Bike Sharing Demand Prediction

Avanish Dixit
Vikas Chaudhary
Rushikesh Borude

Data Science trainees,
AlmaBetter, Bangalore

Abstract: - Bike sharing is an effective way to reduce the carbon emission as well as it is pocket friendly way to commute.

We selected a dataset relevant to Bike sharing Demand from Seoul, South Korea, which included somany features like Temperature, Humidity, Rainfall, Snowfall etc.. For the raw data that is available, First, we have done some EDA on the data, followed by the regress and is the hourly rental bike count. in response to an Our model was able to explain the factors to some extent. Coordinating the hourly rental bike demand**.**

Keywords: - *Data Mining, EDA, Machine learning, Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree, Random forest, Bike Sharing Demand Prediction.*

Problem Statement

Basically, we have to predict the count of the rental bikes. so that we can maintain a steady supply to the consumer.

different features presented in our dataset.

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind Speed - 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 – No Func (Non-Functional Hours), Fun (Functional hours)

INTRODUCTION.

The first bike-share programs began in 1960s Europe, but the concept did not take off worldwide until the mid- 2000s. In North America, they tend to be affiliated with municipal governments, though some programs, particularly in small college towns, center on university campuses. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis

The typical bike-share has several defining characteristics and features, including station-based bikes and payment systems, membership, and pass fees, and per- hour usage fees. Programs are generally intuitive enough for novice users to understand. And, despite some variation, the differences are usually small enough to prevent confusion when a regular user of one city's bike-share uses another city's program for the first time.

In addition to the environmental benefits, the sharing systems will impart healthier habits among the commuting public, who in the hustle of daily routine, often are unable to integrate optimum level of physical activity, which results in a barrage of ailments.
The primary objective was to build a superior statistical model to predict the number of rented bikes with the availability of data and understand the trends and factors affecting the rented bike count on a particular day.

Steps involved:

Exploratory Data Analysis

For our Seoul Bike Dataset, the Pre-Processing was performed on Colab Notebook. The CSV file was loaded using pd.read_csv () function.The DataFrame contains 8760 observations and 14 features.

Null values, outliers treatment:

Our dataset does not contain any null value. And checking the null data is checked using the isnull() function of python. Additionally, we use the sum() function for cross verifying.

After checking the null values we have checked for the outliers.

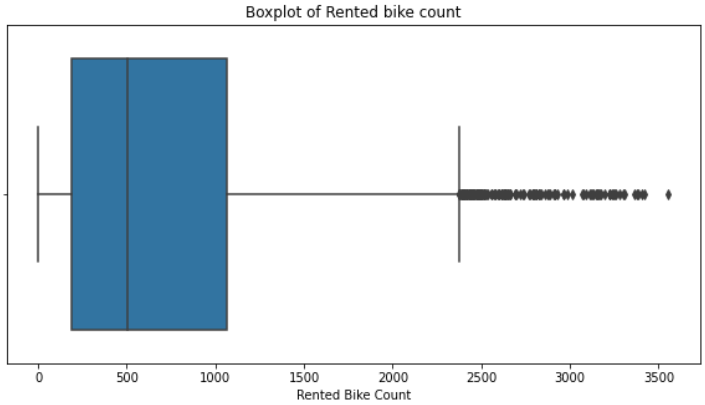


fig: box plot of dependent variable (outlier detection)

After the experiment, we can come to a conclusion that our dependent variable contains the outliers. so simply we did remove those rows with the help of drop() function.

Categorical variables- Seasons, Functioning Day, and Holiday- were converted coded into numerical features with the help of one hot encoding to fit our Linear Regression analysis..

For Categorical variables, we use barplot with the help of matplotlib.

