Project Picked

Increasing-Bike-Share-Efficiency
https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Increasing-Bike-Share-Efficiency

Background

Increasing Bike share efficiency is chosen based on my personal experience. As a public transport user to commute to university, one of the issues I had to face is the last 2 mile problem. At that time I had to walk to the university. This was in 2010 and at that time the Bike sharing was not prominent.

Today in cities Bike sharing is really prominent. To be widely adopted the bike sharing service provided by the companies should be reliable, accessible, and requires a conceptualized network of stations with enough docks for the service to run smoothly even during rush hours. From analysis it could be seen that from 2017 onwards bike sharing is being widely used in Los Angeles.

Bike share principles are the bikes can be borrowed from any docking station across the city. The charge is per minute and the first 30 minutes is free for registered customers.
In this case study I would like to focus on discussing models that will help to increase bike share efficiency. Focus will be a big city , Los Angeles, where bike sharing schemes have started and companies are looking for improvement and optimization.

Problem Statement

In this case study the focus will be to use data analytics and models to increase the Bike share efficiency of the cities, Los Angeles, where the people are environment friendly and chose to use bikes companies like Uber. The companies that provide the bike sharing facilities in these busy cities would have to consider the following

1. Categorize the district in the city as bikeable or non bikeable
2. Optimal number of docking stations based on the demographics that will use the bikes. - This is a major deciding factor for bike share success.
3. Per docking station how many bikes are required on any given day of the week to maintain adequate inventory of bikes
4. Take into consideration the discrete points at a given time period or hour where demand and supply problems occur.
   a. Docking station may be full when a customer brings back the bike.
   b. During rush hour a customer wants to use the bike, but the docking station is empty.
5. User experience and the company revenue can be limited when the origin station is empty, which forces the user to either resort to another means of transport or try to find an available bike in another station. Similarly, if the destination station is full, the user must either wait until one bike is picked up, or return the bike to another station with at least one parking spot available.

This document will focus on analytics and data models used to make good recommendations to the bike sharing company.

Data Analysis and Solution Approach

At a very high level solution approach followed is below

1. Clustering model - To identify bikeable and non bikeable districts
2. Linear regression - To find the optimal number of docking stations
3. Graph theory and topology - To find the optimum location of the docking station
4. Regression with decision trees. - To Maintain adequate inventory of bikes at a docking station at any point in time
5. queuing/simulation theory. To model the incoming and outgoing bikes behavior in a docking station during peak hour behavior.
6. To avoid model performance deterioration real time data will be collected and model will be rerun against the data on a periodic basis.

Clustering model to segment districts in to clusters that bikeable or non bikeable

To understand the bikeability of a city the thought process during data gathering should be bicycle comfort, suitability, friendliness and accessibility based on district geography and demographics. The analysis should also include data on user preference and user patterns.

Given data for these variables for all districts in the county of Los Angeles

SL No	Variables	Data type	Description
1	Latitude	Decimal	Latitude coordinates -> city location
2	Longitude	Decimal	Longitude coordinate -> city location
3	No: of Medical facilities	Integer	
4	Walk Score	Integer variable	Based on the demographics the number of people walks to nearby destination
5	Bike score	Integer variable	Based on the demographics the number of people bikes to nearby destination
6	Public transport score	Integer Variable	Based on demographics number of people in the city taking the pubic transport

7	No of grocery store	Integer variable	This provides a bikeable index based on how near to a housing community the grocery store is
8	No of restaurants	Integer variable	Help to group tourist as well as people sho bike during weekends and have breakfast etc
9	Shopping	Integer variable	This provides a bikeable index based on how near to a housing community or public transport the Shopping is
10	Employment Opportunities	Integer variable	Job opportunities in the city.
11	No of Libraries	Integer variable	Libraries
12	No of Universities	Integer variable	Bicycle sharing users student. Rental for commuting to workplaces and colleges on a daily basis
13	No of Tourist access spots	Integer variable	As the number of spots increase the tourist will use the bikes for recreational rides
14	No of facilities	Integer variable	No Miscellaneous facilities that have a given density threshold of interaction.
15	Public Transportation Infrastructure No of train stations	Integer variable	No of train stations in the district
16	Public Transportation	Integer variable	No of bus stations in the district

	Infrastructure No of bus stations		
17	Public Transportation Infrastructure interconnect score.	Integer variable	Interconnection score between public transportation
18	Bike lane access score near public transportation	Integer variable	
19	No of dedicated bike lanes in the district	Integer variable	
20	No of Roads with bike lanes	Integer variable	
21	Access to Bike lanes acore	Integer variable	
22	Safety and Security Score		This will be crime score in the district, How safe is to use public transportation and bikes
23	user preferences		1) short bike route with 2) recreational cyclists

Additional notes on data: If a district is more than 2 miles radius during data preparation the data points should be split such that latitude and longitude for a data point covers only 2 miles radius.

