ISSN No:-2456-2165

Regression Model to Predict Bike Sharing Demand

Aditya Singh Kashyap
Student: PGDM, Research & Business Analytics
Welingkar Institute of Management
Mumbai, India

Swastika Swastik
Student: PGDM, Research & Business Analytics
Welingkar Institute of Management
Mumbai, India

Abstract:- 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 bikes.

Keywords:- *Data Mining, Linear Regression, Correlation Analysis, Bike Sharing Demand Prediction, Carbon Footprint.*

I. INTRODUCTION

Bike Sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. 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 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, centre on university campuses.

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.

With the onset of Industry 4.0, integration of Internet of Things (IoT) systems with bike-sharing ecosystem has eased the rental process to a significant extent. Real-time tracking of bikes, traffic density, and climate variables aids in gaining useful knowledge about trends, and patterns of renting process, thereby allowing an incisive prediction to meet future demand.

Considering the current ecosystem, bike-sharing can play a vital role in reducing the impact of carbon emissions and other greenhouse gases- major contributors in climate change. Sustainable and clean transport system, if successful, can provide a greener alternative to the traditional car-pool system, and help in reducing traffic congestion, too.

In addition to the environmental benefits, the sharing systems will impart healthier habits among commuting public, who in the hustle of tasking daily routine, often are unable to integrate optimum level of physical activity, which results in a barrage of ailments.

On a positive note, the global Bike-Sharing market size, which was sized at USD 2570.9 million in 2019, is expected to breach the USD 13780 million mark by 2026, with Compound Annual Growth Rate (CAGR) of 26.8% during 2021-2026, as per Market Analysis via MarketWatch.

For our project, we retrieved data from UCI Machine Learning Repository. The dataset contained per day Bike Rental Count with 8760 entries, possessing 14 attributes, out of which 13 variables-12 independent and one dependent-form the part of our Regression Analysis. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Date does not provide relevant information to generate a model to predict the Rental Bike Count.

The primary objective was to build a superior statistical model to predict the number of bicycles that can be rented with the availability of data and understand the trends and factors affecting the rented bike count on a particular day.

II. PROPOSED MODEL

A. Scatter Plot Analysis

For our Rental Cycle Dataset, the Pre-Processing was performed on R Studio. The CSV file was loaded using read.csv() function. The missing data is checked using is.na() function of R. Additionally, which() function was invoked to attain the index numbers of the missing values in

the dataset. The output depicted that there was no missing values in our dataset.

Categorical variables- Seasons, Functioning Day, and Holiday- were converted coded into numerical depictions to fit our Linear Regression analysis. The transformed dataset is loaded as a fresh .csv file using the write.csv() function.

Other than Seasons, Holiday, and Functioning Day, Descriptive Statistics provide detailed information of numerical data in terms of Central Tendency, namely Mean, Median, and Mode.

Data dispersion is also explained via Standard Deviation. Also, the extreme values are represented by Maximum, Minimum, and Range.

For Categorical variables mentioned above, Central Tendency and dispersion become irrelevant. Hence, the description is done with the help of a matrix which shows percentage share of each category in a specific Category. Also, Cumulative percentage shows the validates the part of each category type.

Using Scatterplot, all the set of independent variables are plotted against the dependent variable, Rented Bike Count.

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

In addition to providing initial distribution of continuous data, the scatter plot also aids in identifying any noisy data or outliers which could be removed to gain an optimised linear model.

All the plots are made on RStudio using plot() function.

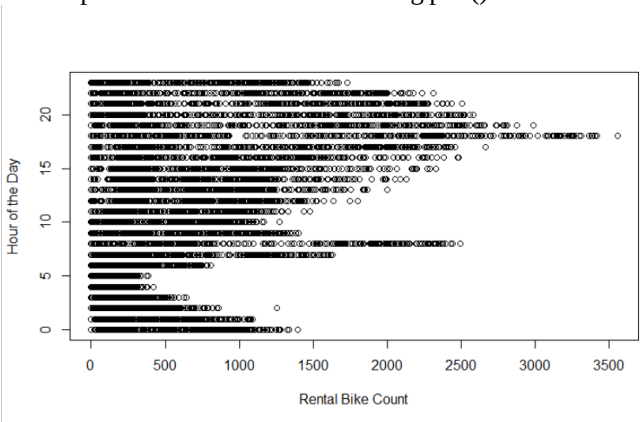

Figure 1. Scatter Plot between Rental Bike Count and Hour of the Day

From the above Scatter Chart, we can observe that data points are closely placed to each other, thereby forming dark linear patterns on the graph.

However, the only secluded point appears to be where the Rented Bike Count exceeds 3500. We have to closely monitor that point over its potential to be an outlier.

