Project: Bike Sharing Prediction

This report documents the analysis and model development for the Bike Sharing Dataset. It contains the following steps:

Table of Contents

About the Bike Sharing Dataset: An Overview

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, the user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure, and arrival position is explicitly recorded in these systems. This feature turns the bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

At first the required libraries has ben set up and read the datatset. Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv. The attributes are as follows:

	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual
0	1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0	3
1	2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0	8
2	3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0	5

•instant: record index

•dteday : date

•season : season (1:springer, 2:summer, 3:fall, 4:winter)

•yr : year (0: 2011, 1:2012)

•mnth : month (1 to 12)

•hr : hour (0 to 23)

•holiday : weather day is holiday or not

•weekday : day of the week

•workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

•weathersit :

- •1: Clear, Few clouds, Partly cloudy, Partly cloudy
- •2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- •3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- •4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

•temp: Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)

•atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)

•hum: Normalized humidity. The values are divided to 100 (max)

•windspeed: Normalized wind speed. The values are divided to 67 (max)

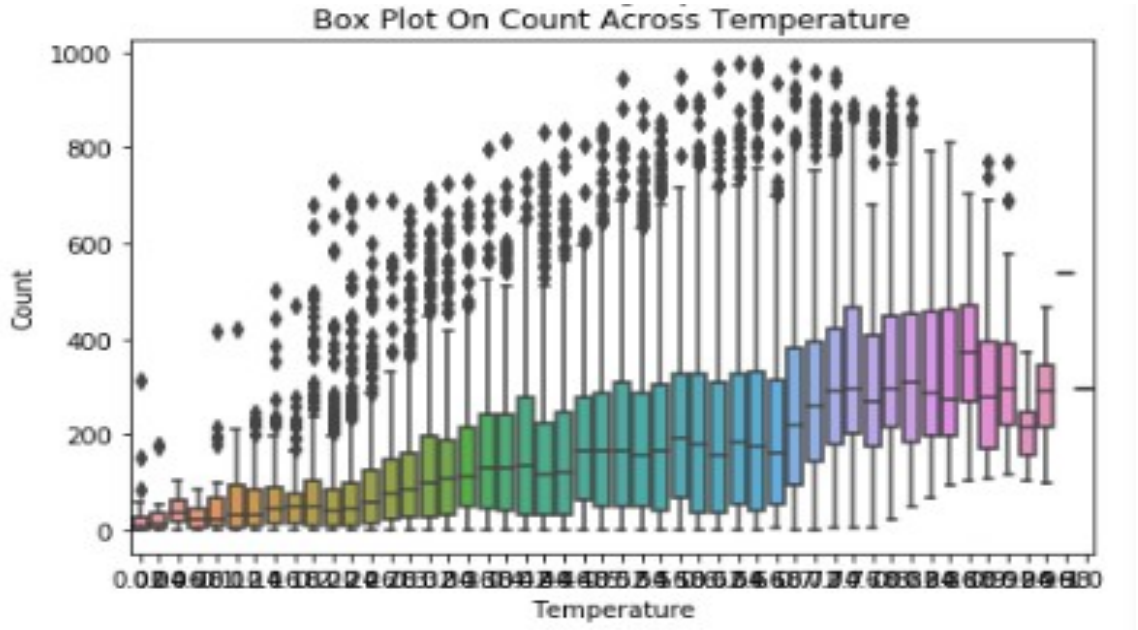
•casual: count of casual users

•registered: count of registered users

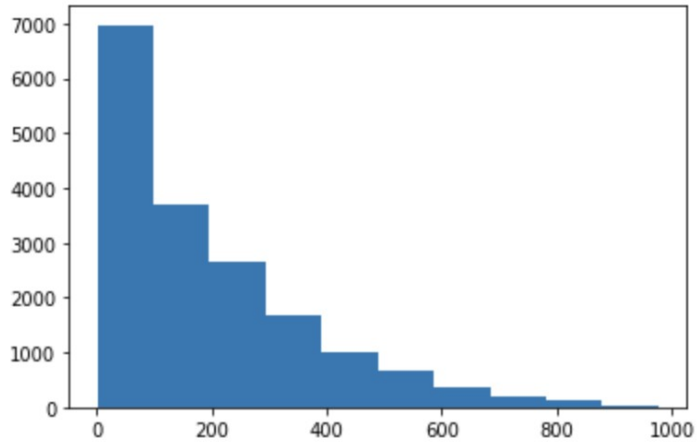
•cnt: count of total rental bikes including both casual and registered

The data has been explored, finding out missing values if there is any and analyzed the outlier as well. The Interpretation is that the working day and holiday box plots indicate that more bicycles are rent during normal working days than on weekends or holidays.

