Predicting Customer Traffic for Coffee Shops

Edwin Sutrisno

June 21, 2020

1 Business Background

“Location, location, location” is the three-word mantra we often hear when considering how to make our investment run for the money. In this article, we look at a problem of deciding the best location to open a coffee shop so that it has the highest chance of getting high customer traffic. We know that location is important, but what factors specifically make a location a good location. To answer that question, we will study some data collected from the most successful coffee shops in West Kuala Lumpur area and try to use a machine learning technique to extract the driving factors behind their success.

2 Data Collection

The main source of the data being used for this study is obtained from the Foursquare's API called Places. From the places API, we obtained locations of several successful coffee shops that become the subject of our study, their customer rating count to represent their traffic, and their neighboring venues. The location of interest in this study is the West Kuala Lumpur area. Because I am a local resident, I can determine that the most successful names for coffee shop in the city are Starbucks and Coffee Bean & Tea Leaf. These two brands will serve as our main subjects.

The data gathering process is broken down into these 3 steps:

1. Get all "Starbucks" and "Coffee Bean & Tea Leaf" locations within 4 km radius from a coordinate in West Kuala Lumpur, Malaysia area (latitude 3.1871 and longitude 101.6276)
2. Get their customer traffic data
3. Get their nearby venues within 120 m radius

With the three pieces of information above, we can start grouping locations for their levels of business traffic as marked by their customer rating count and the categories of venue that are nearby those locations. We will build a regression model to predict customer traffic as "low, medium, or high" based on features of the count of venues of different categories. Hopefully, we can identify characteristic neighbors that support a high-selling coffee shop.

In this study we are interested in analyzing neighborhoods in the West of Kuala Lumpur city, Malaysia. Let's define the coordinate of our search which is 3.1871 latitude and 101.6276 longitude. We can view this location on a map generated using Folium.

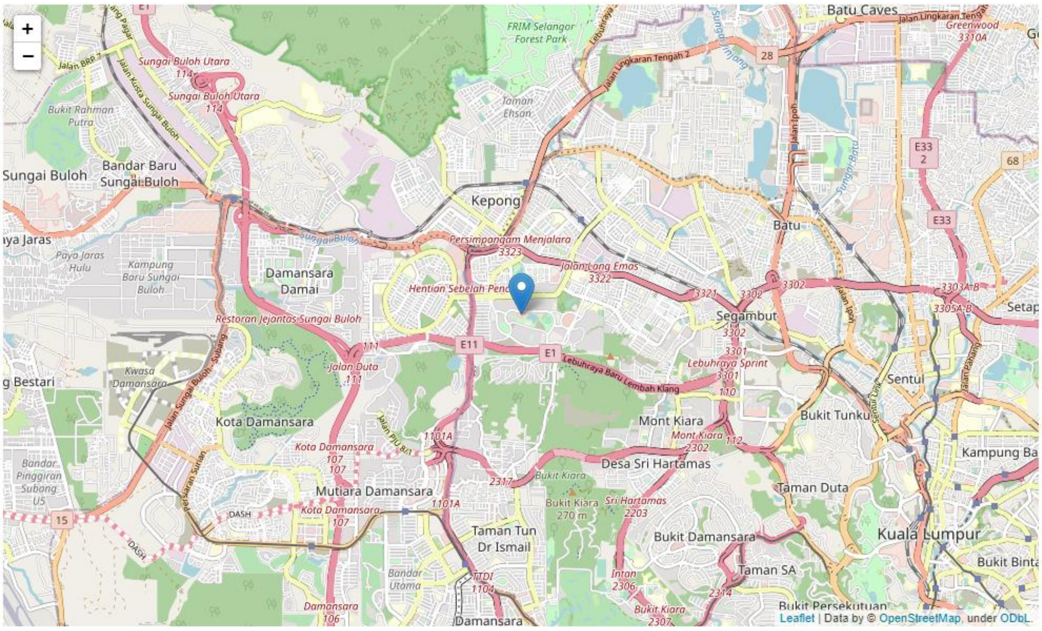
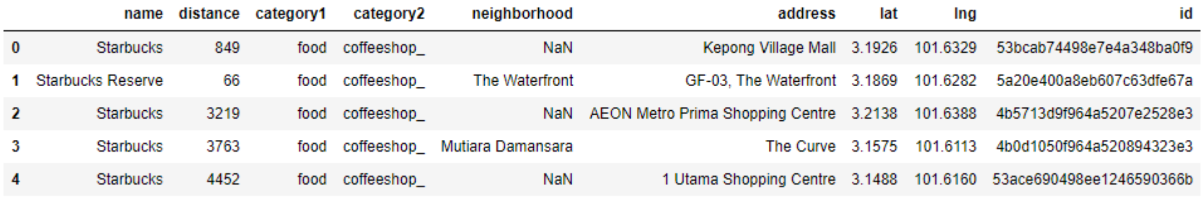


