Predicting Bike Sharing Rental Demands

DATA 1030: Midterm Project Report

https://github.com/rads1306/DATA1030-Radhika-Mehrotra.git

Supervised by Prof. Andras Zsom

Radhika Mehrotra

I. Introduction

Stock procurement is often times a challenge for many businesses because it requires predicting future consumer demand of a specific good or service. This can be extremely challenging because most times, demand patterns are a result of interplay between a multitude of factors. Thus businesses often seek to find indicators that can allow them to optimize stock procurement.

Patterns in consumption coupled with knowledge of how external factors affect consumption serve as useful indicators to solve such a business problem.

In this project, I use Seoul's bike sharing demand data set to understand how certain features of the day such as weather conditions, presence of holiday, and time of the day can affect the demand for bike rental in Seoul.

Target variable: Rented Bike Count

The "Rented Bike count", in other words the number of bikes rented in a given hour on a particular date is the target variable. It follows a continuous distribution. We want to understand the quantitative influence of the other features in in the dataset, on the number of bikes rented and then leverage the quantitative analysis to predict bike rental patterns in the future.

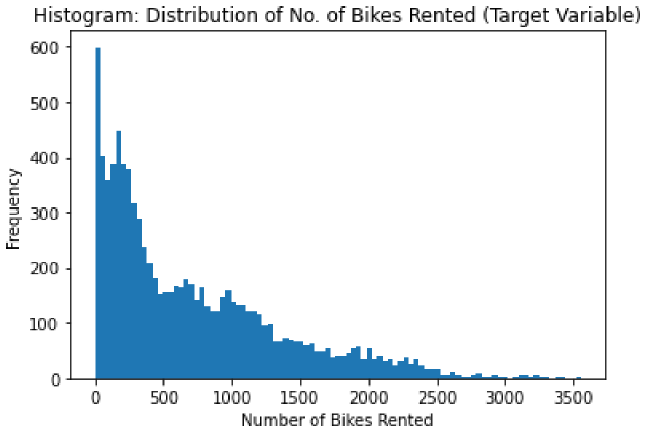

Figure 1: Histogram depicting the distribution of the number of bikes rented

Methodology

To build this predictive model, I plan to leverage tools of regression to tease out a correlation between the factors that affect the quantity of bike rentals.

Understanding the data

Rows: The dataset has exactly one year's worth of data, that has been collected per hour of everyday in the year. The number of rows are thus 365 * 24 = 8760.

Columns: There are 14 columns that each describe the characteristic of the day/hour when the bike rental took place. Most of the columns provide data on the weather of the day.

Thus we have a total of 14*8760 = 1,22,640 data points and no missing values.

Feature Name	**Description**	**Distribution**	**Data Type**	**No. of unique Records**
Date	date on which the bike was rented. The data is collected from December 2017 to December 2018. Format- dd/mm/yyyy	Continuous	object	365
Rented Bike count (Target Variable)	Count of bikes rented at each hour on a given date	Continuous	Integer	2166
Hour	Hour of the day represented in the 24 hour clock system (Takes values from 0 to 23)	Categorical	Integer	24
Temperature	Temperature calculated in degree Celsius at a specific hour for a specific date.	Continuous	Float	564
Humidity	Percentage point level of humidity for a given hour in a given day	Continuous	Float	90
Windspeed	The speed in m/sec at which the wind is blowing for a given hour in a given day	Continuous	Float	65
Visibility	Level of visibility in tens of metres for a given hour in a given day	Continuous	Integer	1789
DewPoint Temperature	The temperature to which air must be cooled to become saturated with water vapour. Measured in degree Celsius.	Continuous	Float	556
Solar radiation	Electromagnetic radiation emitted by the sun captured in MJ/m2	Continuous	Float	345
Rainfall	Amount of rainfall in a given day on a given hour measured in mm	Continuous	Float	61
Snowfall	Amount of snowfall in a given day on a given hour measured in cm	Continuous	Float	
Seasons	The four seasons that occur in Seoul	Categorical	Object	4 (Winter, Spring, Summer, Autumn)
Holiday	Whether the day is a holiday or not. This can mean weekends or national holidays. It is not clearly given in the dataset.	Categorical	Object	2 (Holiday, No Holiday)
Functioning Day	Whether the day is a working day for the Bike renting company or not	Categorical	Object	2 (Yes, No)

