Improvement of Seoul Bike Sharing Linear Model Using Feature Engineering, Model Selection and Gradient Descent

Rajat Kumar Sahu[1], Santosh Susarla[1], Ashish Alden Dsouza[1], Aashish Pathak[1], Sudha B G[2]

[1]PES University, Electronic City Campus, Bangalore, India
[2]Great Learning, Bangalore, India

Abstract . This project presents an improvement of Seoul Bike Sharing Linear Model Using Feature Engineering, Model Selection and Gradient Descent. Nowadays, rental bikes are introduced to the public in almost all the top urban cities in the world for enhancement of mobile comfort ability and environmental factors [1]. The most important part is the prediction of bike counts or bike availability at a certain point at an hourly basis for stable supply of rental bikes. The dataset includes weather information like temperature, humidity, wind speed, visibility, dew point, solar radiation, rainfall etc. This project explores the application of Linear Regression models for predicting bike count for meeting the necessary demands. This paper discusses the models for hourly rental bike demand prediction.

Keywords: Linear Regression, Feature Engineering, Model Selection, Gradient Descent

1 Introduction

Currently the bike renting scheme is well accepted by the public in the top urban parts of the world. Most of these bike rental companies allow people to borrow and return a bike from a common bike rental station. For expanding availability of bicycles for public use, the company allocate a truck that collects bicycles parked in various stations and relocate them to the original station gradually. These rental bikes come with a GPS tracking system for the user to know at which points on the map the bikes are available and for the company to know where the bikes are dropped-off so that they can relocate it to the most demanding stations for the customers. The bike rental has helped see a lot of positive environmental impact with betterment of physical human activity mentally and physically. Which is one of the few reasons why this bike sharing concept is a trend and quite stable business.

2 Dataset Description

We have sourced the data from UCI Machine Learning Repository.The dataset contains the count of public bikes rented each hour in Seoul Bike Sharing System with the corresponding weather data and holidays information used include weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. Since any form of transportation mainly depends upon the Climatic conditions, the corresponding weather information such as Temperature, Humidity, Wind speed, Visibility, Dew point temperature, rainfall, and snowfall for each hour is added. The processed data consists of the total count of rental bikes rented at each hour with date/time variable and Weather information.

3 Methodology

Linear regression - Linear regression model (LM) in the most simplest method, that is equated with the relationship between the Y attribute of the scalar output and one or even more X attributes of the input quantity. The case of an independent attribute is known as simple linear regression, and the method is called as multiple linear regressions when more than one independent attributes are considered. Data is designed using linear predictor functions in linear regression, and from data, the unknown model parameters are estimated. Usually, Linear regression refers to a system where the conditional mean of Y is an affine function of X, given the value of X. The model is assumed as in Eq. (1). $Y = \beta 0 + \beta 1X + s$ (1)

Here $\beta 0$ and $\beta 1$ are two unknown constants representing the intercept and slope, also known as parameters or coefficients, and s is the term of error.

