Clustering Capital Bikeshare Stations and Predicting Usage: A Final Project for Intro to Data Science at General Assembly

Destry R. Saul

August 2013

ABSTRACT

Using historical data from the Capital Bikeshare system in Washington, D.C. and temperature data from NOAA, I have built models for each Capital Bikeshare station to predict the flow of bikes into or out of that station in a given hour. Using Kmeans clustering I have also grouped the stations by predicted behavior. There appear to be three broad categories: stations with no flow, stations with negative morning and positive evening spikes, and stations with positive morning and negative evening spikes.

1. Introduction

Bikeshare systems are gaining popularity throughout the U.S. and as the systems become more heavily used, how to redistribute bikes to maintain a balance of available bikes and empty stalls becomes more difficult and more important. For my final project for General Assembly NY's Data Science course, I have built a model that could be used to predict when a station will run out of bikes or fill up. I began by gather data from the Capital Bikeshare system in Washington, D.C. This is one of the older U.S. systems with data available from 2010 to the present. I also collected temperature data from NOAA, though the models using temperature did not perform better the models using only time. In addition to the supervised learning of models for each station, I used unsupervised learning to cluster the stations by predicted bike flow. These groups could be used to simplify the models to three or five templates for a faster deployed system.

All data, figures, and scripts are available at
https://github.com/destrys/citibike/tree/master/capital_bikeshare.

2. Data Collection and Manipulation

Historical data from the Capital Bikeshare system is available at
http://www.capitalbikeshare.com/trip-history-data. These data are formatted such that

each line contains information on a single trip: duration, start/end times, start/end stations, etc. After downloading the data for 2011 and 2012, I built a DataFrame in python using the pandas module[1]. There are approximately 2.1 million trips during 2011 and 2012. Instead of storing individual trips, I aggregated the trips into one hour bins and calculated the net flow of bikes at each station resulting in about 17,000 rows.

Weather data is available from the NOAA at http://www.ncdc.noaa.gov. I downloaded hourly temperature readings from Air Force Catalog Station Number 997314 which is located at 38.867° N 77.017° E.

3. Supervised Learning: Predicting Station Usage

To predict the change in the number of bikes at a given station I chose to used support vector regression because I am trying to predict a quantity that is nonlinearly related to time. The two features I chose to use were time and temperature.

For evaluating different models, I chose to focus on one station. SVR can be slow, and with almost 200 stations, I could not iterated over the full set. I used station 31111 which is located at 10th and U st NW. This bike flow at this station has a visible pattern when plotted again time of day or the hour of the week (weekhour) as shown in Figure 1.

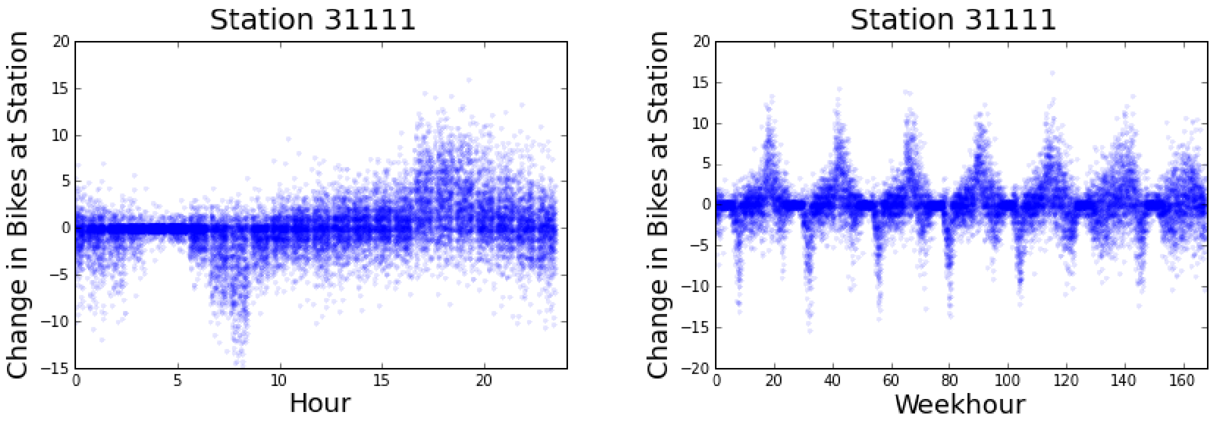

Fig. 1.— *left:* The change of bikes at station 31111 as a function of the time of day, the points have been jittered in both dimension and made partially transparent to aid the eye. *right:* The change in bikes at station 31111 as a function of the hour of the week with zero corresponding to midnight on Sunday night, local time.

