Where to Put an Advertisment

Shuo Yuan Chang, Chunchu Liu, Yuke Liu, I-Ling Yeh, Ruhao Zhuang

December 9, 2019

Business Understanding

Concerning devising proper business strategies, advertising, (along with marketing), is orchestrated to maximize revenues and profits for companies that aim to actualize business in their operations. In this project, let's assume that we are an advertising agency.

As the fact of 47% of Americans say they would be more likely to ride a bike if pathways were physically separated from motor vehicles, our primary goal is to analyze the dataset related to the bike-sharing information to enable us to pinpoint "the place" having the prospective and existing customers for the particular company. Firstly, we explicitly identify "the most-frequently-used bike route." Furthermore, two of the crucial parameters – gender and age – were also carefully examined to implement the most proper market segmentation (the customer preference). With all the information available to us, we are able to help the company boost the numbers on its sales sheets.

Data Understanding

This dataset that we will be used for the final project was found on Kaggle (https://www.kaggle.com/samratp/bikeshare-analysis#NYC-CitiBike-2016.csv). The data was collected from CitiBike, a bicycle sharing system offered only in New York City.

Fifteen attributes that are included in this dataset, which are tripduration, starttime, stoptime, start station id, start station name, start station latitude, start station longitude, end station id, end station name, end station latitude, end station longitude, bikeid, userype, birth year and gender.

Table 1: The Outputof the NYC-Citibike-2016.csv

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype	birth year	gender
0	839	1/1/2016 00:09:55	1/1/2016 00:23:54	532	S 5 Pl & S 4 St	40.710451	-73.960876	401	Allen St & Rivington St	40.720196	-73.989978	17109	Customer	NaN	0
1	686	1/1/2016 00:21:17	1/1/2016 00:32:44	3143	5 Ave & E 78 St	40.776829	-73.963888	3132	E 59 St & Madison Ave	40.763505	-73.971092	23514	Subscriber	1960.0	1
2	315	1/1/2016 00:33:11	1/1/2016 00:38:26	3164	Columbus Ave & W 72 St	40.777057	-73.978985	3178	Riverside Dr & W 78 St	40.784145	-73.983625	14536	Subscriber	1971.0	1
3	739	1/1/2016 00:40:51	1/1/2016 00:53:11	223	W 13 St & 7 Ave	40.737815	-73.999947	276	Duane St & Greenwich St	40.717488	-74.010455	24062	Subscriber	1969.0	1
4	1253	1/1/2016 00:44:16	1/1/2016 01:05:09	484	W 44 St & 5 Ave	40.755003	-73.980144	151	Cleveland Pl & Spring St	40.722104	-73.997249	16380	Customer	NaN	0
5	525	1/1/2016 00:47:07	1/1/2016 00:55:52	474	5 Ave & E 29 St	40.745168	-73.986831	470	W 20 St & 8 Ave	40.743453	-74.000040	22823	Subscriber	1975.0	2

Trip duration of this dataset counted in seconds that how long does one user had rent this bicycle in total during that trip. In this dataset, two temporal attributes that indicate the specific time and date of when people start the trip and when they end up the trip called starttime and stoptime. The geological information is also included in this dataset which are start station latitude, start station longitude, end station latitude and end station longitude used to record the specific geological information where customers rent and returned the bicycles. To identify the different stations, based on the longitude and latitude, the dataset had also recorded the name and given the corresponding station id. The dataset also included bikeid that keeps track of all the shared-bicycles. Besides, three attributes that are related to customers: usertype, gender and birth year.

Based on the undefined usertype, which was recorded as 'Customer' in the usertype column, lead to the missing user information of their gender and birth, which shown as '0' in gender, 'NaN' in birth year. This situation leads to the abstraction of the dataset when applied to Models later.

Table 2: The Attributes Type

tripduration	61- 2363758
starttime	1/1/2016 00:09:55 - 12/31/2016 23:53:42
stoptime	1/1/2016 00:23:54 - 01/01/2017 00:14:41
start station id, end station id	72-3440
start station name, end station name	variables of characters
start station latitude	40.44535 - 40.804213
start station longitude	-74.01713445 - -73.92989109999999
end station latitude	0 - 40.804213
end station longitude	-74 - 0
bikeid	14529 - 27327
user type	subscriber / customer
birth year	'NaN'& 1885-2000
gender	0 - 2

Data Preparation

We accomplished the data cleaning before fully implement the analytics with this dataset.

To efficiently use the resources of the advertisement, in another way to reduce the cost, we need an unambiguous data of the specific time of people cycling by the advertisement board, and how old are they and also the gender affect the advertisement type equivalently. Thus, the frequent month and hours are important, and also the age of the users is also important for us. Hence the original dataset is based on the trip records from the start station to the end station, the repeating data such as Start station name and end station name are somehow useless.

