ISYE 6501 Course Project
Improving the efficiency of bike-sharing systems in major cities

In the not-too-distant past, renting bikes was only something done by tourists in vacation towns looking for something to do that doesn't involve museums or alcohol. More recently, companies and cities have invested in a network of bicycles for commuters, residents, and other members of the public as an alternative to private and traditional public transportation.

Similar to the electric scooter phenomenon that swept the nation, bike sharing allows a consumer to select a bike, pay a fee for use (either by distance or access), and leave the bike for another user once the destination is reached. Unlike an electric scooter, bikes require being returned to a "dock" or "station". This makes station placement critical: bad placement can lead to long walks for users to get or store bicycles, decreasing interest and use.

Cities and operators also want this system to be worth the investment: it should be regularly used by its citizens to improve their quality of life and fund the system independently. Together, these challenges create opportunities for analytics to assist cities and companies to improve their bike sharing systems. This proposal will outline some of the possible models which could be used to help improve a bikeshare system in a few different key areas, the data necessary for those models, how that data would be collected, and how often the model would be updated to provide useful information.

Model Question 1: Where should stations be placed?

The first assumption made with this question is that stations can be moved: in some scenarios, this may not be possible. However, as outlined earlier, station placement can be one of the largest determinants of system use. It's therefore of critical importance that stations be placed in effective areas. For the purpose of this exercise, the assumption will be made that stations can be moved.

K-means clustering (or another clustering algorithm) is particularly well suited to this problem. Using a set of interest points, clustering can estimate the point which minimizes travel distance from the points of interest. This works particularly well on the assumption that distance is a key determinant of bikeshare use. Additionally, K-means clustering can use a pre-determined number of centers: this could be adjusted slightly, either through manual review or by varying the prescribed number of centers within a range of acceptable options. Theoretically, these points of interest could be weighed based on some inherent factors: number of residents in an apartment building, number of visitors to a museum, number of employees in an office building. This would require access to an additional dataset which may or may not be available.

Optimization modeling would also work well on the same dataset to minimize distance between stations. Constraints could be set based on the number of total stations, distance between docks, distance from docks to specific interest points (such as transit stations), and number of people served. With simple and naïve constraints, this may not look very different from kmeans: however, if we are presented with a situation where more complex constraints are required, using optimization should improve the results of the model. This model may also be good when only the addition of docks is required: previously deployed stations can be modeled as constraints.

Collecting "points of interest" is the key element in this model, as defining a point of interest can be somewhat arbitrary: the need(s) which the bike system is designed to serve (ex. commuters going from transit hubs to office buildings, residents traveling between neighborhoods, etc.) will likely drive decisions on what constitutes a point of interest. The physical location data (X, Y, and Z relative to a defined point) should be collected from a topographic map of the city. However, if other data is desired (such as daily attendees at a museum, or number of riders entering or exiting from a specific subway station) to determine the importance of each point, more data sources would be required. These could be omitted in the initial revision of the model before being integrated if necessary. Additionally, if a bikeshare system was already implemented, it may be possible to estimate the value of each point based on ridership at specific entry and exit points of the current system: the downside of this methodology may be incorrectly grading points not well served by the active stations.

While the data may be tricky to collect, it does not need to be frequently updated. As riders plan their lives, it becomes increasingly costly to move stations and adjust other infrastructure to accommodate changes. Therefore, although the model may need to be run multiple times before conclusions can be drawn (to generate a variety of random cluster starting locations and possible answers), once a determination has been reached, the model will not need to be re-run unless further changes to station locations are desired.

Model Question 2: When will bikes break down and require repair or replacement?
Just like all mechanical creations, bicycles can be broken. This can be mitigated by regular maintenance, but occasionally major unexpected damage can occur. It's also possible for bicycles to be stolen or vandalized, requiring replacement. Anticipating these issues is key to a well functioning bikeshare system: a lack of available and trustworthy vehicles can result in riders walking away from the system temporarily or indefinitely.

Collecting usage data from the bicycles would be very straightforward. GPS units on each bicycle can use positional information gathered at roughly 1Hz to determine real-time position, and therefore also calculate speed of travel and total distance. Along with maintenance data logged by employees (such as when the last time the bike was inspected), logistic regression models could be built for each component. A logistic regression model would be a good selection because it outputs a probability: in this case, the probability of failure on the bicycle. Any bike with a probability above a preset threshold would flagged for maintenance. These probabilities can also be weighted based on importance (e.g. the bike chain disconnecting is much more likely to be problematic than a broken pedal reflector) and a score can be assigned, allowing bikes to be prioritized for maintenance. Increasing the sampling rate of the GPS may also allow for tracking of acceleration data, allowing for better modeling of brake use, a particularly catastrophic failure mode.