Literature Review

- *Sathishkumar V E & Yongyun Cho (2020) A rule-based model for Seoul Bike sharing demand prediction using weather data, European Journal of Remote Sensing, 53:sup1, 166-183, DOI:* 10.1080/22797254.2020.1725789

  The increasing popularity in bike sharing for environmental reasons has led to several studies on bike sharing systems. Researchers are coming up with systems that prop up bike sharing systems in metropolitan cities. These systems try to predict bike renting demand not just based on different weather conditions and times of the day but also based on geospatial factors. The predictions have enabled them to devise more efficient systems of managing bike rental stations, procuring optimal levels of stock and redistributing inventory from one geographic zone to another. Research has uncovered that more interesting factors like presence of cycle paths have also affected bike rental demands.

II. Exploratory Data Analysis

This section draws from some interesting graphs that I encountered during the process of EDA.

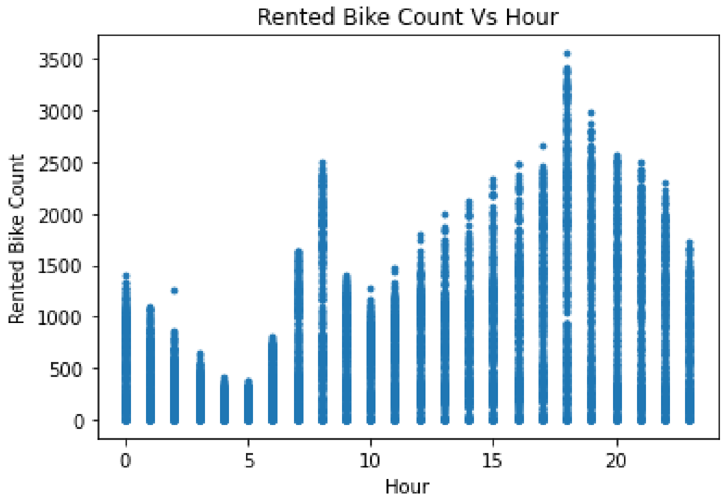

Figure 2: Number of bikes rented at every hour during the day

Figure 2 was interesting to me because it has a unique distribution- a bimodal one. The x-axis represents the time (24 hour clock) where 0 is 12am all the way up to 23 which represents 11pm. The y-axis represents the target variable, that is, the number of bikes that were rented. The scatter plot maps all the 8760 data points and gives us an interesting distribution- there are two peaks- one at 8am and the other at 6pm. Probably this sudden rise in demand at 8am and 6pm each day can be attributed to people using the bikes to travel to and from their workplace/school. This distribution allows the company to identify "peak hours" of bike demand and can influence the way they run their day-to-day operations.

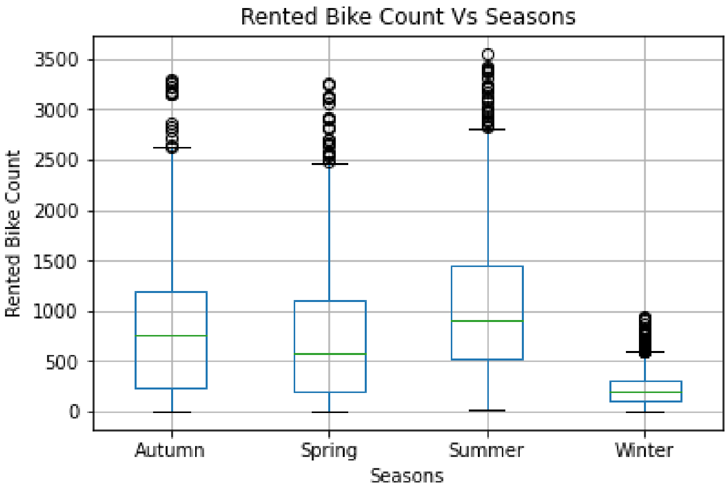

Figure 3: Number of bikes rented in different seasons

Figure 3 represents the difference in the distribution of number of bikes rented for different seasons. The x-axis represents the 4 exhaustive seasons that occur in Seoul and the y-axis is the target variable- number of bikes rented. As intuition suggests, in the winter months the median value of bikes rented is much lower than the median values of the other seasons. Once again the insights from this graph will help the businesses procure limited stock in seasons where the demand for bike rentals is not high.