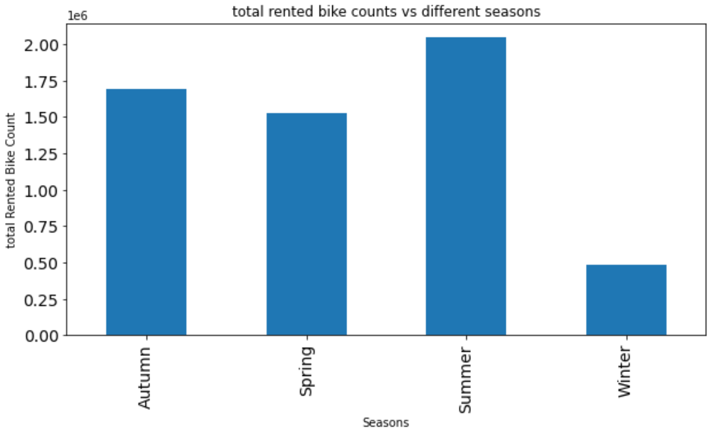


fig: different seasons vs total rented bike

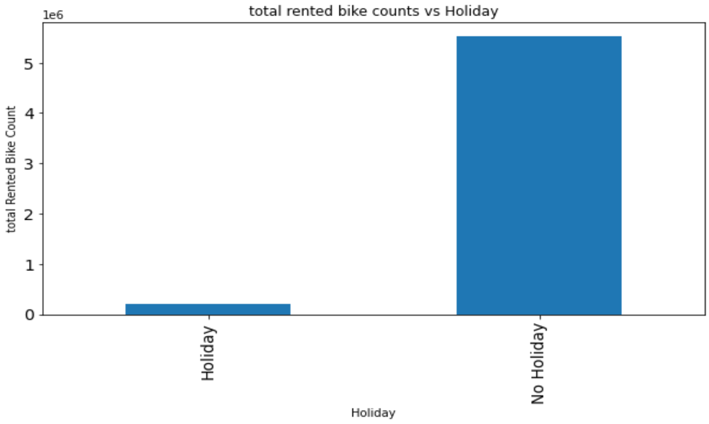


fig: total rented bikes on holiday and non- holiday.

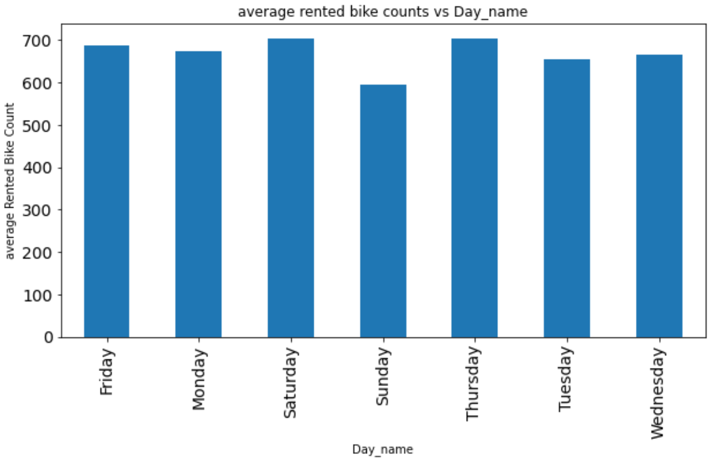


fig: average rented bike on different days.

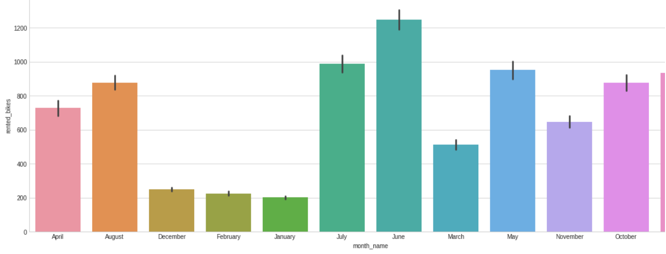


fig: Average rented bike vs different months.

Numerical features- For the numerical features we drew the scatter plot between the numerical features and the dependent variable here it is rented bike count.

The plot displays the data distribution of each dependent variable with respect to the hourly Rental Bike Count.

All the plots are made on Colab Notebook using the sns.plot() function.

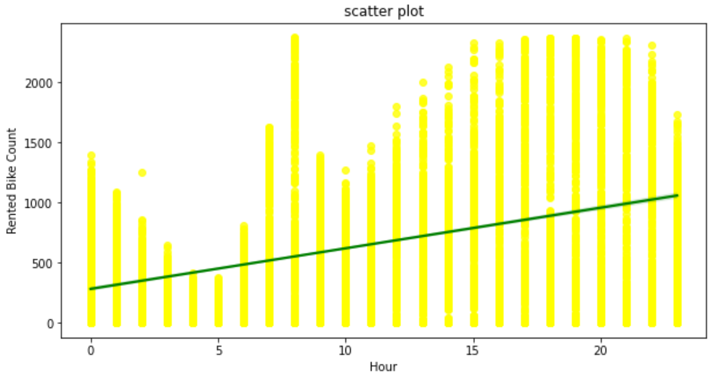


Fig:. Scatter Plot between Rental Bike Count and Hour of the Day with regression line.

