Bike sharing Demand Prediction

Team Members

1)Chandan Kumar Raxit

2)Deepak Kumar Jena

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. This phenomenon has seen its stock rise to considerable levels due to a global effort towards reducing the carbon footprint, leading to climate change, unprecedented natural disasters, ozone layer depletion, and other environmental anomalies. In our project, we chose to analyse a dataset pertaining to Rental Bike Demand from South Korean city of Seoul, comprising of climatic variables like Temperature, Humidity, Rainfall, Snowfall, Dew Point Temperature, and others. For the available raw data, firstly, a through pre-processing was done after which a Here, hourly rental bike count is the regress and. To an extent, our linear model was able to explain the factors orchestrating the hourly demand of rental bike.

INTRODUCTION:

According to recent studies, it is expected that more than 60% of the population in the world tends to dwell in cities, which is higher than 50% of the present scenario. Some countries around the world are practising righteous scenarios, renderings mobility at a fair cost and reduced carbon discharge. On the contrary other cities are far behind in the track. Urban mobility usually fills 64% of the entire kilometres travelled in the world. It ought to be modelled and taken over by inter-modality and networked self-driving vehicles which also provides a sustainable means of mobility. Systems called Mobility on Demand has a vital part

in raising the vehicles' supply, increasing its idle time and numbers.

PROBLEM STATEMENT:

Maximize: The availability of bikes to the customer.

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

- **DATA DESCRIPTION:**

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Data which is collected from a rented bike provider company form Seoul to get analysed, involves usage details of customers from. The data was taken from rented bike Provider Company. It has 8760 rows and 14 columns. Most columns related to hourly bike count for rent. Other column was indicative of weather condition affecting bike count per hour.

- **DATASET PREPARATION:**
- The bike sharing demand prediction dataset from 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-set description**

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

The main goal of the project is to:

Finding factors and cause those influence shortage of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with customer churn, if any. Hourly count of bike for rent will also be predicted.

- **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**: *Number of rented bikes per hour which our dependent variable and we need to predict that*
- **Hour:** The hour of the day, starting from 0-23 it's in 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 day* and 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
- **EXPLORATORY DATA ANALYSIS:**
- If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. we using univariate frequency analysis was conducted 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 a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

- **DATA ANALYSIS:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

- **DATA SOURCING**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

- **DATA CLEANING**

After completing the Data Sourcing, 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 sourcing it into our system.

Irregularities are of different types of data.

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

- **MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of service 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

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

- **BIVARIATE ANALYSIS:**

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

- **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

- **CORRELATION AMONG VARIABLES**:

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

- **GRAPHICAL REPRESENTATION OF THE RESULTS:**

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

- **ALGORITHMS:**

1. LINEAR REGRESSION:

Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning

2. RIDGE REGRESSION:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization

3. LASSO REGRESSION:

Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the

sum of their absolute values (L1 penalty).

4.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.

5. RANDOM FOREST:

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.

6. GRADIENT BOOSTING:

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

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

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, summarize, and investigate your data by Suresh Kumar Mukhiya, Usman Ahmed (z-lib.org)
- https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c