Summary of Analysis

About: This is a summary of the analysis for the project done by Yoav as part of The Data Incubator program in Fall 2015. It can be found at: https://github.com/yshaham/docksvsbicycles

Title: Docks vs. Bicycles

A Short Description of My Project:

I created an App that was deployed on Heroku, whose purpose is to help operations in Citibike, the bicycle sharing service in New York City, by predicting in real time which docking stations are going to be empty or full and when, using historical and current system and weather data and machine learning to make the predictions. A link to my App: https://nameless-plateau-2699.herokuapp.com

Related projects by former Fellows of The Data Incubator:

See the following three links:

1) https://predictbikenyc.herokuapp.com/ - predicting bicycle trip duration.

2) https://citirhythms.herokuapp.com/ - predicting demand for bicycles (but not by station) using a linear additive model fitted to the logarithm of trip counts, and analyzing spatial components of bicycle traffic and how they depend on time of day and day of the week.

3) https://gpine.shinyapps.io/my_app/ - predicting demand for bicycles by station in Washington, DC using a Poisson model.

New Features in My Project:

Predicting demand for bicycles is only half the story, and the second half is the demand for docks when bicycles are returned to stations, hence the title of my project. Also, for the service to be available at a station it is important to examine, not the absolute demand for bicycles, but the difference between the number of bicycles that leave a station and the number of bicycles that return to it. The results of the above projects by former Fellows helped me to identify important features that should be included in my models, and standing on shoulders of giants I was able to address both halves of the story and create an App that makes it easy for anyone to use the results of the analysis (that is, the user does not need to have knowledge in statistics).

For my models I used boosted regression trees with Poisson distribution (I also tested Random Forest

and Poisson GLM, but based on cross validation the results of the boosted regression trees were better).

Data Sources:

1. For Model Training:

1.1. Data on bicycle trips taken between April 2014 (older data may not be reliable, see: http://www.rausnitz.com/blog/2014/04/bad-data/) and August 2015 (latest month available at the time of analysis). The total number of bicycle trips exceeds 10 million. The total size of the data files (as unzipped CSV files) exceeds 2 GB. The data was downloaded from: http://www.citibikenyc.com/system-data

1.2. Hourly (or more frequent) data on weather in New York during the above period was read from daily summaries (for example: http://www.wunderground.com/history/airport/KNYC/2015/07/04/DailyHistory.html?req_city=New+York&req_state=NY&req_statename=New+York&reqdb.zip=10001&reqdb.magic=5&reqdb.wmo=99999&format=1) and was saved to files using the Python program: https://github.com/yshaham/docksvsbicycles/blob/master/Python_programs/get_weather_history.py

2. For Real-Time predictions:

2.1. Citibike system status, i.e. how many bicycles and docks are available in each station, is read in real time from: http://www.citibikenyc.com/stations/json

2.2 Current weather data and the weather forecast are read in real time in accordance with the instructions at: http://openweathermap.org/current and http://openweathermap.org/forecast5

Exploratory Data Analysis

The project proposal that I submitted to The Data Incubator as part of the application to the program was based on the following results from exploratory data analysis done on about 180 MB of data (in unzipped CSV format) on over 950,000 bicycle trips taken in summer 2014. I did this analysis before I became aware of the projects of former Fellows of The Data Incubator.

This analysis made it clear that there is a big difference in bicycle demand over time between workdays and weekends. This can be seen in figure 1. This graph also shows the difference between two groups of users of Citibike, the Subscribers, who use it for their daily commute, hence the demand peaks in morning and afternoon rush hours, and the Customers, who are short term users of Citibike, for example, tourists. The difference between workdays and weekends and the dependence of demand on the time of day became important variables in the predictive model that I trained for this project (based on cross validation the difference between the two groups of users was not included).