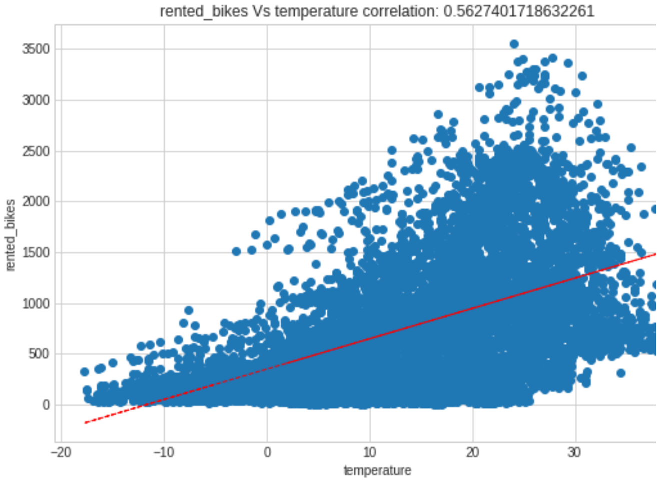


Fig: scatter plot between temp. and rented bike with linear line.

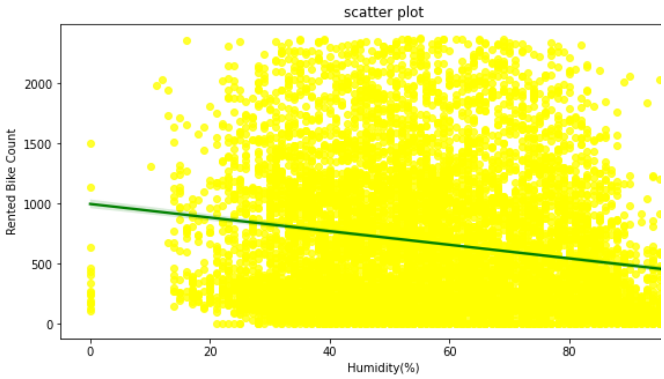


Fig:Scatter Plot between Rental Bike Count and humidity with a linear line.

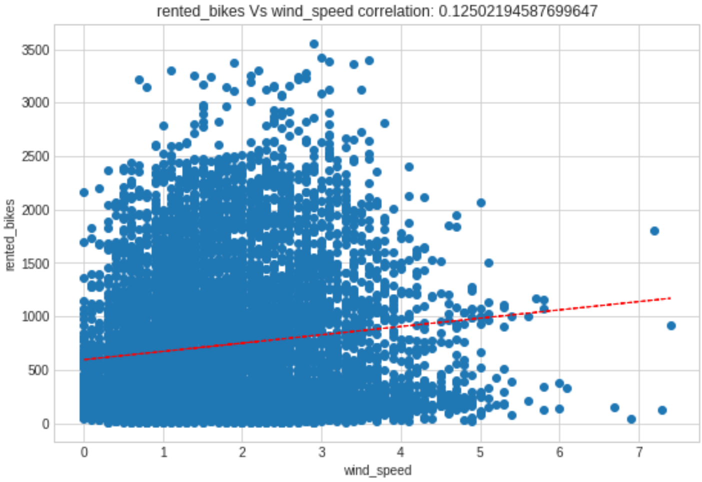


fig: scatter plot between wind speed and rented bike count.