While some issues may result in a customer calling in a problem (such as a crash or collision) other issues may be ignored or unreported. However, a change detection model can be used to identify bikes that may have suffered unexpected damage but are not reported. Under the assumption that riders will only pick bikes that they like and trust based on all features performing as expected or appearing correctly, checkout history data can be used to identify bikes that are frequently passed over and report a possible issue. Using a hypothesis that all properly functioning bikes are used at a rate within a predictable distribution (possibly normal, but this should be confirmed with testing or simulation), a bike being used less frequently could become noticeable. A change detection model could be trained on the combination of this past usage data and maintenance logs to find similar events and point field operatives to examine the bike for possible defect. A change detection model with cyclic factors would also be able to anticipate regular changes in demand (such as weekends, holidays, or seasons).

These models would need to be run regularly by maintenance personnel. A change detection model would ideally be running for each bike in real-time, although this could be limited and batched each night. The logistic regression models would only need to be run when retraining: otherwise, input data from each of the bikes (like miles traveled since last maintenance) could be plugged into the regression equation to produce a probability and decision.

Model Question 3: How many bikes should be at each station to accommodate demand?
Every time a customer approaches a bike dock, at least one bicycle should be available for them to choose. If there aren't any from which to choose, the customer will either wait, possibly for a significant amount of time, or the customer will leave and choose an alternative form of transportation.

Lack of space at a bike dock would be equally, and possibly more, problematic. Without locations to end their rides, riders may resort to a more distant finish location to their end goal. This could decrease further ridership as possible customers have negative experiences while riding. Frustrated or delinquent riders could also leave bikes scattered and not return them at all, possibly increasing damage to the bicycle and removing part of the system capacity. This makes setting appropriate bike and dock capacity crucial.

Prescriptive simulation would very effectively model this situation. Using the known station locations and models for estimating bike maintenance combined with estimated distributions for arrivals and travel times, we can collect extremely granular (possibly to the minute) estimates of bike locations across the network. These estimates can then be used to determine the appropriate number of bikes to be stationed at any given location.

Estimating distributions for both arrivals and travel times are both imperative to a good simulation and somewhat complicated by the changing and addition of dock locations. However, by using data collected from the bikes in a previous implementation of the system, bike checkout times can be discretized into minutes in each day (possibly separated by weekday/weekend/holiday, if necessary) and smoothed to create an estimated distribution for each former station. From there, the distributions can be combined, either naively or using a weighted average, to estimate the distribution for the new station location. Travel time distributions would be estimated using a combination of speed and GPS data from previous rides. In both cases, prior knowledge about the station's neighborhood could be key: residential neighborhoods may experience different patterns than tourist or corporate regions. In the distribution estimation phase, these classifications should be monitored since they could be used to simplify the model if the labels can be used to generalize distributions for similar dock locations.

For the simulation, the bike maintenance model would also require input from the change detection model: this would help simulate major unexpected damage. Additionally, the maintenance model would require probability estimates of bikes being stolen or catastrophically damaged during or at the end of the ride in different situations (abandoned by frustrated customer who can't find a dock versus a traffic accident mid-ride).

For most of this modeling, previous bikeshare ride data would be required, and future bikeshare ride data would need to be collected. A brand-new system would struggle with a lack of history from which to pull insights. However, an alternative could be to use electric scooter travel data, and this similarly mimics travel data and data collection techniques for a bikeshare system. Alternatively, data could be collected on bike use rates along a random sample of streets by monitoring traffic live or remotely.

The simulation would require iterative updates based on incoming data. With new stations, creating a perfectly descriptive simulation will be extremely difficult. Even with a perfect simulation, it would be impossible to know until the stations were put in place. Additionally, traffic patterns, construction, and weather changes could all impact the accuracy of the new simulation. Therefore, as ridership data is collected with new station locations and user habits, the simulation and distributions should be modified to reflect the new realities.

Conclusions

The bikeshare systems across the world have the potential to improve transit options for every resident, commuter, and visitor to the cities which operate them while sustainably supporting the service. Using the optimized K-means clustering, logistic regression, change detection, and simulation models described in this report, the current state of bike sharing can be improved dramatically.