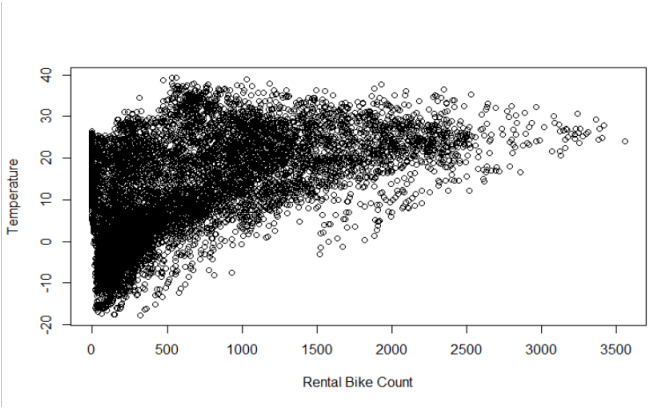

Figure 2. Scatter Plot between Rental Bike Count and Temperature

From the above distribution, Rental Bike Count is spread in form of a cloud which is dense around the region of -20 to 40C. The small tailing clusters towards the higher end of X axis shows that almost all the data points will affect our regression model.

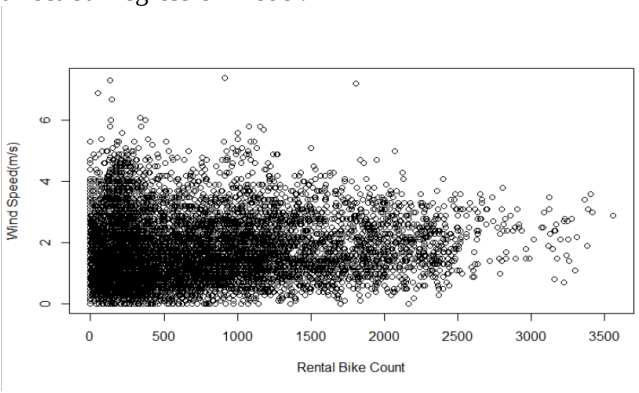

Figure 3. Scatter Plot between Rental Bike Count and Wind Speed

As identified above, in this distribution too, the data points form a prominent cloud around the Wind Speed lying between 0-5 m/s. However, the distribution starts to fade into secluded clusters till breeze of 6m/s.

Post that speed, the data seems to be isolated without any affect from the available data bunch. These points would be considered as potential outliers for our linear regression model.

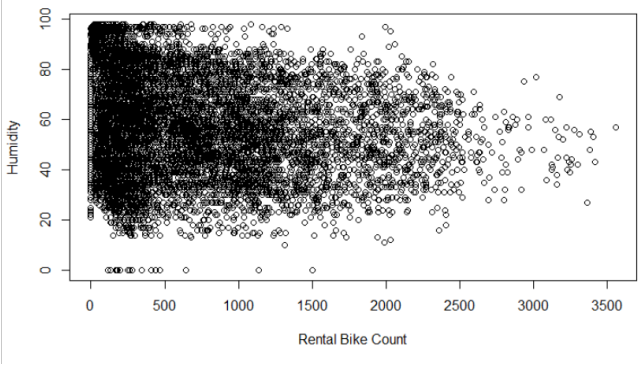

Figure 4. Scatter Plot between Rental Bike Count and Humidity

ISSN No:-2456-2165

From the above Scatter Plot, it is evident that data points form a cloud for Humidity ranging between 20 to 100. Also, it understood that Humidity could not be equal to zero, realistically, meaning a possible data discrepancy.

Additionally, the single data point of Rental Bike Count above 3500 which would be taken into consideration during outlier analysis.

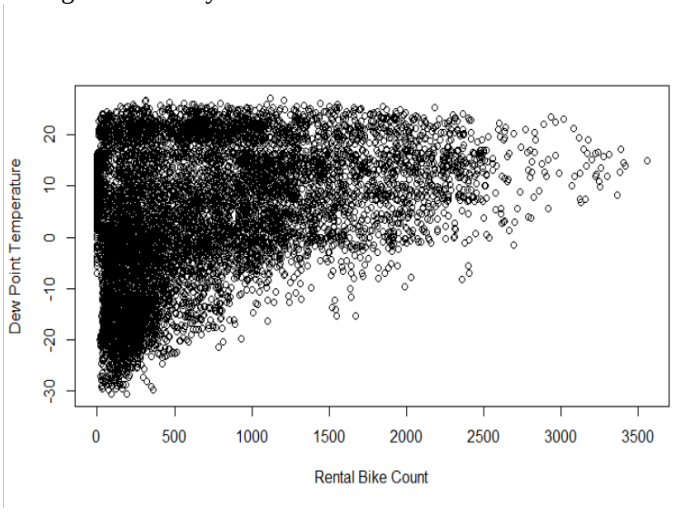

Figure 5. Scatter Plot between Rental Bike Count and Dew Point Temperature

From the above Scatter Plot, the formed data point cloud depicts that Dew Point Temperatures did not make any significant impact until Rental Bike Count reached 500.