3.1 Data understanding & Visualization of the analysis

Data pre-processing :-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8465 entries, 0 to 8464
Data columns (total 14 columns):
Date                       8465 non-null object
Rented Bike Count          8465 non-null int64
Hour                       8465 non-null int64
Temperature(°C)            8465 non-null float64
Humidity(%)                8465 non-null int64
Wind speed (m/s)           8465 non-null float64
Visibility (10m)           8465 non-null int64
Dew point temperature(°C)  8465 non-null float64
Solar Radiation (MJ/m2)    8465 non-null float64
Rainfall(mm)               8465 non-null float64
Snowfall (cm)              8465 non-null float64
Seasons                    8465 non-null object
Holiday                    8465 non-null object
Functioning Day            8465 non-null object
dtypes: float64(6), int64(4), object(4)
memory usage: 926.0+ KB
```

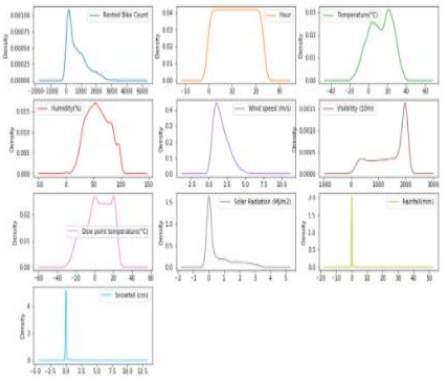

Fig. 1. Data types

Fig 2: - Features - distribution

The shape of the dataset is (8760, 14). We checked the null values and treated it accordingly. Also we found the date to be unique so we will drop it later, following which we checked the distribution of the categorical variables. We checked the rented bike count if zero for any rows and found 5 rows having zero bike count. Then we checked the count of functioning day and non functioning day. And found out that there are 8465 functioning days and 408 holidays. Following which we checked the correlation to see if any variables are correlated with each other.

We segregated the categorical and numerical variables and on grouping the Hour and Rented bike count we found the average bikes required every hour. We plotted a KDE plot and we can say that the data is right skewed also maximum values lie between 0 and 1000.Following which we plotted all the numerical variables. After plotting a count plot for the rented bikes across holidays and non-holidays we saw that the bike count is comparatively a lot more on 'No Holiday' than during a holiday.

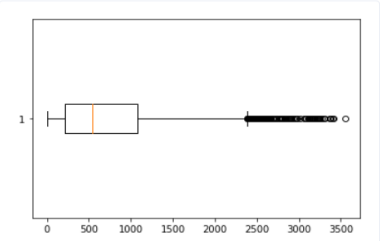

Fig 3: - Box plot showing outliers in data

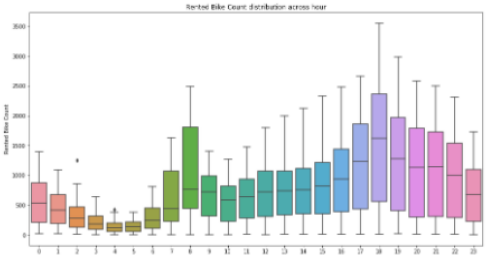

Fig 4: - Rented Bike Count Across 24-hour format

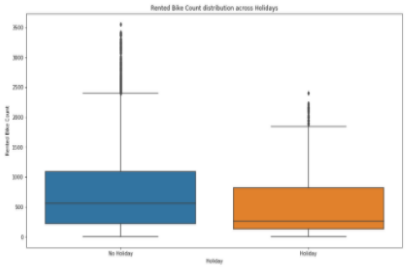

Fig 5: - Box plot for holiday and non-holiday

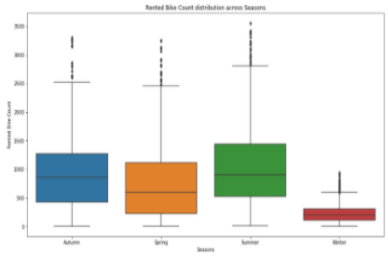

Fig 6: - Box plots for different seasons

We plotted a box plot for the rented bike count and found a lot many outliers in the data. Following which we plotted a heat map and found that the dew point temperature and the Temperature have a very high positive correlation.

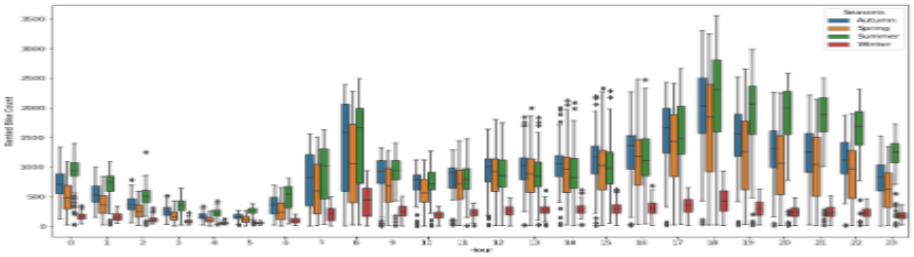

Fig 7: - Combination of different boxplots

Detecting and treatment of outliers:

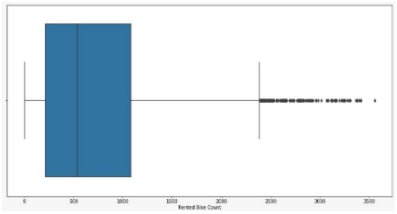

Fig 8:- Before outlier treatment

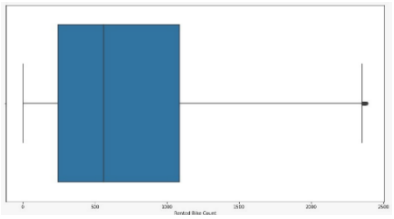

Fig 9:- After outlier treatment

Previously we saw there are several outliers in our dataset so to remove them we dropped the date as it was insignificant. So as per the plotted graph in the Jupyter notebook we can see that the maximum outliers are present in the rented bike count and for wind speed, rainfall, snowfall are equal to 0, hence we dropped rented bike count and visibility.