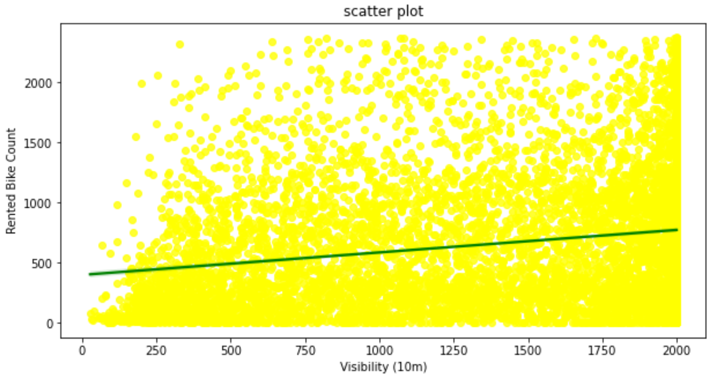


fig: visibility vs rented bike count

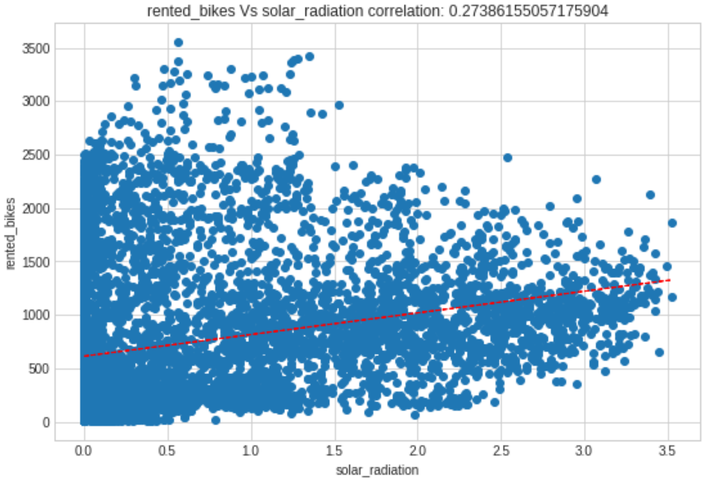


fig: Solar radiation vs rented bike count

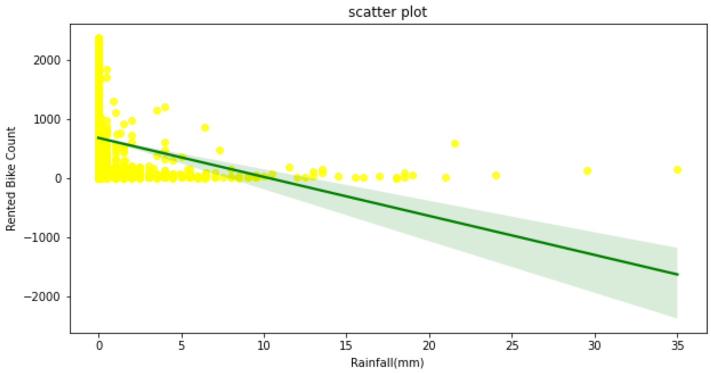


fig: Rainfall vs rented bike count with linear line.

From the above distributions between rentented bike count vs different numerical variables, we can fathom so much information by just looking into the scatter plot carefully like:

1. The range of temperature varies from -20.c to approx 35.c.
2. The maximum percentage of humidity ranges between 20 to 85 %.
3. most of the time the wind speed is between 0 to 5 m/s.
4. Visibility is in the range of 0 to 2000 units.
5. solar radiation is between 0 to 35 mj/m2.
6. maximum times Rainfall lies around in the range of 0 to 25 cm.

correlations between variables

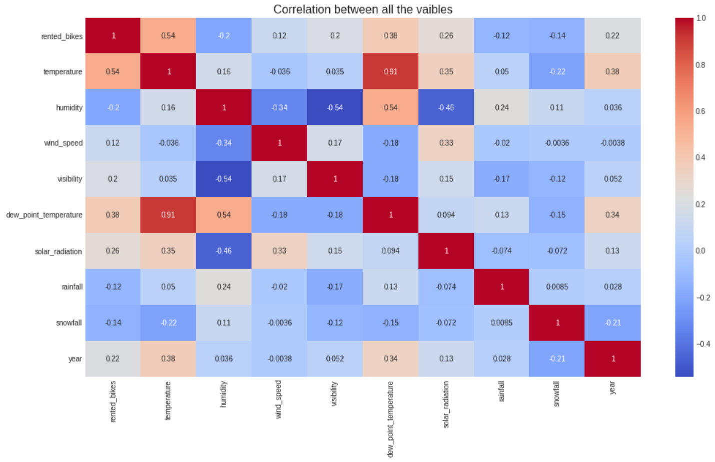


Fig: . Correlation between the variables

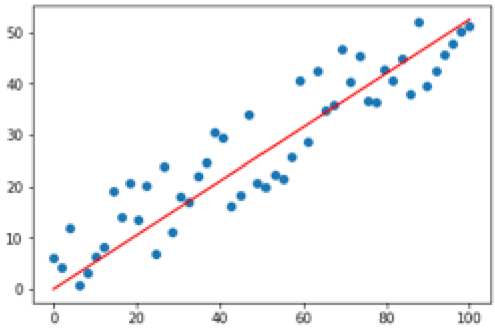
From the above Correlation graph, we can observe that Temperature and Dew Point Temperature are highly correlated, thereby one of the variables would have to be removed from our Regression model, depending on the significance of each variable.

we can see that on the target variable line the most correlated variable to the rent are:

- Hour
- Temperature
- Dew point temperature
- solar radiation

Different models

1. **Linear Regression:**
   Linear regression tries to create a relationship between independent variables and dependent variable by fitting a linear equation to observed data. like, how height will change with respect to age.