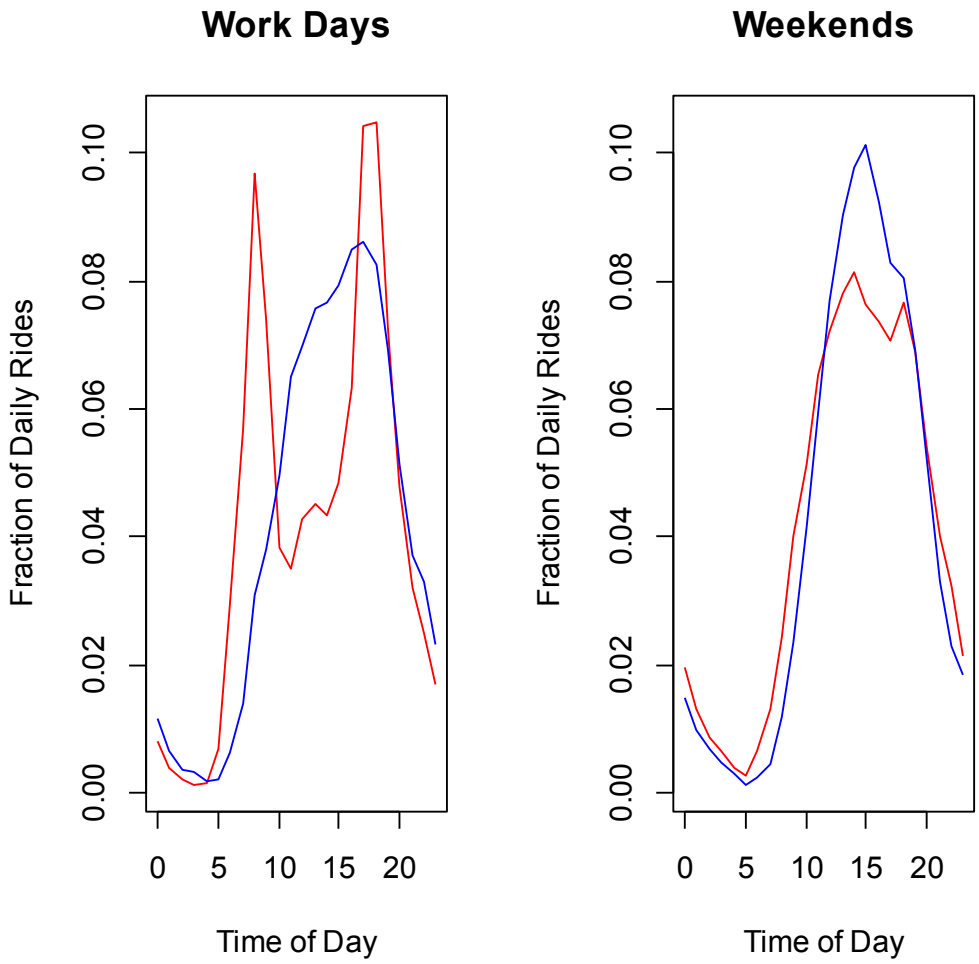

Figure 1: Total demand for bicycles as a function of time of day on an average summer day (workday or weekend; based on data from July 2014). The red line is for subscribers and the blue line is for short term customers. Note that the lines have different normalization factors, and the total for subscribers is actually almost ten times the total for short term customers.

My exploratory analysis also showed the spatial imbalances that are created between stations as bicycles are redistributed according to user demand. This can be seen in figure 2. These imbalances need to be rebalanced by Citibike employees, and the App that I developed in this project can help with managing this challenge.

