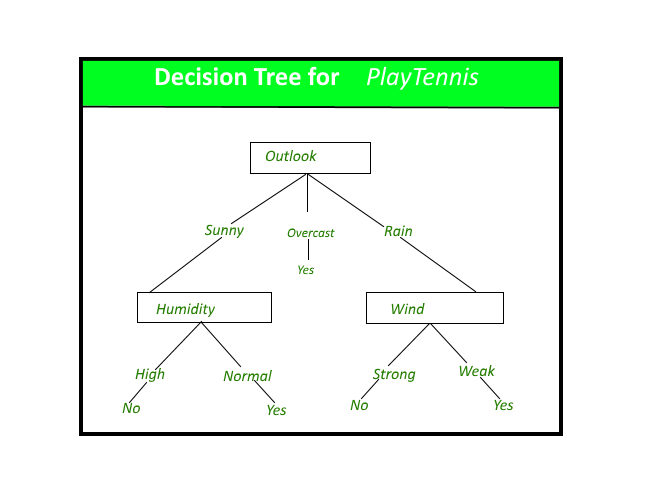
Ridge regression and Lasso, or Least Absolute Shrinkage and Selection Operator, are theoretically quite similar. Moreover, it adds a penalty for non-zero coefficients, but unlike ridge regression, which applies the so-called L2 penalty to the sum of squared coefficients, lasso applies the penalty to 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.

# **ElasticNet Regression**

L1 and L2 regularisation are combined in a weighted manner by the elastic net algorithm. As you can undoubtedly see, only the L1 wt option differs between LASSO and Ridge regression when using the same function. This argument establishes the relative importance of the partial slopes' L1-norm. If L1 wt = 1.0, the regularisation is pure L1, and if the regularisation is pure L2 (Ridge) (LASSO).

**DECISION TREE:**

The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label. By dividing the source set into subgroups based on an attribute value test, a tree can be "trained". It is known as recursive partitioning to repeat this operation on each derived subset. Decision trees categorise instances by arranging them in a tree from the root to a leaf node, which gives the instance's categorization.As demonstrated in the above diagram, to classify an instance, one tests the attribute given by the root node of the tree before continuing down the branch of the tree that corresponds to the attribute's value. The subtree rooted at the new node is then subjected to the same procedure once more.



**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.

**GRADIENT BOOSTING:**

The terms "gradient boosting" and "boosting" are two of its sub-terms. Gradient boosting is a boosting method, as we already know. See how the word "gradient" relates to this situation.

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 flexible technique that can be used for regression, multi-class classification, and other tasks because it is based on minimising a loss function.

**XGBoost:**

XGBoost is a distributed gradient boosting toolkit that has been tuned for quick and scalable machine learning model training. A number of weak models' predictions are combined using this ensemble learning technique to get a stronger prediction. Extreme Gradient Boosting, or XGBoost, is one of the most well-known and commonly used machine learning algorithms because it can handle enormous datasets and perform at the cutting edge in many machine learning tasks including classification and regression.

Its effective handling of missing values, which enables it to handle real-world data with missing values without requiring a lot of pre-processing, is one of the key characteristics of XGBoost. Moreover, XGBoost includes integrated parallel processing capability, allowing you to train models on huge datasets quickly.

Applications for XGBoost include click-through rate prediction, recommendation systems, and Kaggle competitions among others. Additionally, it is quite adaptable and enables performance optimization by allowing for fine-tuning of numerous model parameters.

**CONCLUSIONS:**

With the use of various prediction models, the simplicity of operations will be increased, bicycle sharing systems could 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 achieved some good accuracy and results. Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), R2, and Adjusted R2 have been used to compare the accuracy and performance of the various models. The likelihood of developing a successful system rises if these systems incorporate analytics.

**REFERENCES:**

* https://book.akij.net/eBooks/2018/May/5aef50939a868/Data\_Science\_for\_Bus.pdf
* Hands-On Exploratory Data Analysis with Python Perform EDA techniques to understand, summarise, and investigate your data by Suresh Kumar Mukhiya, Usman Ahmed (z-lib.org)

***Project by***

Ravi Kumar