After that, there is lightening of the data shade which suggests that Dew Point Temperature made visible effect as number of Rental Bike Counts increased till 3500, where a single data point, again, appears to be secluded.

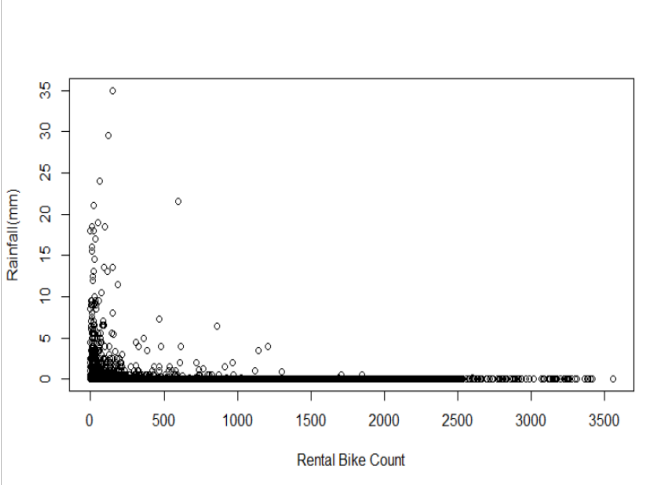

Figure 6. Scatter Plot between Rental Bike Count and Rainfall

As Cycle is an open way to commute between places, Rainfall, almost, will have an inverse relation with the Rental count. The Scatter Plot above also suggests that a significant number of counts lies along the dates when Rainfall was equal to 0 mm.

The scattered clusters above 10 mm to 35 mm show a minimal rise of Rental Bikes from zero. The single data point at 35mm remains an exception and suggests orders for recreational purposes or any other relevant cause. Hence, points above 20 should be considered during outlier analysis.

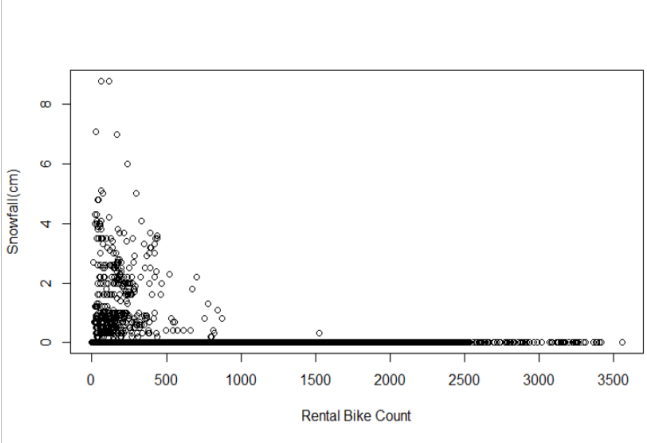

Figure 7. Scatter Plot between Rental Bike Count and Snowfall

As observed in the case of Rainfall, 0 cm Snowfall dominated the Rental Count distribution and clusters lying till 4 cm. Similar to Rainfall, a few data points above 6 cm suggest Rental cycles for recreation or any other relevant cause.

Above scatterplots present a clear picture about how outliers affect the regression model for our dataset. On monitoring data points on multiple visualisations, R was used to trim our dataset initially containing 8700 entries, ending up on 8567 observations.

B. Correlation Analysis

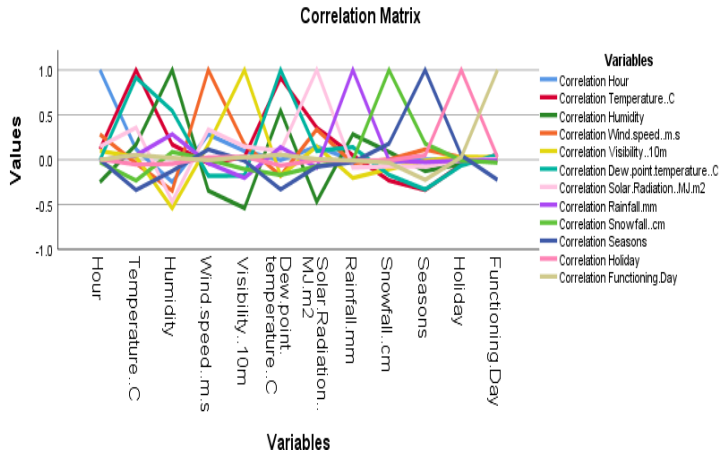

Figure 8. Correlational Analysis of the dataset variables

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.