Use a **clustering** algorithm

To group or segment the district based on is bikeable or is not bikeable. Analyzing the results of this clustering model will help to reveal hidden information.

Data Collection

Data for the variable could be collected from data research companies, cities, real estate companies , using people surveys etc.

There are government websites that provide the public transportation score, walking score, biking score etc. Because of the amount of data we need for this purpose we may have to buy data from data research companies. One other resource to get the information on No of grocery stores, No of Libraries, No of Tourist access spots etc could be collected from the business license data from the district/city.

As the data involves different facets of a district, geography , demography, employment, public transportation etc, the data collection would be in different stages. Data preparation is really the first step before we approach bike sharing scheme modeling.

Optimal number and location of docking stations based on the demographics that will use the bikes.

Optimal number and location of docking stations within a given mile radius is really important in the success of bike sharing schemes.

Given the prediction {bikeable districts from the clustering model} for the districts,

Given the data for the below variables from 2017 onwards from different companies that provide bike sharing schemes for different major cities like San Francisco , Newyork , Los Angeles etc.

SL No	Variables	Data type	Description
1	Latitude	Decimal	Latitude coordinates -> city location
2	Longitude	Decimal	Longitude coordinate -> city location

3	No: of Medical facilities	Integer	
4	Walk Score	Integer	Based on the demographics the number of people walks to nearby destination
5	Bike score	Integer	Based on the demographics the number of people bikes to nearby destination
6	Public transport score	Integer	Based on demographics number of people in the city taking the pubic transport
7	No of grocery store	Integer	This provides a bikeable index based on how near to a housing community the grocery store is
8	No of restaurants	Integer	Help to group tourist as well as people sho bike during weekends and have breakfast etc
9	Shopping	Integer	This provides a bikeable index based on how near to a housing community or public transport the Shopping is
10	Employment Opportunities	Integer	Job opportunities in the city.
11	No of Libraries	Integer	Libraries
12	No of Universities	Integer	Bicycle sharing users student.

			Rental for commuting to workplaces and colleges on a daily basis
13	No of Tourist access spots	Integer	As the number of spots increase the tourist will use the bikes for recreational rides
14	No of facilities	Integer	No Miscellaneous facilities that have a given density threshold of interaction.
15	Public Transportation Infrastructure No of train stations	Integer	No of train stations in the district
16	Public Transportation Infrastructure No of bus stations	Integer	No of bus stations in the district
17	Public Transportation Infrastructure interconnect score.	Integer	Interconnection score between public transportation
18	Bike lane access score near public transportation	Integer	
19	No of dedicated bike lanes in the district	Integer	
20	No of Roads with bike lanes	Integer	
21	Access to Bike lanes	Integer	

	acore		
22	Safety and Security Score	Integer	This will be crime score in the district, How safe is to use public transportation and bikes
23	user preferences	Integer	2) short bike route with 2) recreational cyclists
24	No of registered users	Integer	
25	No of casual riders	Integer	
26	Self-selection score	Integer	People tendency to live in a neighborhood with bike access
27	Density of population	Integer	
28	Data point from the GPS installed in the bike share companies	Integer	
24	**Number of docking stations per 2 mile radius.**	Integer	This is the **response** variable in the sample data collected.

Use

Linear regression

to predict the number of docking stations required per bikeable district in the city

Given docking stations are available from the prediction,

Use network analysis in **Graph theory and topology**

To find the **optimum location of the docking station.** With the geographic and demographic data , the type of network analysis, network partition could be used to divide the 2 mile radius of the district into zones or categories based on proximity to specific points in a network. This result could be used in the placement of docking stations

Data collection

To be collected from different bike providers . The bike providers will have collected data using GPS and sensors in the bikes and docking stations. This data may need to be requested from the providers or gathered from the data research companies.The data could be from major cities that have prominent bike sharing services running

Maintain adequate inventory of bikes

Once the number of stations required in a district is identified, Next step in the modeling will be per the demand at a docking station, at any point of time the scenario of the docking station being full or empty should be avoided.
This analysis should be performed in docking stations where there is a general demand and supply issue.

Given data on variables collected from 300 to 1000 m around each docking station.

