Bike Sharing Report

Background

Bike sharing systems are very popular in recent years. Users can easily rent a bike from a position and return it in another position. They play a very important role in city transportation.

The dataset is generated by the system. As a data scientist, we are going to do a exploratory data analysis and create a model to predict the total usage based on the other features.

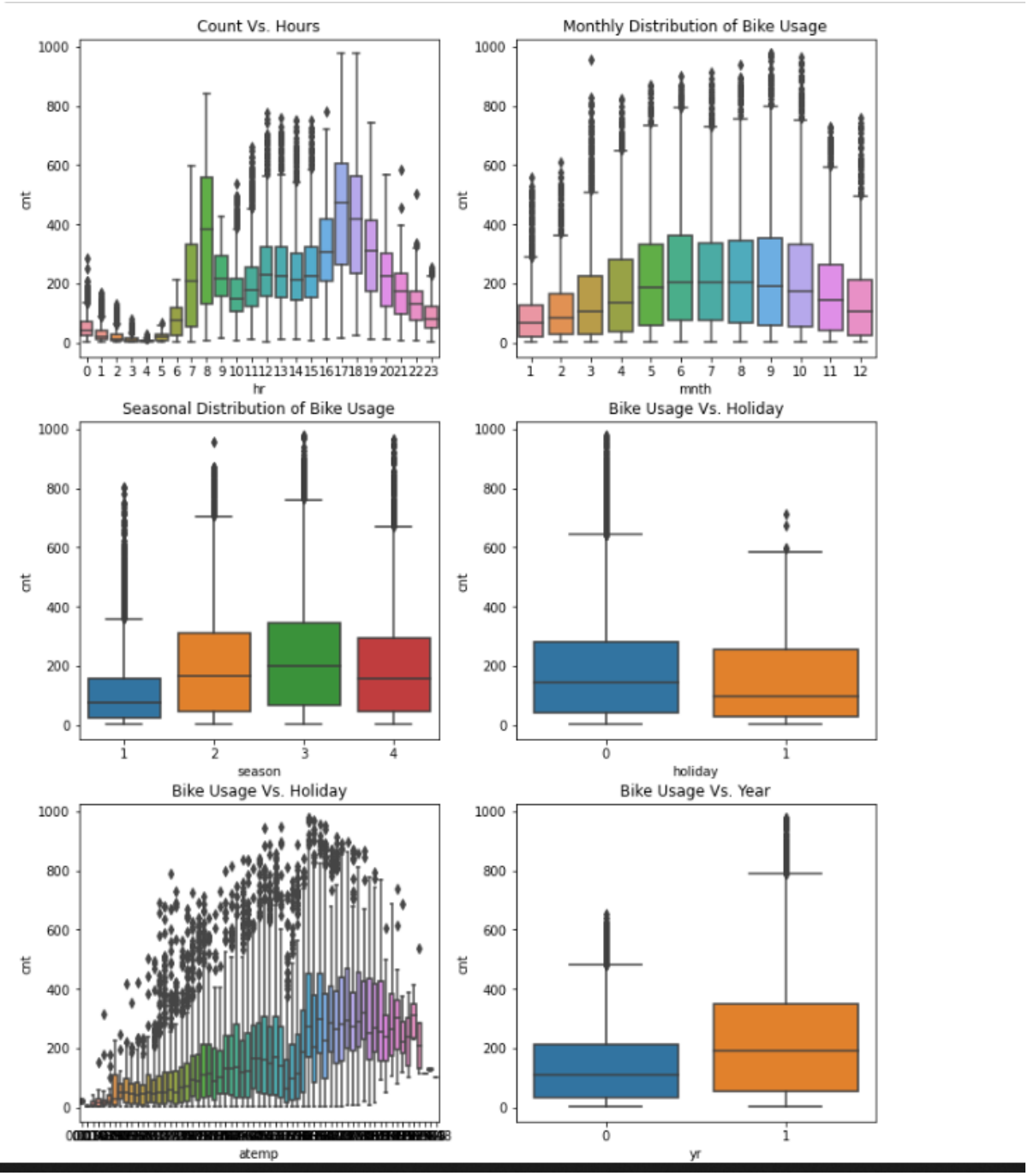
Dataset

Bike sharing rental process is highly correlated to many factors, such as temperature, hours of the day, etc. The core dataset is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C, USA which is publicly available in http://capitalbikeshare.com/system-data. There are two dataset which are hourly and daily basis. I would use the hourly basis in this report because I would like to observe the hourly difference. The explanation of the columns will be in the data dictionary.

EDA(Exploratory Data Analysis)

I found out 8 am in the morning, 5 and 6 pm in the evening are the peak hours for bike usage. From monthly distribution plot, I found out 4 to 12 months are having more bike usage. The same distribution also occurs in the seasonal distribution plot. Bike usage in 2-4 seasons are higher. I also make a plot about bike usage with weather condition. There is a significant increase with better weather condition. I also noticed that there is a significant increase as the

temperature increase. The bike usage in the year 2012 is slightly increased compared to the year 2011. There is a slightly different bike usage between workdays and holidays.



Features Selections

I created a correlation heatmap for numerical feature selection. Temp and atemp are two same columns in different scale. I just kept the temp. In this report, our target is cnt column. Casual and registered column are a part of cnt column. So, I removed these two columns. Based on the correlation map, I removed the windspeed column which shows nonsignificant. All categorical columns show significant relationship with cnt column based on the boxplots we observed previously. Totally we selected season, yr, mnth, hr, holiday, weekday, workingday, weathersit, temp, hum for further prediction.

Preprocessing/Model

I did normalization for numerical columns and one hot encode for categorical columns. Then I split data into train, validation and test set. After that, I created a simple linear regression model, the RMSE is about 100 which is our base model. At that point, I feel like I want to try regular neural network and see how it works. Because regular neural network can handle complex problem, it can learn detect the relationships automatically. The RMSE for regular neural networks is about 51 which is a huge improvement compared to our base model. Our model is performing well on validation set based the loss plot. And residual plots is pretty

normally distributed about 0.