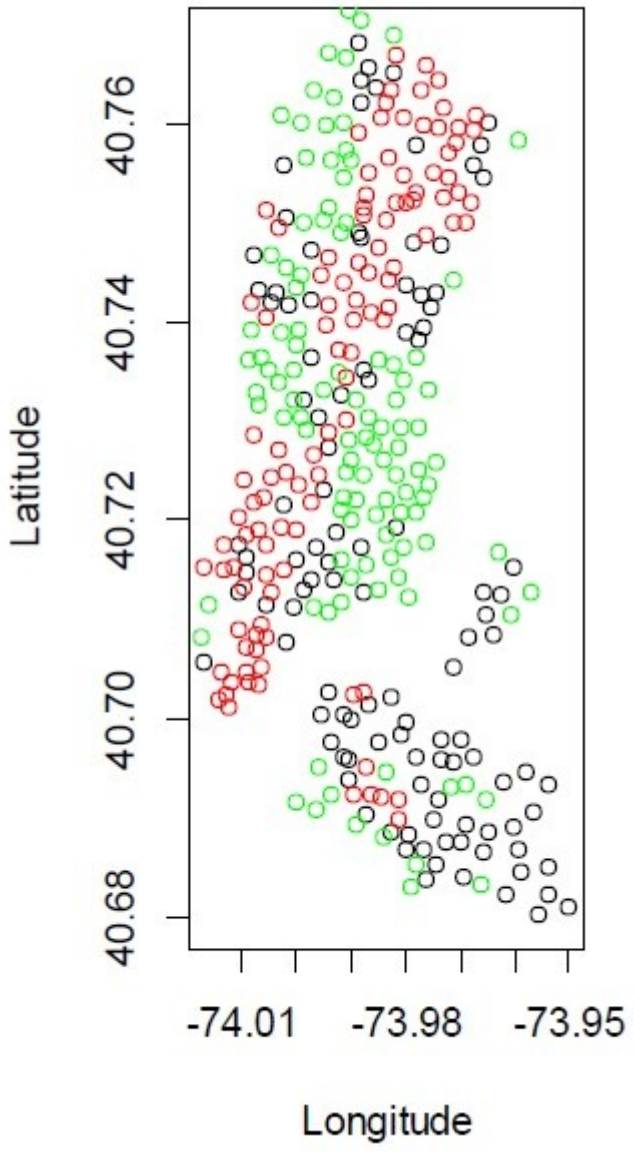

Figure 2: Spatial imbalances during afternoon rush hour on an average summer day (based on data from July 2014). Stations in red are those where more bicycles leave than return, stations in green are those where more return than leave, and stations in black are approximately balanced.

As I showed in the video summary, figure 3 describes this challenge of imbalanced demand at the station level. If the station is full around 6 am then around 9 am it will become empty unless more bicycle are delivered to the station by Citibike employees. However, if they deliver too many bicycles at 9 am, for example they make the station full again, then around 6 pm they need to return to this station and take some bicycles away to prepare a sufficient number of docks for returning bicycles. So, when it comes to managing this challenge it is important to avoid over-corrections of problems.

