**BIKE SHARING DEMAND PREDICTION**

**Pratyush Kumar Rath**

**Gaurav Singh**

**Bikesh Kuma Maharana**

**AlmaBetter, Bangalore**

**ABSTRACT:**

A bike sharing system is a service in which bikes are made available for shared use by individuals for a short period of time at a cost or for free. Many bike-sharing systems allow people to borrow a bike from a "dock," usually controlled by a computer, where the user enters payment information and the system unlocks it. This bike can then be returned to another dock that belongs to the same system. Rental Bike Sharing is a process where bikes are acquired on several basis – hourly, weekly, membership etc. This phenomenon has seen the inventory increase to a significant level due to global efforts to reduce the carbon footprint which leads to climate change. , unprecedented natural disasters, depletion of the ozone layer and other ecological anomalies. In our project, we decided to analyze a data set related to the demand for bicycle rentals from the South Korean city of Seoul, which includes climate variables such as temperature, humidity, precipitation, snowfall, dew point temperature, and more. The available raw data was first pre-processed, after which the number of bike rental hours is regressed here. Our linear model was able to explain to some extent the factors that drive the hourly demand for bicycle rentals.

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

**INTRODUCTION:**

According to recent studies, it is expected that more than 60% of the world's population tends to live in cities, which is more than 50% of the current scenario. Some countries around the world are practicing fair scenarios, providing affordable mobility and reducing carbon emissions. On the contrary, other cities are far behind. Urban mobility typically fills 64% of all kilometers driven in the world. It should be modeled and adopted by intermodal and networked self-driving vehicles that also provide a sustainable means of mobility.

Systems called Mobility on Demand play a vital role in increasing the supply of vehicles, extending their downtime and number. MOD bike sharing systems are already firmly established in an effective role in short commuting and as "last mile" means of mobility on intermodal routes in several cities. Certain issues prevail in the maintenance, design and management of bike sharing systems: station design layout; fleet size and station capacity; detection of broken, lost or stolen bikes; pricing; monitoring customer traffic and activities to promote virtuous behavior; and marketing through campaigns etc. Balancing the system is the most difficult endeavour: During the day some stations they are likely to be crowded with a stream of bikes while other stations remain empty, making pick-up and drop-off difficult. So, several manual techniques are used to restore the balance, such as shifting wheels between trucks, cars and even volunteers. Data analysis techniques and studies focus on dynamic systems and optimization methods are used to supplement the knowledge base on the use of optimal rebalancing policies.

Today, bike sharing systems are flourishing in multiple cities around the world. For a short trip, renting a bike is faster than walking. In addition, it is environmentally friendly and comfortable to drive.

**PROBLEM STATEMENT:**

Maximize: The availability of bikes to the customer.

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

**The main goal of the project is to search factors and causes that affect the lack of a bike and the time delay of a rental bike. Based on the data provided, the objective of this paper is to analyze the data and determine what variables, if any, correlate with churn. The hourly number of bikes to rent will also be predicted.**

**DATA DESCRIPTION:**

The data description phase begins with initial data collection and continues with data familiarization activities. The activity of this step is to identify data quality issues, discover first insights into the data, and detect interesting subsets to generate hypotheses from hidden information. The data that is collected from the Seoul bike rental company to be analyzed includes customer usage details. The data was taken from the bike rental company. It has 8760 rows and 14 columns. Most columns related to the hourly count of bicycles for rent. Another column gave the weather conditions affecting the number of laps per hour.

**DATASET PREPARATION:**

The bike sharing demand prediction dataset provided by bike renter 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**  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**: 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 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 explain EDA simply, it means trying to understand the given data much better in order to make some sense out of it. Using univariate frequency analysis, we described the key characteristics of each element including minimum and maximum value, mean, standard deviation, and more. It was also used to create a distribution of values and identify missing values and outliers.

EDA is the process of examining an available data set to discover patterns, detect anomalies, test hypotheses, and validate assumptions using statistical measurements. In this chapter, we'll discuss the steps involved in performing cutting-edge exploratory data analysis

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

From this EDA we have found that in the month of June there are more number of bike rented and in the month of January it is the least.Next we have created a new column as weekdays and weekend , by this we have found that in weekends more number of bike rented. By hourly calculation and analysis we can conclude that in weekdays the peak hours are 7am to 9am and 5pm to 7pm. It mat be because of the office timing. But in weekends in the evening time form 3pm the rent if bike increasing up to midnight. On seasonal analysis we found in summer most bike are rented while in winter least bikes are rented. In holiday also the average bike count is high as compare to non holiday.

**STEPS EVOLVED IN DATA CYCLE**

1. **DATA ANALYSIS:**
2. **DATA SOURCING**
3. **DATA PREPROCESSING:**
4. **DATA CLEANING**
5. **DATA TRANSFORMATION:**
6. **DATA DEDUPLICATION:**
7. **MISSING VALUES:**
8. **DROPPING MISSING VALUES:**
9. **HANDLING OUTLIERS:**

**ALGORITHMS:**

**1. LINEAR REGRESSION:**

Linear regression is a supervised machine learning model mostly used in forecasting. Supervised machine learning models are those where we use training data to build a model and then test the accuracy of the model using a loss function.