Figure 1 Search location in the West of Kuala Lumpur, Malaysia

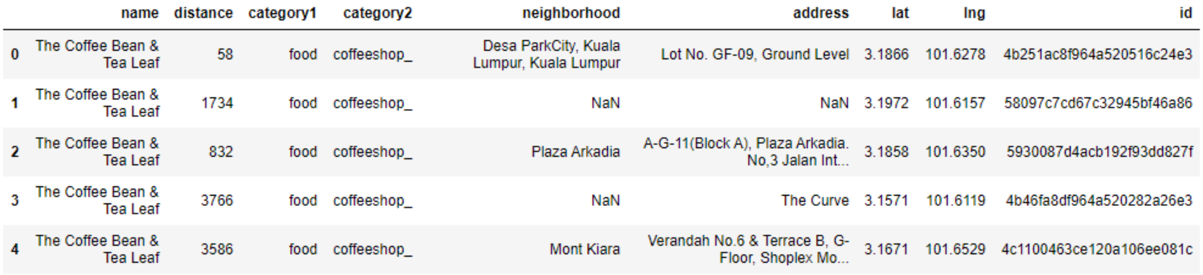
2.1 Getting coffee shop locations

After defining the center of search location we find the nearby Starbucks and Coffee Bean locations using the Foursquare's "search" API. A few rows of the search results are shown in below tables:



	name	distance	category1	category2	neighborhood	address	lat	lng	id
0	Starbucks	849	food	coffeeshop_	NaN	Kepong Village Mall	3.1926	101.6329	53bcab74498e7e4a348ba0f9
1	Starbucks Reserve	66	food	coffeeshop_	The Waterfront	GF-03, The Waterfront	3.1869	101.6282	5a20e400a8eb607c63dfe67a
2	Starbucks	3219	food	coffeeshop_	NaN	AEON Metro Prima Shopping Centre	3.2138	101.6388	4b5713d9f964a5207e2528e3
3	Starbucks	3763	food	coffeeshop_	Mutiara Damansara	The Curve	3.1575	101.6113	4b0d1050f964a520894323e3
4	Starbucks	4452	food	coffeeshop_	NaN	1 Utama Shopping Centre	3.1488	101.6160	53ace690498ee1246590366b