[1] http://pandas.pydata.org/

To fit the models I used the sklearn[2] module and specifically the sklearn.svm.SVR() class. I fit multiple models of increasing complexity. The results for each model are listed in the table below. The 'no model' model assumes no change in the number of bikes and is useful as a baseline. All of the models performed much better than the 'no model'.

[h] model	training MSE	mean cross validation MSE
no model	7.47	-
hour	5.76	5.82
weekhour	5.09	5.29
weekhour, temp	5.12	5.67
weekhour, temp, year	5.07	5.67

The model with the best mean cross validation mean-squared-error (MSE) was the model using just the hour of the week. I was surprised that using temperature did not significantly improve the model. The difference in the training and cross validation scores for those models leads me to think that the cause may be due to overfitting. Figure 2 shows how the predicted flow of bikes depends on temperature for a specific time. The variability suggests overfitting because it seems unphysical that at 62 degrees twice as many bikes will return to the station than when it is 58 degrees. One solution for this would be to bin the temperature into 5 degree bins, but I did not get to trying that.

4. Unsupervised Learning: Station Clustering

Besides the physical location of the stations, I did not know anything about the neighborhoods they we in. After fitting models for all of the stations, I used Kmeans clustering to see if some stations have the same usage patterns. To do this, I first predicted the flow of bikes for each station as a function of the weekhour fo a constant temperature of 70 degrees during 2012 using the model with weekhour, temperature, and year. The unclustered models are shown in Figure 3.

Using three clusters produced very distinct groups. One with little bike flow, one with positive morning/negative evening, and one with negative morning/positive evening. The clustered models and there locations on a map of Washington are in Figure 4. The maps were

[2] http://scikit-learn.org/

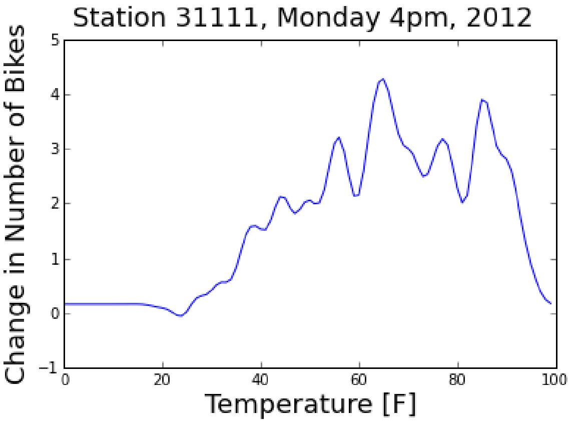

Fig. 2.— Predicted change in the number of bikes at station 31111 for Mondays at 4pm during 2012 using the model with weekhour, temperature, and year. The spiky behavior suggests that the model is overfitting.

produced using cartoDB[3]. The negative morning/positive evening stations are mostly in the Northeast while the positive morning/negative evening stations are centrally located. The low flow stations are all of the distant stations, with some scattered throughout. I also used 5-cluster Kmeans to see if the low flow stations in the interior were more similar to the other interior stations. This resulted in strong and weak clusters of each of the positive/negative groups. One use of this clustering would be to create template models for each cluster, simplifying from the 200 models here to 3 or 5 that may work just as well for balancing purposes.

5. Conclusion

The models I have built are a significant improvement over having no model, but there are many addition features that could be added. Some features I have though of are: school and work holidays, precipitation, depth of snow on the ground, and station full or empty status. Even without these features, this work could be used to build a tool to detect when a station needs balancing.

The clustering was effective, producing distinct groups of stations. These group are related to how people are regularly using the stations. This could be used to reduce the

[3] http://cartodb.com/

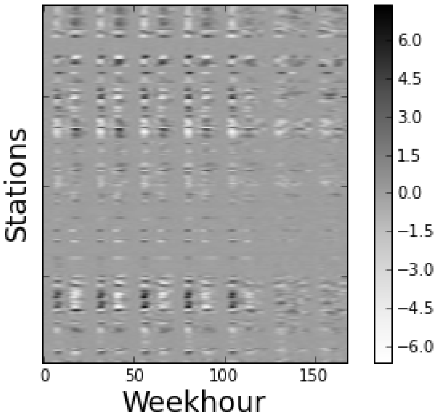

Fig. 3.— Each row of this image is a model for a station for a constant temperature of 70 degrees in 2012. Color represents the flow of bikes for that station.

number of models needed to represent the whole system, or to infer similarities between the neighborhoods.