Linear regression is one of the most well-known time series forecasting techniques used for predictive modeling. As the name suggests, it assumes a linear relationship between a set of independent variables and a dependent variable (variable of interest).

EQUATION OF BEST FIT 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 perform 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

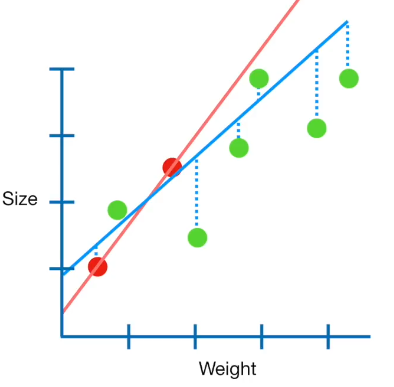
• β0 is the intersection of the line.



**2. RIDGE REGRESSION:**

Ridge regression is a model tuning method used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the problem of multicollinearity occurs, the least squares are unbiased and the variances are large, resulting in the predicted values ​​being far from the true values.

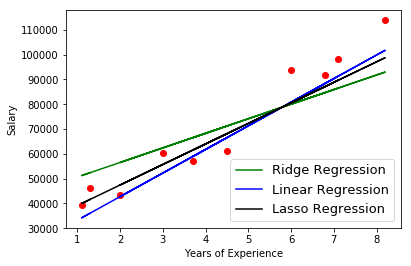
we came to the conclusion that we would like to reduce the complexity of the model, i.e. the number of predictors. We could use forward or backward selection to do this, but that way we wouldn't be able to say anything about the effect of the removed variables on the response. Removing predictors from the model can be thought of as setting their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus forcing them to be small all the time. This way we reduce the complexity of the model while keeping all the variables in the model. This is essentially what Ridge Regression does.



**3. LASSO REGRESSION:**

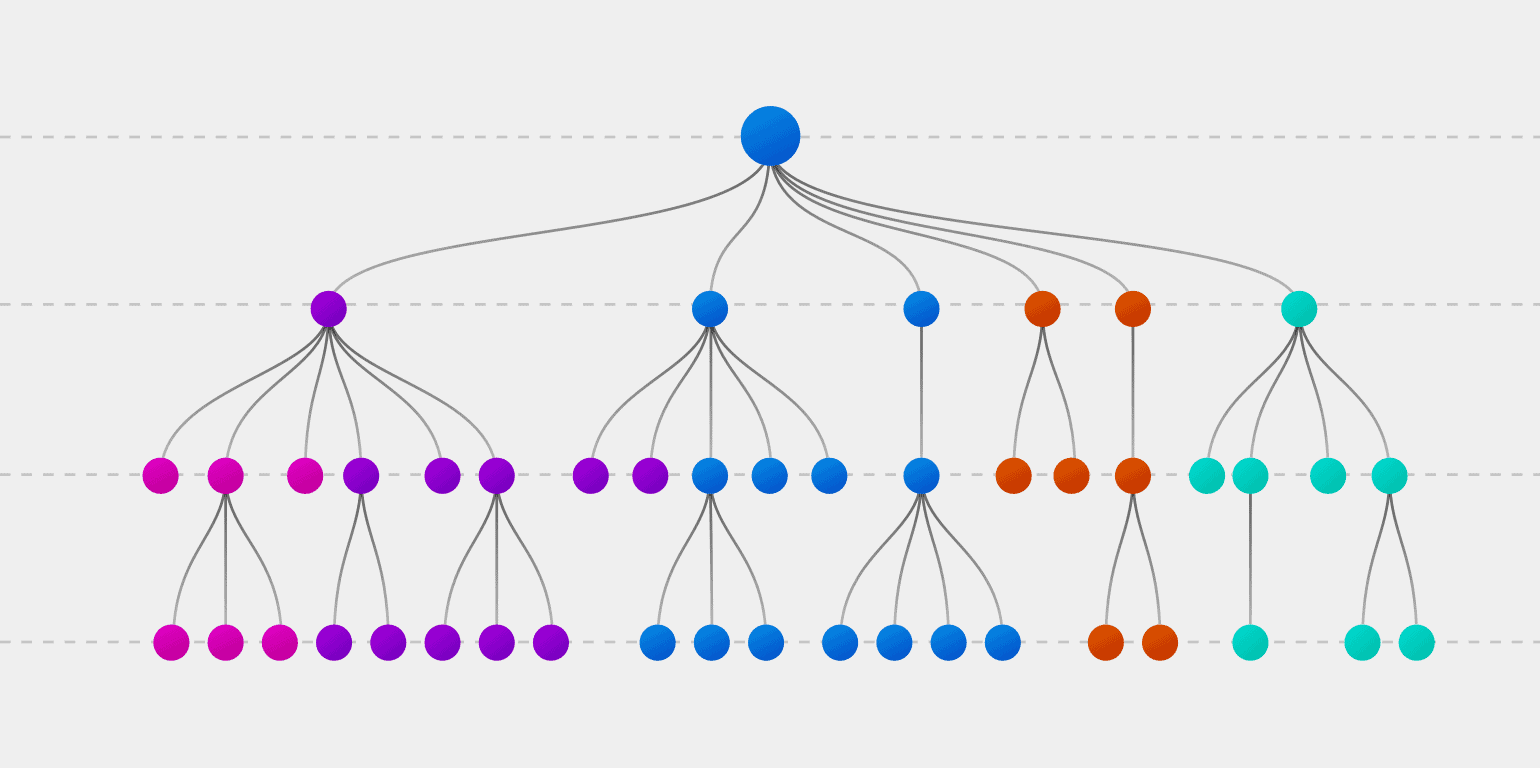
Lasso, or the least absolute shrinkage and selection operator, is conceptually quite similar to ridge regression. It also adds a penalty for nonzero coefficients, but unlike ridge regression, which penalizes the sum of the squares of the coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values ​​(the L1 penalty). As a result, for high values ​​of λ, many coefficients are exactly zero under the lasso, which is never the case for ridge regression.