First, we eliminated the columns recorded bikeid hence this attribute was regarded as useless and distracting. We kept all the information that is related to the stations except the station names. For each corresponding latitude and longitude, we merged them into station latitude and station longitude. Due to the trip track in the old dataset, we added two attributes called start frequency and stop frequency to esteem the frequent times. To count for the frequency of this station as a start station or an end station, we counted the frequent month and specific hour in the day to that associated station as well.

Based on the original dataset, we obtained the information of user type, users' birth year and gender. Meanwhile, the dataset also shows that these three attributes are affiliated. The birth year equals to NaN or gender equals to 0 are customers in the user type. Therefore, we dropped all the rows that user type equals to 'customer'. Then we discarded the gender and age group and calculated the gender ratio in each specific station and the average age.

The final dataset that we use to analyze shows in the table below.

Table 3: The Output of NYC-Citi-bike.csv After Cleaning

station name	start fre	end fre	frequent month	frequent hour	gendra	frequent gender	age	station latitude	station longitude
1 Ave & E 15 St	567	513	5	8	0.379562	1	37.0	40.732219	-73.981656
1 Ave & E 16 St	479	491	8	8	0.425595	1	36.0	40.732219	-73.981656
1 Ave & E 18 St	797	758	11	8	0.367067	1	35.0	40.733812	-73.980544
1 Ave & E 30 St	870	847	8	8	0.344668	1	38.0	40.741444	-73.975361
1 Ave & E 44 St	498	518	7	17	0.245000	1	43.0	40.750020	-73.969053

Modeling

As an advertisement company, we want to make a price in each location and know what features in each location. We set up the model to group the location where have the same feature. According to different groups, we can make a price to achieve the profit maximum and target the right customer to decrease the cost. To identify the location in which class they belong to, we choose the cluster model.

How did we choose the cluster model finally? In the beginning, we use the regression decision tree to find each group's features. In the regression decision tree, we cannot find the obviously feature in each group. Because we want to make a price in each location depending on the frequency of people borrow a bike in each location, so we try to separate these locations depend on frequency to several levels and use a classification decision tree. We can find each group's feature using a classification decision tree, but there is a problem that we did not know these groups are the optimal way to separate. Eventually, we choose the cluster model and then classification into four groups. We don't have the target variable in our dataset, want to find the optimal breakpoint to separate the location, and each group has the same feature let us easily target the right customers.

We need to guess and find the optimal number of the group by using the cluster model, so we use the silhouette_score to calculate the distance of each group center. We try to separate from two groups to seven groups to calculate the score. Figure 2 shows the silhouette_score for each number of the group, from two groups to seven groups. According to the figure, separating two groups can get a higher score. As can be seen, the score went continuously down after separating four groups. We want to group located in a small range of various, so we choose to divide into four groups.

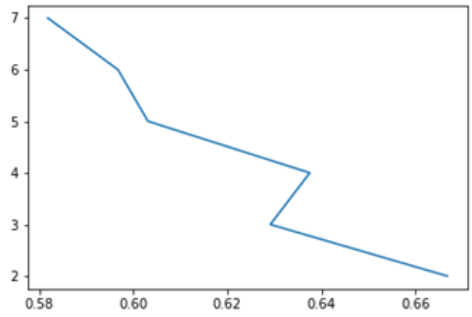

Figure 1: the silhouette_score for each number of group

According to the result of analyzing, we get the benchmark to separate these locations into four groups. We get the mean of frequency in each group's area: 105, 890, 458, 1579, then we use the benchmark to divide into four groups. Table 4 displays each group's features. According to the table, the frequency is perfectly to divide into four groups. As can be seen, each group has a specific hour and age. If a company wants to target customers whose age is within 37 to 41 years old. We can recommend it to make an advertisement in the group4 area during the afternoon at five and six O'clock. This result can let user easily to target the right customer.

Table 4: the result of analyze

	Group1 (0)	Group2 (2)	Group3 (1)	Group4 (3)
Frequency	2-280	284-823	837-1223	1294-2778
Age	30-50	33-44	34-43	37-41
hour	6-23	7-9/16-19	6-9/12/17-18	17-18

According to the station latitude, longitude to draw each group's area. Figure 3 shows each group's location. As can be seen, the more frequency location is closer to the city center. We can use the group distribution to make a price for each group to achieve profit maximum.

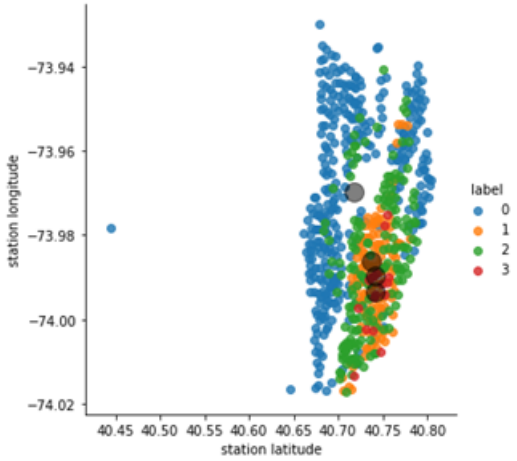

Figure 2: group distribution

Evaluation

We will evaluate our results from 2 angles, model evaluation and business evaluation.