The hourly box plots show a local maximum at 8 am and one at 5 pm which indicates that most users of the bicycle rental service use the bikes to get to work or school.

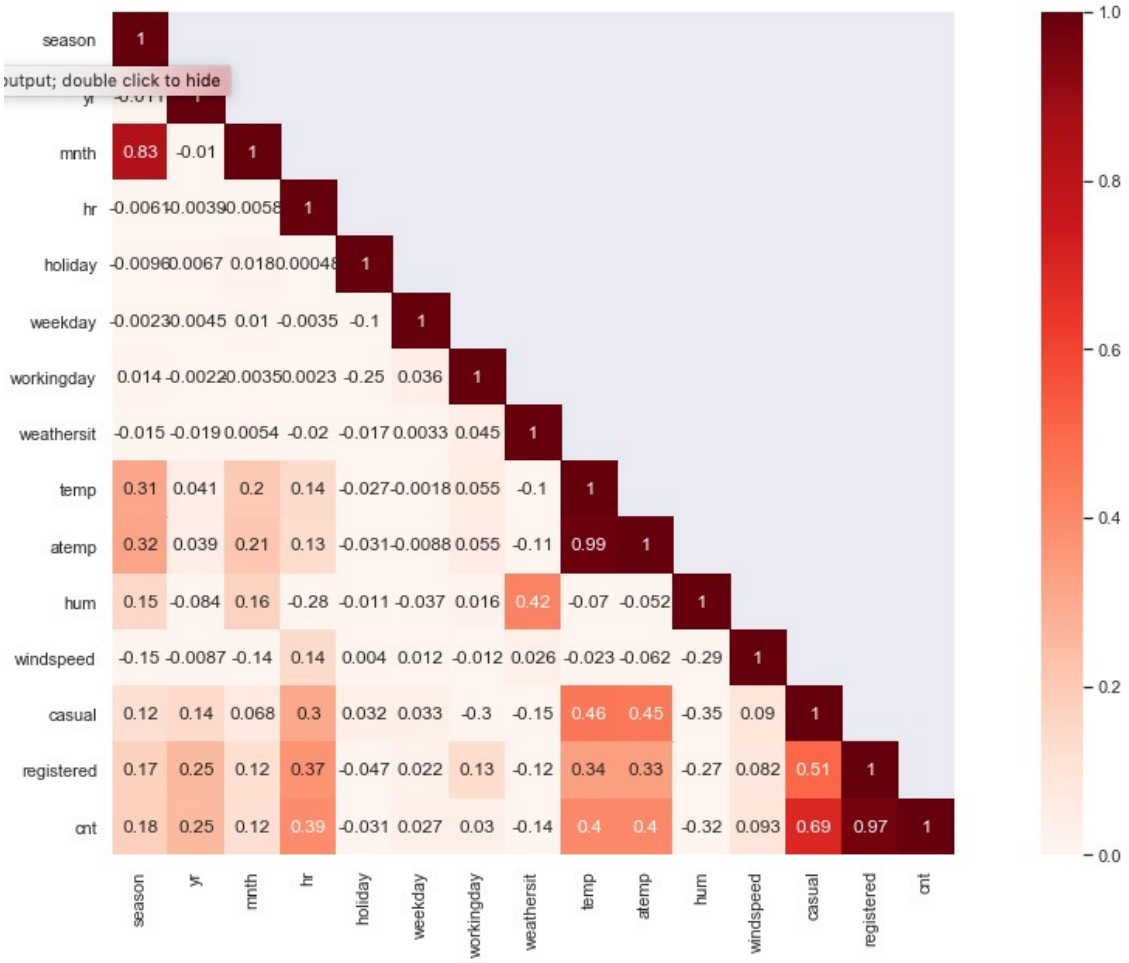


Another important factor seems to be the temperature: higher temperatures lead to an increasing number of bike rents and lower temperatures not only decrease the average number of rents but also shows more outliers in the data.