III. RESULTS AND DISCUSSIONS

A. Regression Parameters

Before the outlier treatment we obtained a Regression Model on the dataset containing 8760 observations. The parameters saw a slight improvement after Outlier Treatment and Correlation analysis.

To refine our model, we cleaned a few outliers to obtain an efficient Regression line. Rental Orders above 2500 were removed from our dataset, owing to the scattered distribution leading to noisy data.

Similarly, Rainfall, Snowfall, Solar Radiation, and Wind Speed entries exceeding 10mm, 4cm, 3.5 MJ/m2, and 5m/s respectively were removed from our dataset, too. Our final dataset comprises of 8567 observations.

R Value is the coefficient between the Predicted and Observed values of the dependent variable. 0.753 suggests a high positive correlation between the Original and Forecasted Rental Bike count.

R-Square Value is the goodness-of-fit and a statistical measure of how close the data are fitted to the regression line. The table value of 0.567 suggests that our linear regression mos able to determine 56.7% of changes in the Rental Bike Count.

Adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. It calculates R-Square of only Independent Variables those are statistically significant.

A minute difference between R-Square and Adjusted R-Square suggests all our Independent Variables being significant, despite both values being on a relatively lower side.

R-square change, which is just the improvement in R-square when the second predictor is added. The R-square change is tested with an F-test, which is referred to as the F-change. A significant F-change means that the variables added in that step significantly improved the prediction.

B. Hypothesis Formation

- $H1_0$: There is no relationship between Rental Bike Count and Independent Variables.
- $H1_1$: There exists a relationship between Rental Bike Count and Independent Variable.
- $H2_0$: There is no statistical significance between Rental Bike Count and Explanatory Variables
- $H2_1$: There exists some statistical significance between Dependent Variables and Explanatory Variables and not all coefficients are Zero.

Inferences:

- Since P-Value (all variables from table) is less than 0.05 for H1, we reject $H1_0$, meaning there exists a relationship between independent and dependent variables.
- Since P-Value (0.000, from ANOVA table) is less than 0.05 for H2, we reject $H2_0$, meaning our model is statistically significant.

C. Linear Regression Equation

Rental Bike Count = 2413.629 + 25.381*Hour – 15.050*Humidity + 12.966*Wind Speed + 29.671*Dew Point Temperature - 64.133*Solar Radiation – 106.550*Rainfall + 49.574*Snowfall -107.074*Seasons -97.747*Holiday - 932.492*Functioning Day

IV. CONCLUSION

We calculated a linear regression, which clearly shows that,

- Functioning Hour is a significant negative predictor (estimate = -932.492).
- It also shows that Seasons compared to humidity is a significant negative predictor (estimate = -15.050) and
- Snowfall compared to Wind speed is a significant positive predictor (estimate = -12.966) of bike rentals.
- The Visibility and Temperature was not included because its p-value (0.580) & (0.528) respectively exceeded alpha value of 0.05 making it insignificant.

Regression models with low R-squared values can be perfectly good models for several reasons.

Some fields of study have an inherently greater amount of unexplainable variation. In these areas, your R2 values are bound to be lower. For example, studies that try to explain human behaviour generally have R2 values less than 50%. People are just harder to predict than things like physical processes.

Fortunately, if you have a low R-squared value but the independent variables are statistically significant, you can still draw important conclusions about the relationships between the variables. Statistically significant coefficients continue to represent the mean change in the dependent variable given a one-unit shift in the independent variable. Clearly, being able to draw conclusions like this is vital.

As observed in our case, 0.56 is a relatively low value but statistical significance aids us to understand the factors affecting the Rental Bike Count better.

To extract better results and patterns from the datasets, advanced algorithms like Classification Trees, Random Forest, K Nearest Neighbours, could be implemented.

REFERENCES

[1]. Sathishkumar V E, Jangwoo Park, Yongyun Cho (2020), 'Using data mining techniques for bike sharing demand prediction in metropolitan city', The International Journal for the Computer and Telecommunications Industry.
[2]. Sathishkumar V E and Yongyun Cho (2020), 'A rule-based model for Seoul Bike sharing demand prediction

using weather data', European Journal of Remote Sensing.
[3]. 'Bike Sharing: It's about the community', Cycling Industries Europe - https://cyclingindustries.com/fileadmin/content/documents/170707_Benefits_of_Bike_Sharing_UK_AI.pdf
[4]. Seoul Bike Sharing Demand Data Set, UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#
[5]. Hadi Fanaee-T, and Joao Gama (2013), 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence.
[6]. Bike-Sharing Service Market Market Size 2021-2026, MarketWatch - https://www.marketwatch.com/press-release/bike-sharing-service-market-market-size-2021-2026-comprehensive-study-development-status-opportunities-future-plans-competitive-landscape-and-growth-2021-01-11
[7]. Data Source :http://data.seoul.go.kr/