Bike Sharing Demand Prediction

Debashish Das
Lucky Jain
Vivek Katolkar
Data science trainees,
AlmaBetter, Bangalore

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. Many 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. Then, 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 a global attempt to reduce carbon emissions, which has resulted in climate change, extraordinary natural disasters, ozone layer depletion, and other environmental oddities, this phenomenon has seen its stock climb to significant levels. 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 bikes.

Keywords- Bikesharing demand prediction, exploratory data analysis, feature engineering, Retention, Higher subscriber Base, Telecommunication, Data mining.

INTRODUCTION:

Recent estimates predict that, compared to today, more than 60% of the world's population would live in cities. This is an increase from the current 50%. Some nations all across the world are putting virtuous scenarios into practise, providing transportation at a reasonable cost and reducing carbon emissions. Other cities, on the other hand, are well behind schedule. 64 percent 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. Bike-sharing MOD systems are already firmly holding the effective part in short commuting and as ‘last mile’ mobility resources on inter-modal trips in several cities. Certain issues prevail in the maintenance, design, and management of bike-sharing systems: layout of the station design; fleet size and capacity of the station; detecting broken, lost, or theft bikes; pricing; monitoring of traffic and customer activities to promote behaviour virtuously; and marketing using campaigns etc. System balancing is the hardest endeavour: In the daytime, some stations are likely to be crowded with bike flow, while leaving other stations empty, which hampers pick-up and drop-off, respectively. So, to restore the balance, several manual techniques, like shifting bikes through trucks, cars and even by volunteers are employed. Data analysis techniques and studies focus on dynamic systems and optimisation methods

are utilised for complementing the knowledge base of employing optimum rebalancing policies.

Today, bike-sharing systems are blooming across more cities around the world. To complete a short trip renting a bike is a faster way when compared to walking. Moreover, it is eco-friendly and comfortable too compared to driving.

PROBLEM STATEMENT:

Increase: Increase the number of motorcycles that are available to the consumer.

Reduce: Reduce the amount of time spent waiting to rent a bike.

Finding the variables and causes that affect the lack of bikes and the wait times for renting a bike is the project's main objective. This paper analyses the data using the supplied information to see whether any variables are associated to customer churn. A anticipated hourly bike rental count will also be provided.

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

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

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 this chapter.

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

- **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 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 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

- **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

- **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

- **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.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. 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:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the is null() and not null() functions from the pandas library to determine 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:**

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. If we are analyzing the dataset closely, sometimes, the mean/average might not be the best representation of the

data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.
Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

- **STANDARDIZING VALUES:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies' cars, then the whole column should be either in meters/sec scale or miles/sec scale.

- **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. Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include 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).

- **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

b) Numeric - Categorical Analysis:

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

- **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. We may show both the distribution of single variables and the relationships between two variables using the pair plot.

- **CORRELATIONAMONG 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:

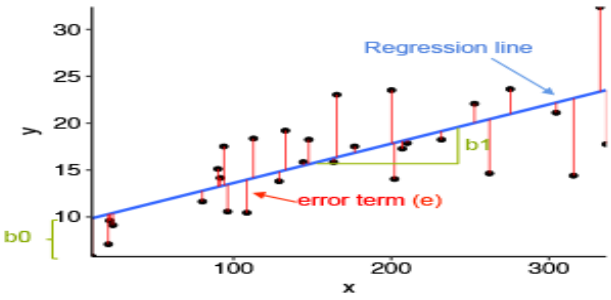
A supervised machine learning model that is primarily utilised in predicting is linear regression. Supervised machine learning models are ones that are constructed using training data and then tested for accuracy using a loss function.

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.



2. 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.

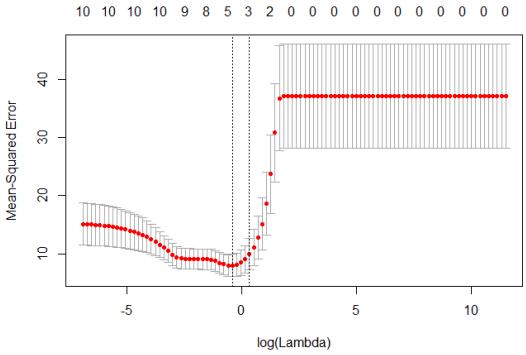
We have come to the conclusion that we want to reduce the model complexity, or the number of predictors. For this, we could choose forward or backward, but we wouldn't be able to tell. anything about the removed variables' effect on the response. Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does.

$$L_{ridge}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j-1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2.$$

3. LASSO REGRESSION:

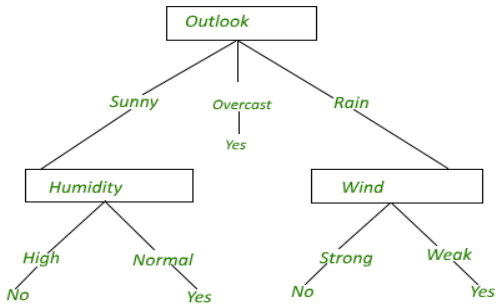
Ridge regression and Lasso, or Least Absolute Shrinkage and Selection Operator, are theoretically quite similar. Additionally, it adds a penalty for non-zero coefficients, but unlike ridge regression, which applies the so-called L2 penalty to the sum of squared coefficients, lasso applies the penalty 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. The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:

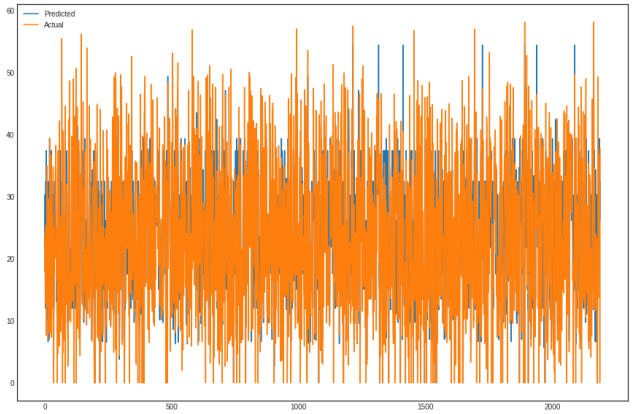
$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m}|\hat{\beta}_j|.$$



4.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. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.





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 most number of times a label has been predicted out of all.

	Feature	Feature Importance
0	Temperature	0.31
1	Humidity	0.16
34	Functioning_Day_Yes	0.15
10	Hour_4	0.03
4	Solar_Radiation	0.03
5	Rainfall	0.03
24	Hour_18	0.03
25	Hour_19	0.02
11	Hour_5	0.02
46	weekdays_weekend_1	0.02
9	Hour_3	0.02

6. GRADIENT BOOSTING:

Gradient and boosting are the two concepts that make up the umbrella term "gradient boosting." Gradient boosting is a boosting method, as we already 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 flexible technique that can be used for regression, multi-class classification, and other tasks because it is based on minimising a loss function.

	Feature	Feature Importance
0	Temperature	0.31
34	Functioning_Day_Yes	0.16
1	Humidity	0.15
4	Solar_Radiation	0.04
5	Rainfall	0.04

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Dicision tree regression	5.614	57.131	7.558	0.629	0.62
	5	Random forest regression	0.804	1.579	1.257	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95
Test set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Dicision tree regression	5.990	66.906	8.180	0.576	0.57
	5	Random forest regression	2.215	12.669	3.559	0.920	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92

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:

- Data science for business: what you think about data mining
- 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