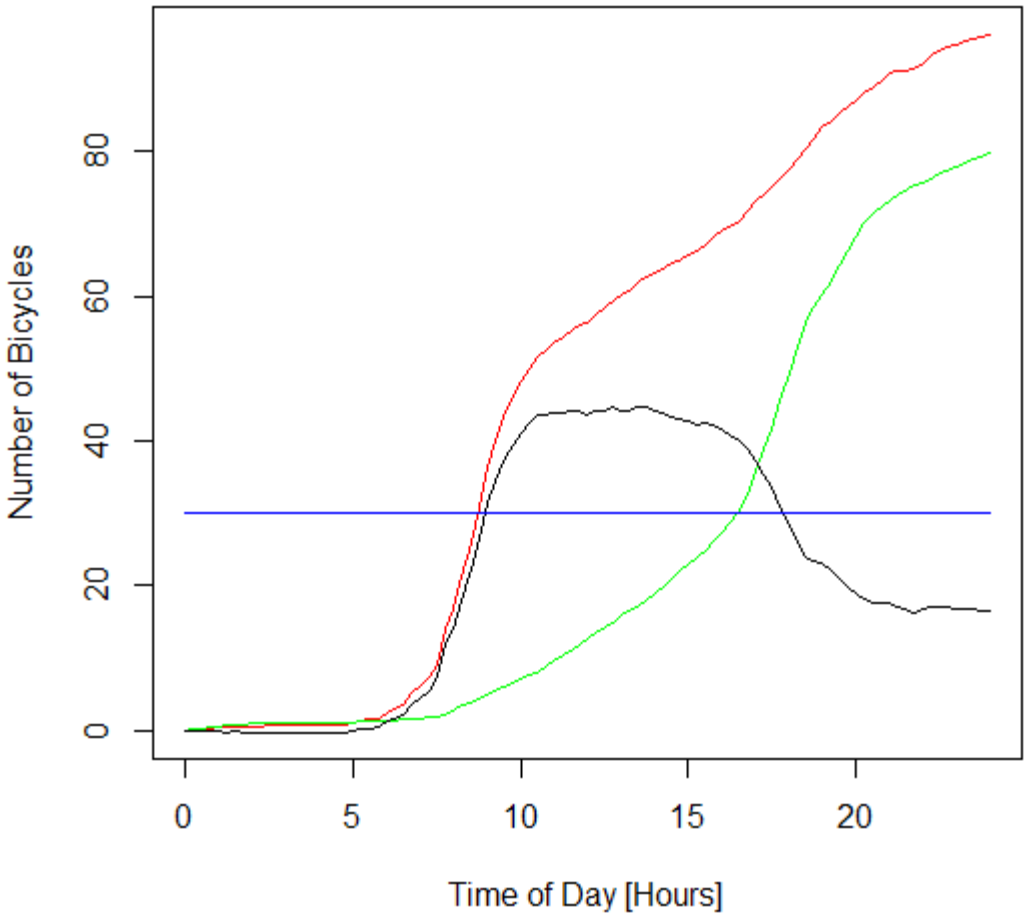

Figure 3: Imbalances at the station level on an average summer work day (based on data from July 2014 for station 502). The red line is the cumulative number of bicycles leaving the station, the green is the same for arriving in the station, and the black line is the difference. The blue horizontal line is the total number of docks in the station.

In my exploratory analysis I also saw the difference in the number of bicycle trip taken between summer and winter, which made me aware of the fact that there is also a strong dependence on the weather. After I became aware of the projects that were done by former Fellows of The Data Incubator, I learned that they too included temperature, rain amounts and snowfall in their analysis and got statistically significant effects, so all that lead me to include weather data in my predictive model – see below.

Predictive Model(s)

The predictive model is actually made of many independent models. I first tried a combination of a predictive model for the total demand for bicycles or docks together with a model of the fractional demand per station, but cross validation showed that an independent model per station gives better results. Actually, there are four models per station: demand for bicycles, demand for docks, and for each of them there are separate models for workdays and for the weekend. Since there are over 300 stations for which the App provides predictions, over 1200 models had to be trained. Hyper-parameters for training the boosted regression trees models were determined by examining several examples at various levels of demand. The optimal value of the interaction depth hyper-parameter as determined by cross validation was two, however the difference compared to an all stumps model was very small, and since there are other sources of error, for example the weather forecast, and since using an all stumps model significantly simplified deployment on Heroku, I decided to use the all stumps model.

The variables that are used in the models are:

- workday or weekend (holidays, for example Labor Day, were also considered weekend days, and were easy to identify, because they did not have the double peaks of demand during morning and afternoon rush hours that are typical of workdays – see figure 1);
- time of day;
- temperature;
- day of year;
- rainfall;
- snowfall.

Decisions about variables, such as the maximum period during which rainfall at a certain time affects demand, were made based on cross validation.

The number of Citibike stations was increased recently. The data that were available for those stations when model training was done were for about one month or less, so these stations appear on the map that is produced by the App, but no predictions are made for them (see discussion in Future Improvements below).

Figure 4 compares model predictions and actual observations for the total demand for docks during a nice spring day, and one can see that the fit is quite good. Of course, the decision to present this specific day was biased, so this fit is better than average. The RMS for all days is 57 (in docks per 15 minutes), whereas for this specific day it is 35, so on average the noise is bigger by about two-thirds than that in figure 4.