As it is visible from the below figures that "count" variable is skewed towards right. It is desirable to have Normal distribution as most of the machine learning techniques require dependent variable to be Normal. One possible solution is to take log transformation on "count" variable after removing outlier data points. After the transformation the data looks lot better but still not ideally following normal distribution.

Correlation Analysis:



At the end of the descriptive analysis chapter, we can note the following points:

- Casual and registered contain direct information about the bike sharing count which is to predict (data leakage). Therefore they are not considered in the feature set.
- The variables "temp" and "atemp" are strongly correlated. To reduce the dimensionality of the predictive model, the feature "atemp" is dismissed.
- The variables "hr" and "temp" seem to be promising features for the bike sharing count prediction.

It is quiet obvious that people tend to rent bike during summer season since it is really conducive to ride bike at that season.Therefore June, July and August has got relatively higher demand for bicycle.

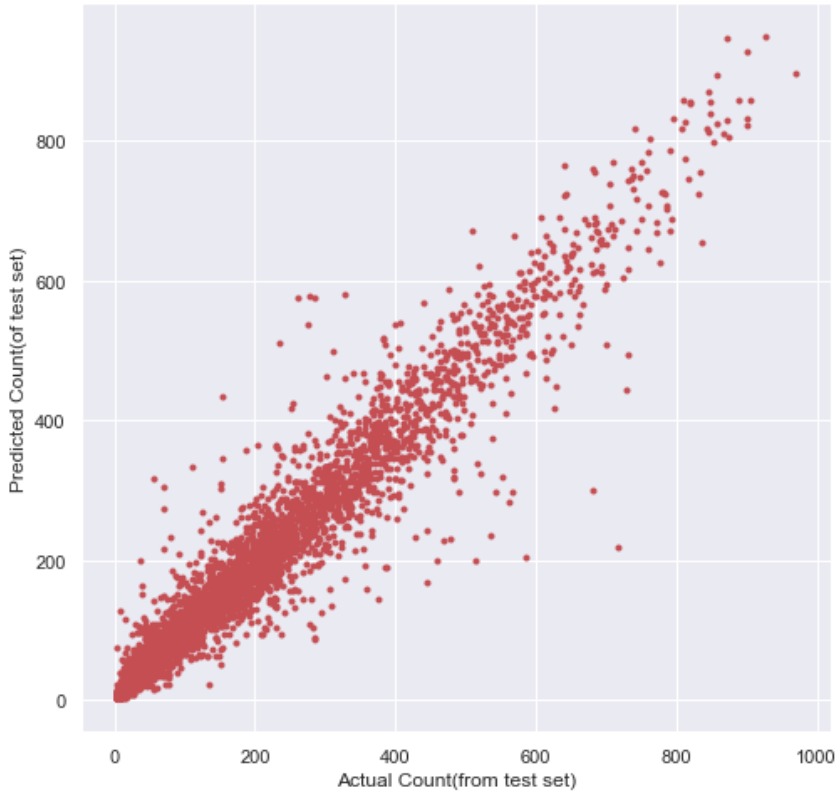
- On weekdays more people tend to rent bicycle around 7AM-8AM and 5PM-6PM. As we mentioned earlier this can be attributed to regular school and office commuters.
- Above pattern is not observed on "Saturday" and "Sunday".More people tend to rent bicycle between 10AM and 4PM. The peak user count around 7AM-8AM and 5PM-6PM is purely contributed by registered user.

I have developed a linear regression model to predict the log(count) based on all other features except registered, casual and atemp.

The variance of the predictions is very high as seen in the below plot.

We develop a Random Forest model with 20 estimators to predict the log(count) based on all other features except registered, casual and a temp.

The variance of predictions is much lesser than that of linear regression in the below plot.



The feature importance plot shown above shows that 'hr', 'temp','working day' are the 3 most important to predict count.