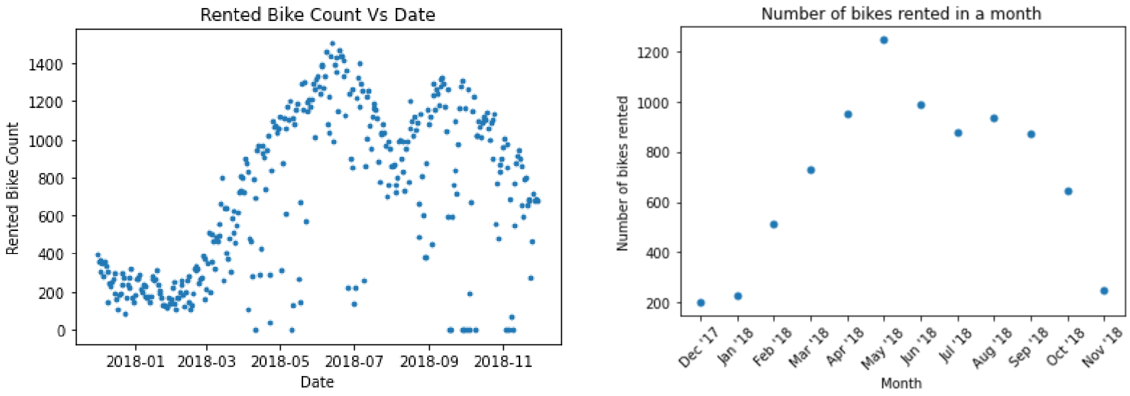

Figure 4: Number of bikes rented across days/months

In Figure 4, the two plots are essentially showing the same thing; the only difference is the level of granularity at which the EDA was done. The plot on the left was done at a day level and the one on the right at a monthly level. The x-axis is populated with dates from December 2017 all the way up to November 2018. The y-axis represents the number of bikes rented. The plot is an interesting one because of the way it is shaped. It can be seen that towards the end of the year 2017, the demand for bike renting started growing and then it grew almost exponentially until July of 2018. Dates in July of 2018, witnessed a sudden drop in demand and this persisted until almost September of 2018. This could probably be attributed either a policy change or a change in consumption patterns (like a drop in fuel prices). Nevertheless, after a short period of a dry spell for the rental company, the demand for bike renting sees a steep growth again and then a decline. These cyclical fluctuations allow us to inspect deeper into what might be underlying it and allow the company to drive business decisions accordingly.

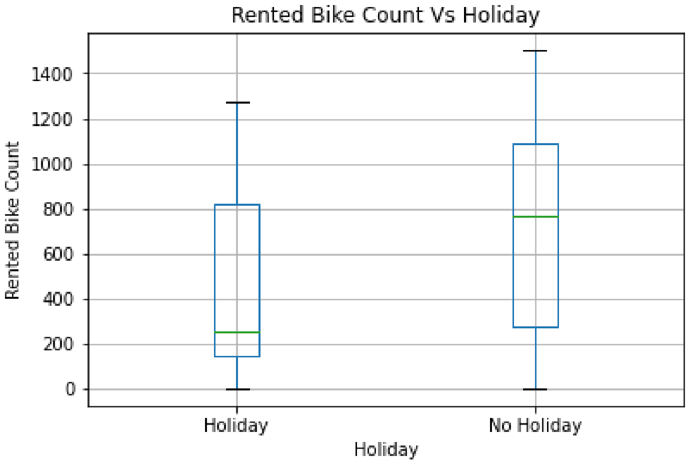

Figure 5: Number of bikes rented on a holiday vs not on a holiday

In Figure 5 the x-axis represents the two categories that a day can be binned into: holiday or not a holiday and the y-axis tells you the number of bikes rented. The plot is interesting because of the stark difference in the median number of bikes rented on a holiday versus a day that is not a holiday. This also aligns with what I mentioned earlier, that a lot of the people are probably renting the bikes to get to and from their workplace. This provides an interesting insight for the bike company. They can leverage this data to promote subsidised weekend passes or advertise schemes or fun events that would nudge people in the city to ride the bikes more on weekends or national holidays.