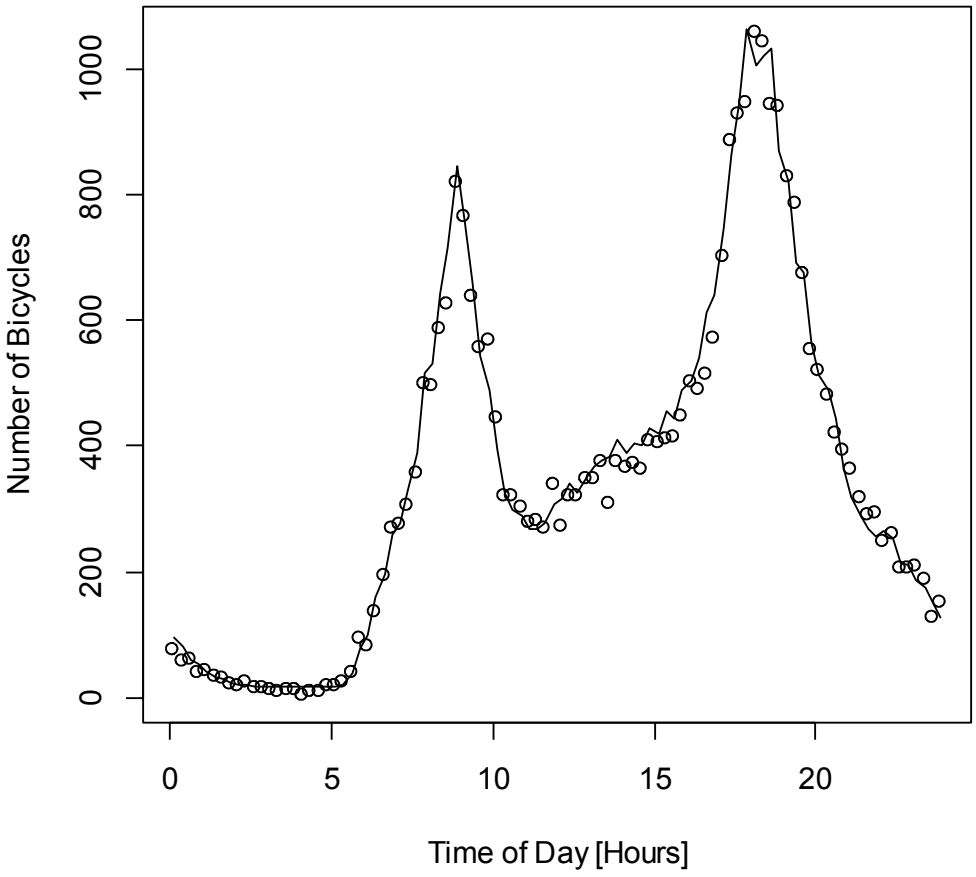

Figure 4: Observed (circles) and predicted (solid line) total demand for docks (per 15 minutes) as a function of the time of day on May 19, 2015.

Figure 5 shows the effects of certain variables on the predicted demand for docks. These effects are presented for four days with unrealistic weather conditions so that it would be easier to see each effect separately. All days are in early spring and with constant temperature, constant rainfall rate and constant snowfall rate throughout the day. The black line is for a day with a temperature of 25 degrees Celsius and no rain, and the red line is for a day with the same temperature and rainfall rate of 10 millimeters per hour. The blue line is for a day with a temperature of 0 degrees Celsius and no snow, and the green line is for a day with the same temperature and snowfall rate of one inch per hour.

