#### **Predicting idle bikes in dockless bike sharing systems**

Katelyn Walker

11/5/18

**Project Description**

Bike shares provide one solution to the “final mile” problem of public transit - helping people to get from their homes, workplaces, or other locations to the nearest transit stop, which may not always be in convenient walking distance. In recent years, dock-based bike shares have been quite successful in major cities, where they receive the heaviest use during commute hours. Dockless bike shares provide a more flexible alternative to dock-based models, and may be well suited to expanding beyond city cores and transit lines. Initial evidence from Washington DC shows that dockless bikes may have different usage patterns than dock-based bikes. Lime bikes in DC received more ridership during afternoons and weekends, indicating a pattern more indicative of leisure than commuting [1]

However, because they are dockless, these bikes can be taken anywhere, and may be left in areas with little foot traffic and few future riders. Bikes which sit idle in quiet areas are not available to riders elsewhere who might need them, and aren’t making money for the company. Identifying when a bike is likely to be left idle would allow the company to target that bike for repositioning.

In addition, areas where bikes are frequently left idle are perhaps undesirable additions to the bikeshare system, or perhaps need to be treated differently for bike sharing to be successful (as an example, in lower income neighborhoods, subsidies and alternative payment methods may be needed to encourage more riders [2]). A next level goal for this project would be to add features to the model that reflect demographic, land use, and topographic information for an area, and see which of those features are good predictors for having less bike idle time. This information could help a company to target their next expansion area.

[As an aside - given my planning and geospatial background, I’m super comfortable working with geospatial analysis on this sort of land use data, so it would be pretty cool to integrate that into machine learning]

The results of this project will be presented in slide/blogpost format with lots of fun maps and visualizations.

**The data and modeling pipeline:**

1. EDA and visualization. Using Geopandas (or perhaps other libraries if you have suggestions - but geopandas seems pretty cool!), make heatmaps of bike idle time. Also make plots (and possibly more heatmaps) of time-of-day usage.
2. Create the target. The target for modelling will be the time an individual bike sits before its next rental (referred to as idle time). From the scraped data, calculate how long a bike sits at a single location.
3. Create the features. This information will come from a few different sources
   1. From within the scraped data, features might include location, bike density in an area, and time of day/week. In order to be an informative feature, I may want to aggregate locations into regions.
   2. Land Use - as an indicator of residential vs. commercial vs. industrial areas, city zoning GIS files will be loaded into Geopandas. Each bike location can then be identified as being in a particular city zone.
   3. Demographic and economic data - from census blocks, also loaded into Geopandas. Using the same procedure as for land use, assign demographic and economic data to each bike location based on the block it falls in.
   4. Topographic data - this would be an interesting one to include (does hilly-ness affect bike shares?), though it would be a more time consuming feature to create. I know I can create a TIN or raster slope model in GIS and then load those blocks into Geopandas, but slope is more about your route than where you are at a given moment. I’m still pondering this one.
4. Fit a model. I plan on trying a variety of regression and tree-based models and seeing what cross validates best. I may also cross validate some ensembling of models.
5. Interpret feature importance, based on coefficients/partial dependence/or whatever the appropriate metric is for the model I end up with.

**A few challenges I anticipate:**

* Calculating the target (bike idle time) is slightly more complicated than it appears, because bikes may be picked up for repositioning or battery charging. I think I may be able to detect when a bike is picked up, but I’m mulling over whether I should adjust my training target value when that happens.
* I’m also going to be learning some new geospatial libraries as part of this capstone, which I’m sure will have plenty of challenges, but that’s part of the fun!

**Libraries I plan on using** (other than what we’ve learned so far):

Geopandas: <http://geopandas.org/index.html>

Geospatial Learn - another library I want to look into (claims to be like sci-kit learn but with geospatial data): <https://pypi.org/project/geospatial-learn/>

Data:

I have two weeks worth of scraped data from Santa Cruz, a week and a half from SF, and about a week from Sacramento (so far - I’m still scraping every minute!). A sample:

,bike\_id,is\_disabled,is\_reserved,jump\_ebike\_battery\_level,lat,lon,name,datetime

0,bike\_23360,0,0,44%,36.97179833333333,-122.02465833333333,1391,1540308547

1,bike\_20799,0,0,63%,36.976315,-121.97773166666667,0958,1540308547

2,bike\_21411,0,0,66%,36.952708333333334,-122.06464166666666,1341,1540308547

3,bike\_23330,0,0,85%,36.983068333333335,-121.99496333333333,1473,1540308547

4,bike\_20790,0,0,91%,36.962896666666666,-122.04521166666666,0887,1540308547

5,bike\_20842,0,0,100%,36.975073333333334,-122.02463166666666,0897,1540308547

6,bike\_23339,0,0,62%,36.97158833333334,-122.02166666666666,1343,1540308547

7,bike\_21415,0,0,37%,37.000098333333334,-122.06282333333333,1358,1540308547

8,bike\_21406,0,0,38%,36.96893166666667,-122.02107,1301,1540308547

9,bike\_23399,0,0,81%,36.98741166666667,-122.03894166666667,1540,1540308547

10,bike\_23335,0,0,69%,36.956015,-122.03936333333333,1404,1540308547

11,bike\_23385,0,0,80%,36.96757,-122.020495,1542,1540308547

12,bike\_20645,0,0,71%,36.958265,-122.03955333333333,0936,1540308547

I also will be getting GIS files from city websites (for zoning) and the US Census Bureau’s TIGER database. These GIS shapefiles then load directly into a Geopandas dataframe. Here is an example of the Santa Cruz zoning data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **OBJECTID** | **ZONING** | **SHAPE\_LENG** | **SHAPEarea** | **SHAPElen** | **geometry** |  |
| **0** | 23 | CC | 0.005469 | 1.125082e+05 | 1772.315833 | POLYGON ((6115648.111000001 1818541.042999998,... |
| **1** | 24 | PK | 0.002634 | 4.084068e+04 | 881.121721 | POLYGON ((6114770.856000006 1817930.811000004,... |
| **2** | 25 | RM | 0.013191 | 5.012602e+05 | 4361.125116 | POLYGON ((6114341.5 1818988.125, 6114368.22599... |
| **3** | 26 | CT | 0.020528 | 1.400059e+06 | 6890.346622 | POLYGON ((6114905.5 1819702.375, 6115071.5 181... |
| **4** | 27 | CC | 0.003188 | 5.038172e+04 | 1046.563196 | POLYGON ((6115484.502000004 1817765.548999995,... |

Santa Cruz zoning: <https://data1-cruzgis.opendata.arcgis.com/datasets/282fb222c0464c7ea832d293b31800ad_42>

San Francisco zoning: <https://data.sfgov.org/City-Infrastructure/Parcels-With-Planning-Department-Zoning/6b2n-v87s>

Sacramento zoning: <http://data.cityofsacramento.org/datasets/zoning/data?geometry=-121.492%2C38.556%2C-121.431%2C38.567>

Census TIGER data: <https://www.census.gov/geo/maps-data/data/tiger.html>

**References:**

[1] <https://grantmckenzie.com/academics/Dockless2018.pdf>

[2] <https://mobilitylab.org/2018/07/18/bikeshare-has-an-equity-problem-and-philadelphia-is-tackling-it/>