Figure 2 Sample data collected from Starbucks search

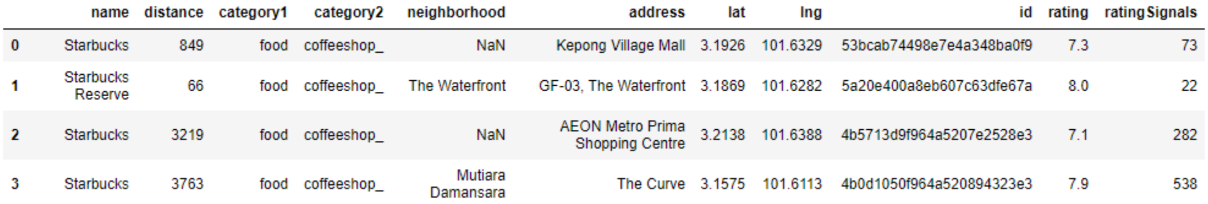


	name	distance	category1	category2	neighborhood	address	lat	lng	id
0	The Coffee Bean & Tea Leaf	58	food	coffeeshop_	Desa ParkCity, Kuala Lumpur, Kuala Lumpur	Lot No. GF-09, Ground Level	3.1866	101.6278	4b251ac8f964a520516c24e3
1	The Coffee Bean & Tea Leaf	1734	food	coffeeshop_	NaN	NaN	3.1972	101.6157	58097c7cd67c32945bf46a86
2	The Coffee Bean & Tea Leaf	832	food	coffeeshop_	Plaza Arkadia	A-G-11(Block A), Plaza Arkadia. No,3 Jalan Int...	3.1858	101.6350	5930087d4acb192f93dd827f
3	The Coffee Bean & Tea Leaf	3766	food	coffeeshop_	NaN	The Curve	3.1571	101.6119	4b46fa8df964a520282a26e3
4	The Coffee Bean & Tea Leaf	3586	food	coffeeshop_	Mont Kiara	Verandah No.6 & Terrace B, G-Floor, Shoplex Mo...	3.1671	101.6529	4c1100463ce120a106ee081c

Figure 3 Sample data collected from Coffee Bean & Tea Leaf search

2.2 Getting customer traffic data

Because we are trying to predict customer traffic, we also need to get the data of customer ratings. To do that, we use Foursquare's "venue" API of querying the details of a venue given its id. One of the returned keys that is important to us is the "ratingSignals" which can indicates how often people visited that venue. Of course, using this feature does not guarantee that it is 100% accurate. For example, people may visit but not rate the venue. So, this is going to be a rough approximation and it is the only data available from Foursquare.



	name	distance	category1	category2	neighborhood	address	lat	lng	id	rating	ratingSignals
0	Starbucks	849	food	coffeeshop_	NaN	Kepong Village Mall	3.1926	101.6329	53bcab74498e7e4a348ba0f9	7.3	73
1	Starbucks Reserve	66	food	coffeeshop_	The Waterfront	GF-03, The Waterfront	3.1869	101.6282	5a20e400a8eb607c63dfe67a	8.0	22
2	Starbucks	3219	food	coffeeshop_	NaN	AEON Metro Prima Shopping Centre	3.2138	101.6388	4b5713d9f964a5207e2528e3	7.1	282
3	Starbucks	3763	food	coffeeshop_	Mutiara Damansara	The Curve	3.1575	101.6113	4b0d1050f964a520894323e3	7.9	538

Figure 4 Sample of Rating Signals data to represent customer traffic

After getting the data, I did a manual check and found that Foursqure returned many false locations where the venues don't exist at the given coordinates. Some venues were duplicated locations but of different id's such that they seemed to be different venues. Simple but laborious Google Maps search can give us the answer if a venue is valid or not. There was no hard rule to automate the data cleaning process. After cleaning there are 22 coffee shop locations in the dataset. Let's visualize again on the map, but this time with all of the coffee shop locations shown.