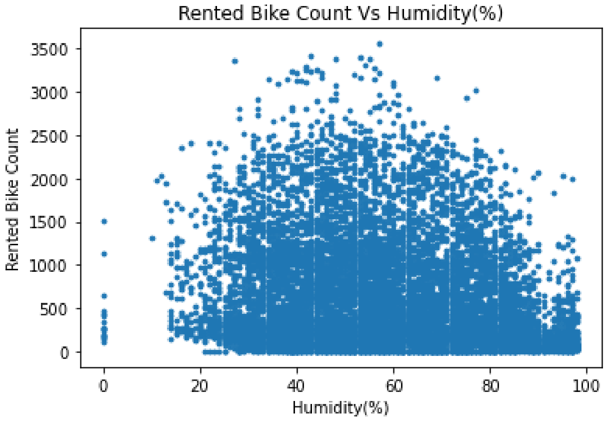

Figure 6: Number of bikes rented given the level of humidity

Figure 6 is interesting because I'm not sure why it is distributed like that. On the x-axis is the humidity level in percentage terms and on the y-axis is the number of bikes rented. The plot is imbalanced in the sense that when humidity levels are low, then there is almost no demand for the bikes. I'm not sure of the exact reason but my best guess is that the low humidity months are the winter months and hence the number of bikes rented is low.

III. Data Splitting and pre-processing

Data Splitting

Initially my dataset was a time series one because in order to predict of number of bikes rented on a particular day *u,* I would have to base my predictions on historical data from days *v,w, z* (that occurred prior to *u*) only.

However, my project doesn't require me to work on time series data because I am predicting the number of bikes rented given certain weather conditions. Thus, I'm not modelling a time-series predictive model, and hence I decided to drop the "Date" column which would have otherwise made it a time series dataset. Instead, I will use all other feature columns (Hour, Temperature, Humidity, Wind Speed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, Snowfall, Seasons, Holiday, Function Day) as part of my feature matrix (X). My target variable (y) as mentioned before is the column "Rented Bike Count".

Now that I have removed the time factor in my dataset, we know that each data point in every one of my columns is IID because they all are generated from the same distribution and the way it is generated is not dependent on the value generated preceding it; hence my dataset is IID.

I used the basic splitting strategy (train-test-val strategy) where I allocated 80% of my data to the training set on which the predictive model will be trained and I split the remaining 20% equally for the purpose of testing and validating the results of the model. I used a large percentage of values in the training set because I have a dataset of over 8000 points which is a good amount of data and leaves a good amount of data points for both the test and the validation set.

Data Pre-processing

For the data pre-processing part I used both the one-hot encoder and the standard scalar because my dataset has both categorical and continuous features. I applied the One-hot encoder on "Seasons", "Holiday" and "Functioning Day" because they are all categorical but unordered in the sense that no season is preferred over the other or can be ranked greater than the other and similarly for holiday and functioning day. I have 8 features after pre-processing the categorical variables.

For the remaining larger chunk of the feature variables (Hour, Temperature, Humidity, Wind Speed, Visibility, Dew Point Temperature, Solar Radiation, Rainfall, Snowfall) I used the standard scalar because they are all continuous features and they all follow a tailed distribution. I have the same 10 features after pre-processing the continuous variables.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Sathishkumar V E & Yongyun Cho (2020) A rule-based model for Seoul Bike sharing demand prediction using weather data, European Journal of Remote Sensing, 53:sup1, 166-183, DOI: 10.1080/22797254.2020.1725789