A linear regression line has an equation of the form Y = a + bX, where X is the independent variables and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0)

2. **Lasso,Ridge,Elastic net models :**
   There are three popular regularization techniques, each of them aiming at decreasing the size of the coefficients:

   a. Ridge Regression, which penalizes sum of squared coefficients (L2 penalty).
   b. Lasso Regression, which penalizes the sum of absolute values of the coefficients (L1 penalty).
   c. Elastic Net, a convex combination of Ridge and Lasso.
   d. The size of the respective penalty terms can be tuned via cross-validation to find the model's best fit

The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

The cost function for Lasso regression:

$$\min_{\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i^t\beta\right)^2\right\}$$

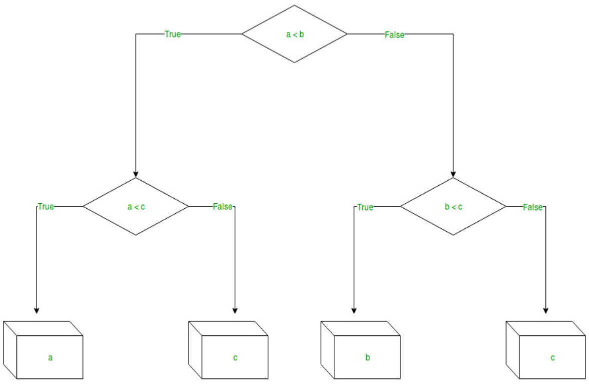
$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \le t_1 \text{ and } \sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \le t_2.$$

The cost fuction for Elastic net:

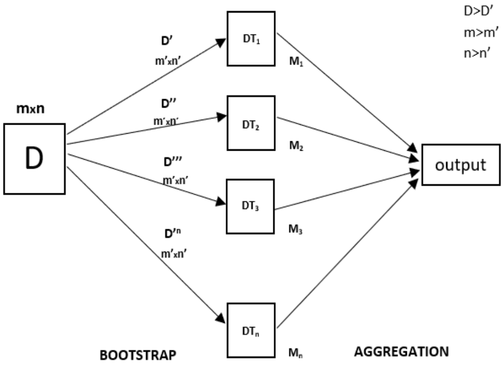
$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|),$$

The penalty terms or we can say hyper parameters can be controlled with the help of GridsearchCV.
There is one major difference between Lasso and Ridge is, in Lasso regression a coefficient will be diminished to Zero but in Ridge it is otherwise.

3. Decision tree regressor:
Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.



3. **Random Forest Regressor:**
A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.



4. **GradientBoostingRegressor:**
Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The gradient boosting algorithm (gbm) can be most easily explained by first introducing the AdaBoost Algorithm.The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. Here, the idea is to improve upon the predictions of the first tree. Our new model is therefore *Tree 1* + *Tree 2*. We then compute the classification error from this new 2-tree ensemble model and grow a third tree to predict the revised residuals. We repeat this process for a specified number of iterations. Subsequent trees help us to classify observations that are not well classified by the previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models

Model performance:

Model can be evaluated by various metrics such as:

1. **Mean Square Error (MSE)**

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

2. **Root Mean Square Error (RMSE)**

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

3. **Mean Absolute Percentage Error (MAPE)**

The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

4. **Mean Absolute Error (MAE)**
Absolute Error is the amount of error in your measurements. It is the difference between the measured value and “true” value.

5. **R^2 Score**

The R^2 score varies between 0 and 100%. It is closely related to the MSE (see below), but not
the same. R^2 defines the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

6. **Adjusted R-Squared**

R2 shows how well terms (data points) fit a curve or line. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will

increase.
Adjusted R2 will always be less than or equal to R2.

CONCLUSION

After the treatment of the outliers, checking for the null values, duplicate values we did EDA on the given dataset.then we apply different regression models on dataset like Linear regression, Lasso, Ridge, Elastic net. from all the four models we get the approximate same R2score **0.69**. MSME(104898.75), RMSE(323.88).

After applying linear models we moved to the advanced model like Decision tree, GBM, Random Forest. We got the highest R2score from the RandomForest model i.e, **0.85.**