Capstone 2 milestone report

Executive Summary

This milestone report is intended to demonstrate the initial steps taken towards the overall project goal of finding bike share count predictive algorithm based on unique features. This is part of the Capstone class in the Springboard data science specialization. The dataset used in developing a model was downloaded from the UCI repository.

The problem

Given the Bike Sharing dataset with hourly level information of bikes along with weather and other attributes, model a system which can predict the bike count.

The client

The client could be a business interested in renting bikes and wants to know how many bikes could be shared in an hour or so. It could also be a government entity who wants to decrease a carbon foot print in the corresponding metropolitan area.

Data Set Information

Bike sharing systems are new generation of traditional bike rentals where entire process from membership, rental and return has become automatic. Through these systems,

user can easily rent a bike from a position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their key 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 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 notable events in the city could be detected via monitoring these data.

Data wrangling

The provided data files were processed using panda's tools. Summary statistics such as dataset size were gathered to better understand the contents of the files. Missing and unknown objects were removed. I have standardized the column name to make it more readable. I type cast the attributes to reflect the intended values of the columns.

Visualization of the hourly distribution counts give us different interesting relationships. The season wise hourly distribution of counts shows that in all the seasons there are peaks around 8AM and 5PM which are considered to be rush hours.

The analysis of hourly counts by weekday wise hourly distribution shows that on average weekdays have a higher usage than weekends. The analysis of a random hourly distribution of counts shows that early hours and late nights have low counts, but

significant outliers as do afternoon hours. Peak hours have higher medians and counts with virtually no outliers.

The monthly distribution of total counts shows that the highest counts occur from June to October of the year which makes senses as it’s the fall season of the year.

If we have other datasets that could explain the nature of our bikes or their efficiencies, it could help as improve our bike share count. Moreover, having a dataset that explains the hourly weather condition could also be very helpful.

I investigated the effects of the features have on bike share count. The collinearity between features was also investigated to determine which features are more important to predict bike share count.