The only difference in the comb and lasso loss functions is in the penalty conditions. Under the lasso, loss is defined as:

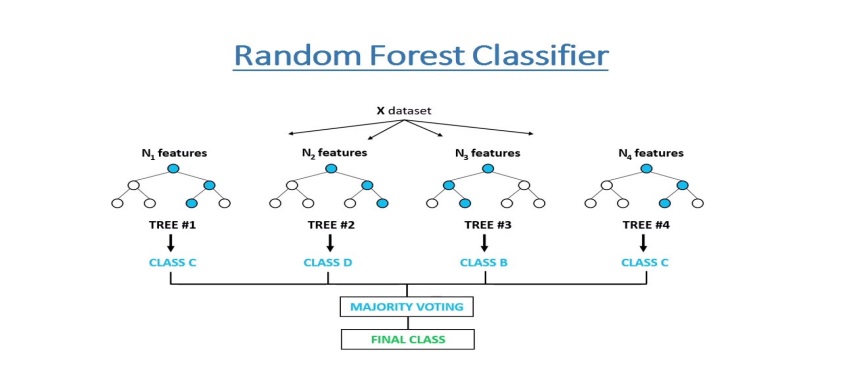


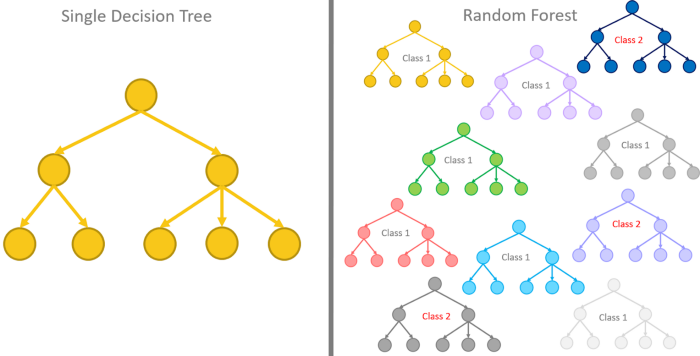
**4.DECISION TREE:**

A decision tree is the most powerful and popular tool for classification and prediction. A decision tree is a tree structure similar to a flowchart, where each internal node denotes a test on an attribute, each branch represents the result of the test, and each leaf node (terminal node) has a class label. A tree can be "learned" by dividing the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. Decision trees classify instances by ordering them in a tree from the root to some leaf node that provides the instance's classification. An instance is classified by starting at the root node of the tree, testing the attribute specified by that node, and then moving down the branch of the tree corresponding to the value of the attribute, as shown in the figure above. This process is then repeated for the subtree rooted at the new node.



**5. RANDOM FOREST:**

Random Forest is a wrapper type of decision tree algorithm that creates a series of decision trees from a randomly selected subset of the training set, collects labels from those subsets, and then averages the final prediction depending on how many times the label has been used. predicted of all.



**6. GRADIENT BOOSTING:**

The term gradient gain consists of two sub-terms, gradient and gain. We already know that gradient boosting is a boosting technique. Let's see how the term "gradient" relates here.

Gradient boosting redefines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak pupils using gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. Since gradient boosting is based on the minimization of 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:**

Bike sharing systems may be the new boom in India, with the use of various predictive models to streamline traffic. Four algorithms are applied to a bike-sharing dataset to predict the number of bikes that will be rented per hour. We got good results and accuracy with random forest. Accuracy and performance were compared between 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 likelihood of building a successful system increases

**REFERENCES:**

* Data science for business : what you think about data mining
* <https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf>
* <https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c>
* W3SCHOOLS
* GEEKS FOR GEEKS
* ALMA BETTER
* PANDAS DOCUMENTATION

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*