SL No	Variables	Data type	Description
1	Population	Integer	Population density
2	Number of jobs	Integer	Number of jobs
3	Number of students in campus	Integer	Number of student in university at any point in time

4	number of student residences near a station	Integer	
5	number of people using public transport near the docking station	Integer	
6	no of shopping malls	Integer	
7	No of tourist access spots	Integer	
8	no of tourist people visiting	Integer	
9	Hours	Integer	0-23
10	Day of the week	Categorical Variable	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
11	Morning Peak hours 6:00 AM -10:00 AM	Boolean	
12	Evening Peak Hours 3:00 PM - 6:00 PM	Boolean	
13	Hours 10:00 AM-3:00 PM	Boolean	
14	Weather	Categorical	Spring, Summer, Fall, Winter
15	Weekend or Week day	Boolean	

16	Number of bikes ready for servicing	Integer	
17	Number of arrivals	Integer	
18	Number of departures	Integer	
19	Number of registered bikers	Integer	
20	Number of casual riders	Integer	
21	Number of bikes in the docking station	Integer	**response**

The data is time series data. This data will have an increasing trend and seasonal effect. As a part of initial data preparation the Exponential smoothing could be used to smooth the data. As there are many variables during data preparation identify the collinearity among the features.

Use **Regression with decision trees**. Branching strategy should be there because from the analysis it could be seen that per season , per week data or weekend, per peak hour or no peak hour required no of bikes on any day will be different. Going forward with a branching base model will help us to capture the different predictions.

Examples of branching would be

1. One branch summer, week day , peak hour evening
2. One branch summer, week day , peak hour Morning
3. One branch summer, week day , non peak hours

To Predict

Number of bikes required in a docking station at different time intervals , using the arrival and departure of bikes given any date of the year.

Arrival and departure of bikes to and from the biking station could be modeled using queuing/simulation theory.

We can simulate a model to watch the incoming and outgoing bikes behavior in a docking station during peak hour behavior. The discrete point for this simulation will be

1. Someone brings back the bike
2. Someone wants to make use of bike sharing

And could be considered as a stochastic simulation with randomness.

Given prediction from the linear regression model {adequate inventory of bikes} the response follows the poisson distribution for arrival and departure of bikes in the docking station
In the docking station the is capacity k .
No of bikes available at docking station at any hour = xi(t) element of {0..k}

xi(t) >0 -> make use of bike
xi(t)<k -> bring back the empty bike

Arrival rates of the bikes is i/per hour
Departure rate of the bikes is j per hour
Ruote time of a bike is average of z minutes

Use **Arena Software to build a simulation system** varying the number of bikes in the docking station

To predict number of bikes require per hour so that average user experience is not hampered due to

1. There is no shared bike sin the station when a user arrives
2. There is no parking space in the bike station when a user arrives

Data Collection

Data for simulation model could be collected from bike sharing providers who publish data in their apps about the availability of bikes and empty parking spots which could be used for modeling.

Data Collection and Preparation

Could be collected from bike sites which record the bike sharing trips from major cities. Per the analysis for this project it could be seen that more than 1000 cities in 60 countries have adopted the bike sharing programs. There is a cost factor to get this data from different companies who own bike sharing systems but it could be achievable given the advantages of creating the model. Automated data can be collected from the docking station . how many times per day the bike was taken and where was the ending destination etc. Service provide firms can also gather dataset using meteorological surveys and people's lifestyles.

During data preparation for the linear regression models, optimal number of docking stations and maintaining adequate inventory of bikes variable selection techniques like stepwise regression for initial variables analysis and Lasso and Elastic net should be considered for variable selection. This analysis will provide the significant variables for predicting demand for shared bikes.

Model Refresh and Rerun

Frequency of refresh of clustering model to segment the places in the city as bikeable or non bikeable should be based on the business related or transport related development in the city.

Let us see that the city is going through business related or transport related development then the clustering model to segment the places in the city as bikeable or non bikeable could be run at least once a month.

For the linear regression models , Optimal number and location of docking stations based on the demographics that will use the bikes, and Maintain adequate inventory of bikes, an automated job should be set up that runs the prediction with real time data at a given time interval. Based on the prediction results from the automated jobs and the output interpreters like R-squared, AIC, BIC etc the model rerun could be planned.

For simulation model refresh and rerun we should closely monitor the user data and dynamics in the biking station of interest. Especially user complaints lodged in the com[any website, review etc. If users are going through the issue of bikes not available in a docking station during departure or no empty docks during arrival then this simulation model should rerun frequently to get the optimal number. At this point in time the docking station has a lot of randomness and variability.

Improvements

Analyze if any other models are more suitable for each step. As we start collecting real time data some variable data will be more analyzed and could be done to find a least expensive variable that correlates the data. As we learn the model more, find the features that explain more variance and provide more weightage to that coefficient. Develop domain knowledge.
Tune the parameters for each algorithm used for modeling.

References

https://tram.mcgill.ca/Teaching/srp/documents/MarkO.pdf
https://www.huduser.gov/portal/sites/default/files/pdf/Creating-Walkable-Bikeable-Communities.pdf

https://bikesharingworldmap.com/reports/bswm_mid2021report.pdf