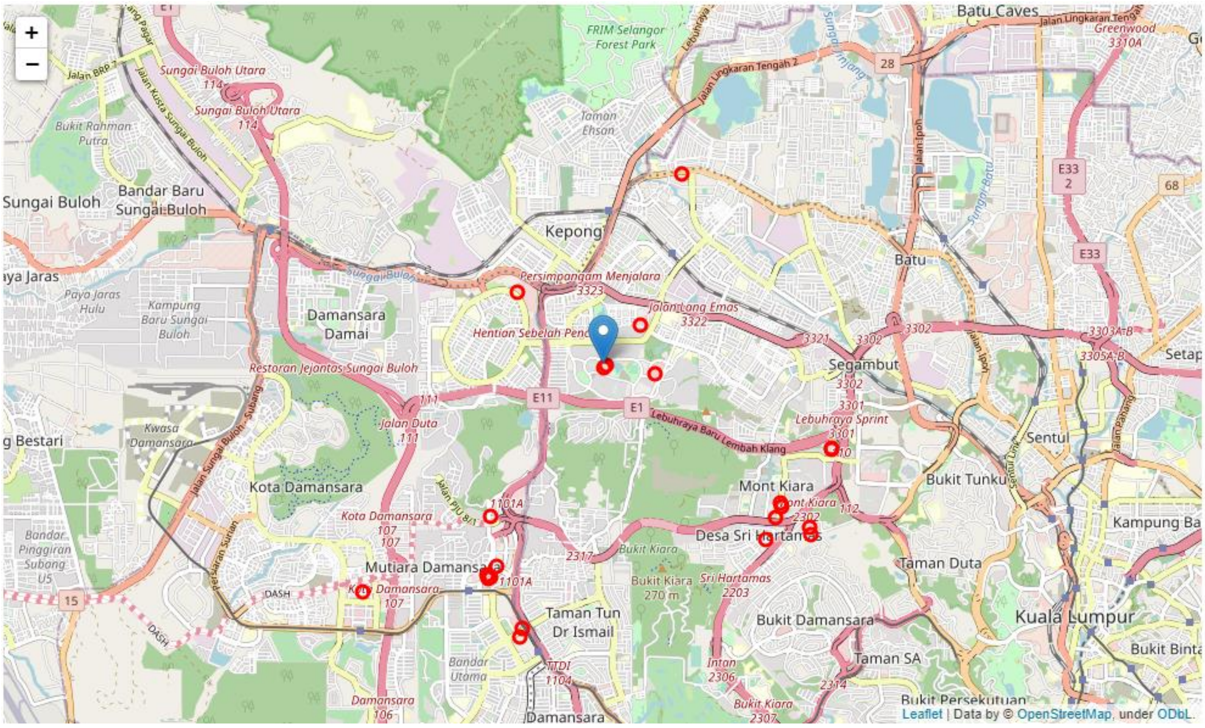


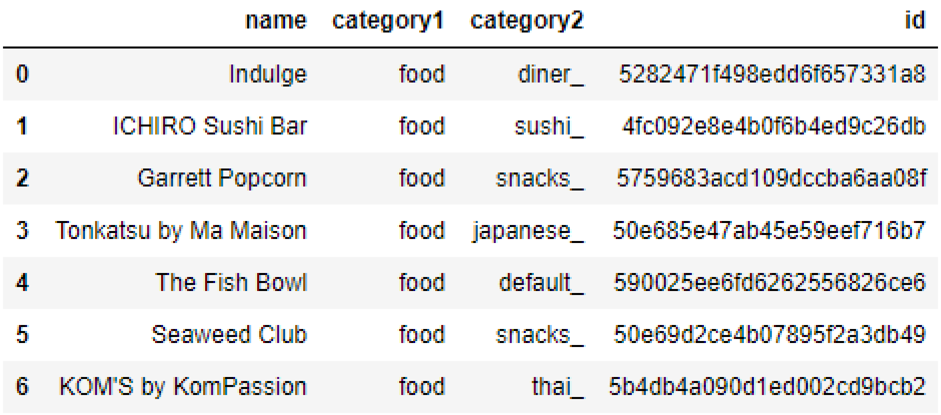
Figure 5 Locations of coffee shops to be modeled

3 Methodology

In the previous section, we collected data of coffee shops and their number of customer ratings (i.e. customer traffic). That dataset established our prediction target. The next step to build a model is to define our predictive features. In this study, we are going to build our features from the types of venues that are adjacent to the coffee shops. We are working on a hypothesis that certain types of venues and the number of them may promote traffic to our coffee business.

3.1 Building the Features

To get the nearby venues and their categories, we use the Foursquare's "explore" API. We write a for loop that calls the search for nearby venues for every coffee location in our dataset. For example, for one of the coffee shops, its neighbors are the following.



	name	category1	category2	id
0	Indulge	food	diner_	5282471f498edd6f657331a8
1	ICHIRO Sushi Bar	food	sushi_	4fc092e8e4b0f6b4ed9c26db
2	Garrett Popcorn	food	snacks_	5759683acd109dccba6aa08f
3	Tonkatsu by Ma Maison	food	japanese_	50e685e47ab45e59eef716b7
4	The Fish Bowl	food	default_	590025ee6fd6262556826ce6
5	Seaweed Club	food	snacks_	50e69d2ce4b07895f2a3db49
6	KOM'S by KomPassion	food	thai_	5b4db4a090d1ed002cd9bcb2