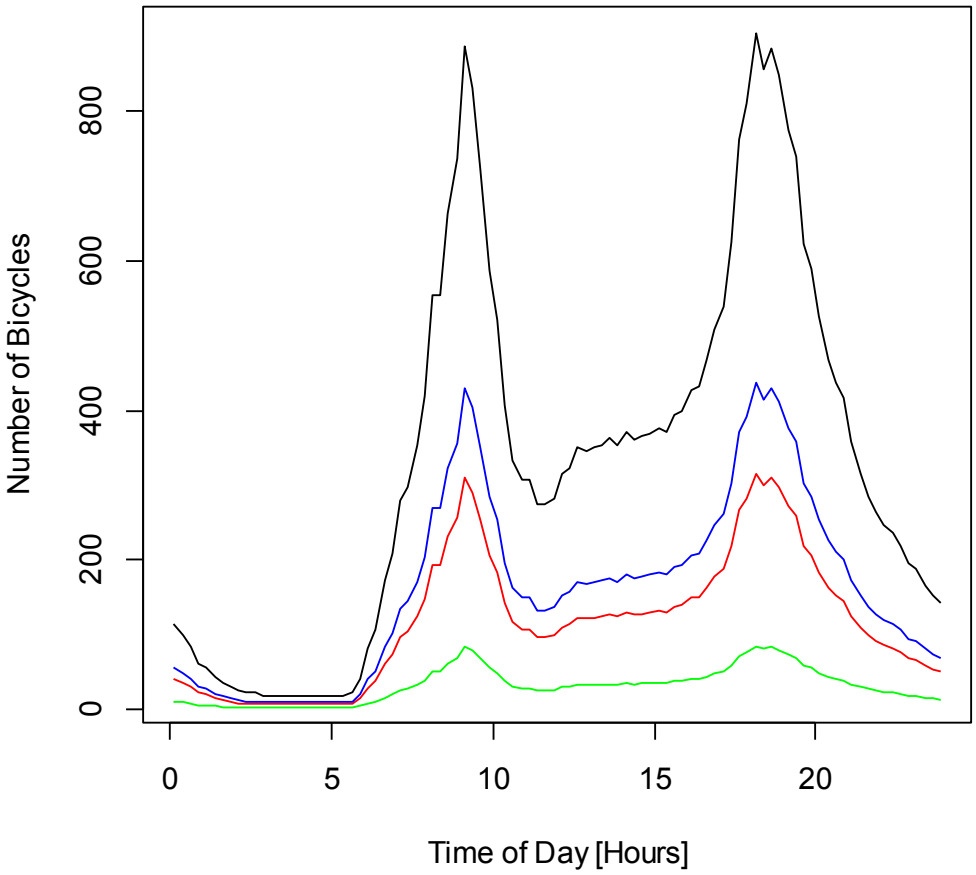

Figure 5: Illustrating the effects of changes in weather conditions on the predicted total demand for docks (per 15 minutes). See explanations in the previous paragraph.

Future Improvements

1. Current predictions distinguish between workdays and weekends automatically, but do not have an automatic way to know that a certain non-weekend day is actually a holiday and should be treated as a weekend day. While training the model this was taken into account manually. So, getting the program to read in real time data about holidays from the web would be an improvement.
2. Currently the model gives predictions up to three hour into the future, and even in that short time more than a third of the stations are predicted to run out of bicycles or docks during afternoon rush hour. So, just for the map that the App creates it does not look like it is necessary to give predictions for longer periods of time. However, as mentioned in the discussion about figure 3, it is important to avoid over-corrections when making more bicycle or docks available at a certain station, so predictions for longer times would make it possible for the App to recommend the best numbers of bicycles or docks that should be made available, and that would lead to improvements in operations.
3. As mentioned, Citibike added many new stations recently, and with the current modeling approach it would take a long time until it would be possible to give good predictions for the new stations. It would be helpful to consider other approaches that could give more reliable predictions faster, for example, using unsupervised learning methods to find old stations that are similar to the new ones and base the predictions on them, possibly by using the ideas on credibility from Actuarial Science.
4. The increase in the number of stations is just one of the trends in the bicycle trips data. A similar trend exists in the number of users, and testing different ways to extrapolate such trends into the future (in addition to re-training the model on more recent data) could lead to an improvement. Again, methods from Actuarial Science may be useful for this.
5. Currently there is about a two month lag between the most recent date about which there is bicycle trip data and the date for which predictions need to be made. It would be great to incorporate this App into the computer systems of Citibike and improve predictions by having access to the most recent bicycle trips data (almost) in real time.