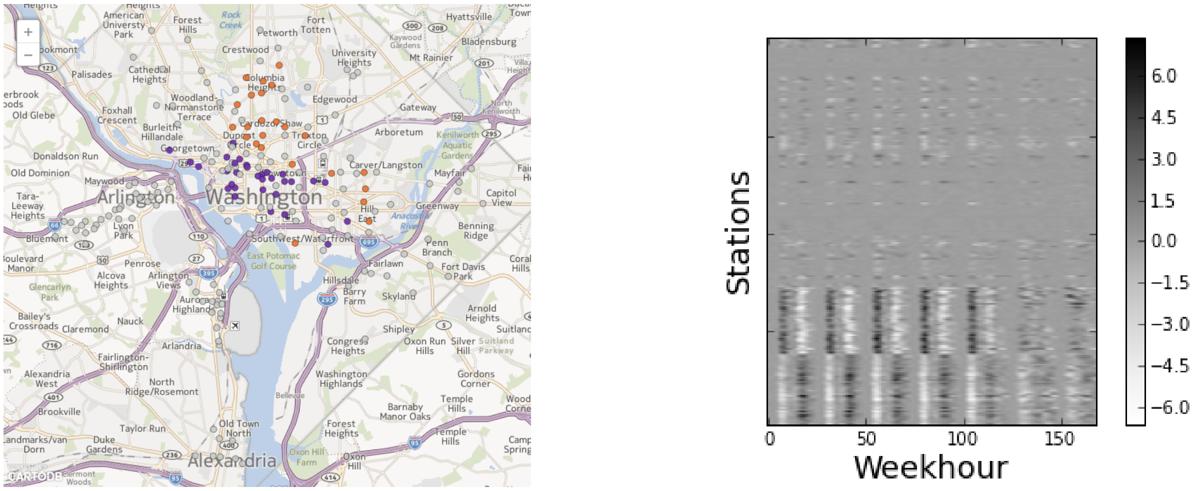

Fig. 4.— Results of 3-cluster Kmeans. *left:* Station map colored by group. Grey for the models with little flow, orange for the negative morning/positive evening, and purple for the positive morning/negative evening.*right:* Models sorted by group.

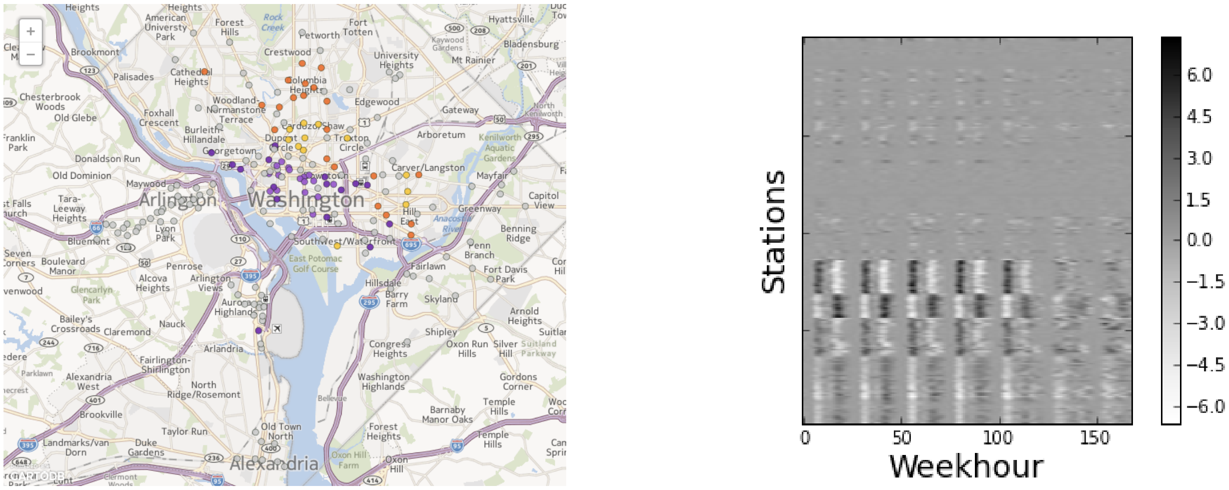

Fig. 5.— Results of 5-cluster Kmeans. *left:* Station map colored by group. Grey for the models with little flow, orange for negative morning/positive evening, and purple for the positive morning/negative evening. The darker colors represent the stronger models while the light colors are the weaker models (the weaker models are the lower two sets in the right image.)*right:* Models sorted by group.