Hubway System Data

Exploratory Data Analysis and Machine Learning

Background about Hubway:

Hubway is the bike sharing system in Boston and its neighboring towns that was found in 2011. You can subscribe and be part of one of their plans which lets you rent cycle for 30 minutes and you have to pay for every extra 30 minutes after that.

Analysis:

Cleaning and EDA -

Hubway has around 5.8 million users, what I wanted to do was check if I can predict the number of trips and what variables will be the most defining factors in this case. The Hubway System data website has all their system data publically available which can be readily accessed and once I had that I merged it with the weather data for the same time duration(that I got from Kaggle) to check if temperature, pressure, and humidity can be some of the defining variables for my problem.

But before I could start with the Regression issue, there were some fundamental issues with that had to be resolved for any algorithm to work efficiently. One of the significant anomalies that I encountered was that there were trips that were going on for over 40000 minutes(which kind of seems impossible and one of the reason behind this can be that a lot of times people don’t dock their bike correctly and the timer goes on). Also, there were rides for less than 3 minutes starting and ending at the same station which seems that somebody accidentally docked the bike in and out. A lot of these issues can be relevant and can point to other significant problems faced by Hubway, but for my Prediction Analysis, this was the data that I did not need.

Multi-Variable Regression

After having a clean dataset, I wanted to predict the number of rides.
In my first attempt, I tried taking the number of days and the years, but I did not get a model that would perfectly fit our curve.

I was sure that temperature should be one of the defining factors in this, and after adding that variable in my model, I saw the observations which were pretty good, but for all the three variables the t-statistic was pointing out that I should accept the Null Hypothesis (which will void the entire point).

On more research, I found that Machine Learning considers days as a categorical variable, so I used dummy variables for days, and once I had those values, I re-run the Regression Algorithm which gave us the desirable amount.

On further analysis, we also saw that all the weather factors such as temperature, humidity pressure, and wind speed have no impact on our model.

Algorithms and Packages Used:

Plotly, Matplotlib, Seaborn – To plot the graphs and charts

Statsmodels- Linear Regression – To create and run the model

References

https://www.thehubway.com/system-data

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html

https://plot.ly/pandas/getting-started/#hosting-on-plotly