First, we will evaluate the model :

If we just use the traditional method like classification, we can just compute the frequency of each location. So, If we want to determine the advertisement price for each location using the decision tree, we need to manually determine the price of each location based on people's frequency of each location. But the amount of bikes' location is very big. Figure 4 is a decision tree to classify locations.

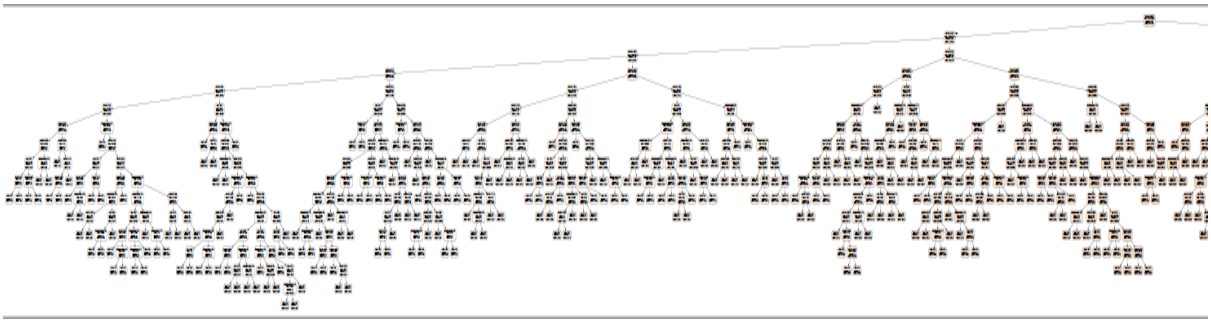

Figure 3: decision tree

The tree is very complex and indirectly, there are too many branches and leaves in this tree, this is because there are so many locations need to be classified. So, we think using this way will make the pricing strategy more complex.

But in our cluster model, we can group the locations first and then fine-tuning the result. After that, we draw the decision tree and set prices based on each group. From Figure 5, we can see the tree is

more direct to analyze. So, it is much easier to set prices in every location. Our cluster model is more appropriate than the classification model. We can work out a better pricing strategy based on our cluster model, so it will help us to make more profits.

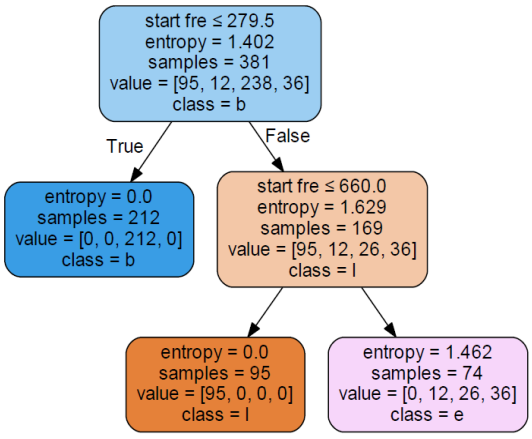

Figure 4: decision tree

And then, we evaluate the model from the business angle:
The current method to detect the frequency of people is through the passenger flow counter. But using this device cost numerous money because we must put a lot of counters in different locations if we want to detect the flow of people at every spot. And this way has another disadvantage. You know, if we want to determine the age and gender of the people, we need to use some other technologies to process our data like face recognition using neural networks. So it will make the process of monitoring people flow more complex. And the cost will be increased if we use a lot of counters in different locations. So, we used the shared-bicycle data, we can easily calculate which age group has a higher frequency in a certain place. And with these results, we can target advertisements for people of a specific age more purposefully.

Deployment

Now that we have our model, we can deploy it to decide where to put advertisements. Assume that we are going to advertising for a man's suit company. Firstly, we divided the downtown of NYC into four parts, which is based on the frequency of bicycles arriving at each location. In each part, we can classify people by age, gender, passing frequency, and distribution period. With these classifications, we can put a specific advertisement into a certain location, which can help us get more profit with less cost. For the problem at the beginning, we need to figure out what kind of people are our target customers. Our decision is men aged from 35 to 41. With our model, we can get the distribution of our target customers. The result is, group1 18%, group2 30%, group3 41%, group4 11%. We now know that our target customers mostly appear in group2 and 3. To reduce the cost of publicity and also get a good profit, we can only put this advertisement in group2 and 3. Correspondingly, our prices for advertising in these two areas are much higher.
Our model can be used more than just advertising. Because our goal is classifying users by region, to make the original dataset fitting our model, we drop many attributes which are useless at this stage, like user type, bike id, trip duration, and so on. If we collect more useful information, our model can make a more specific decision. For example, if we get some effective data of user type, we can roughly get the consumption level of users in each group, then we can put more specific advertisements. Besides, if we want to choose a new bike station to relieve the pressure of bicycle demand, we can analytic the trip duration between two bike stations.