Capstone proposal: Seid Ahmed

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.

Attribute Information:

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

- instant: record index
- dteday : date
- season: season (1: spring, 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 (extracted from [Web Link])
- 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

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.

General outlines

Import the required libraries

Download the data from the website

Analyze the dataset and check if there are missing values

Check the Data types if there is a non-numerical value.

Check the skewness of the data distribution and scale the data if its skewed.

Explore the target with unique features to determine which feature is more important and decisive.

The scatter plots and heat map could give as a good hint how the features are related.

We evaluate the dataset with some baseline algorithms with the non-scaled data.

After checking the scores of the algorithms, we will proceed to evaluate the same algorithms with a scaled data.

We tune the algorithms we chose based on the scores we get.

We prepare a pipeline to show the reproducibility of the above process.

Present the model.

Deliverables

A code that do all the above