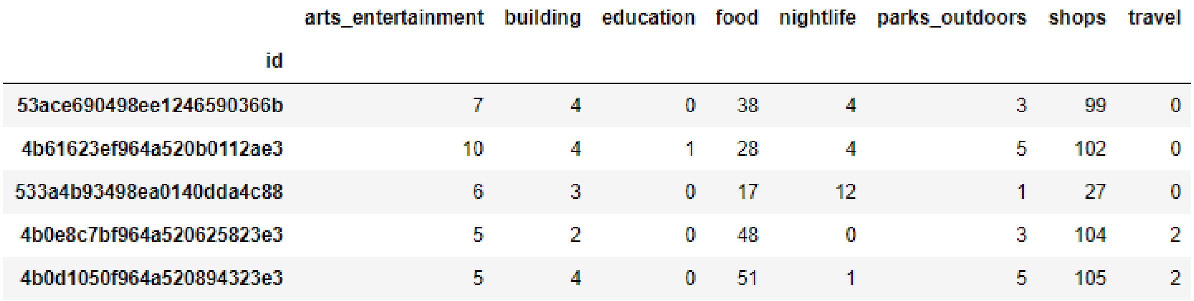
Figure 6 Sample of neighboring venues to a coffee shop

With the neighbors dataset, we identify all the unique categories that exist in the data set. Here we are only using the category1 because if we include category2, the feature dimension will get out of hand. We only have 22 coffee shops, so we want to keep the feature table from getting too wide.

The neighbors dataset contains 8 “category1” categories:

1. arts_entertainment
2. building
3. education
4. food
5. nightlife
6. parks_outdoors
7. shops
8. travel

Using the available categories above, we will develop a feature table that counts the number of venues for each category. So, each row will represent one coffee shop and each column is the count of venues for each category.

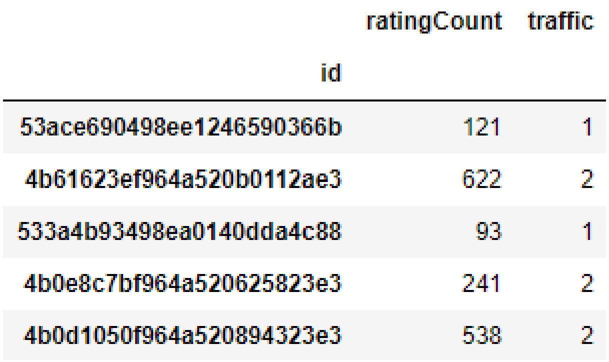


	arts_entertainment	building	education	food	nightlife	parks_outdoors	shops	travel
id								
53ace690498ee1246590366b	7	4	0	38	4	3	99	0
4b61623ef964a520b0112ae3	10	4	1	28	4	5	102	0
533a4b93498ea0140dda4c88	6	3	0	17	12	1	27	0
4b0e8c7bf964a520625823e3	5	2	0	48	0	3	104	2
4b0d1050f964a520894323e3	5	4	0	51	1	5	105	2

Figure 7 Features table for the first five coffee shops

3.2 Building the Target Variable

In the previous section, we created a features table that will be used to train a classification model. Now, let's define our target which is the rating count obtained from before. Because the rating signal count is somewhat a broad approximation and it has not a defined upper limit, it is not appropriate to use the raw values to define how busy a coffee shop is. To normalize the data, we create 3 level map of traffic of low, medium, and high, where a rating count < 50 is low traffic (0 score), rating count between 51-149 as medium traffic (score 1), and rating count >= 150 high traffic (score 2). Here is a sample of the mapping output for the target variable.



	ratingCount	traffic
id		
53ace690498ee1246590366b	121	1
4b61623ef964a520b0112ae3	622	2
533a4b93498ea0140dda4c88	93	1
4b0e8c7bf964a520625823e3	241	2
4b0d1050f964a520894323e3	538	2

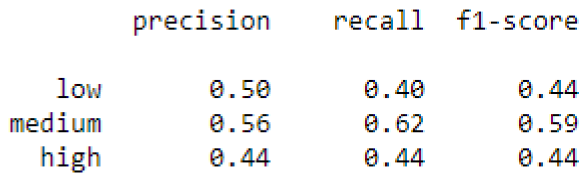
Figure 8 Target variable "traffic" calculated by mapping rating count to 0,1,2

3.3 Building a Classifier Model

For classification, a decision tree model was chosen because it offers the transparency to tell us how the features are split to determine the traffic prediction. The decision tree is set to have a maximum depth of 10. Because of our data only has 22 shops, we want to maximize our testing volume by performing a leave-one-out cross validation.

4 Results

The leave-one-out cross validation result has an overall accuracy of 54%. The accuracy is not very high, but this is a challenging problem with many data quality issues from Foursquare. Nevertheless, it would be very difficult for a human to make guesses and achieve this level of accuracy. Let's review the accuracy metrics in more detail by plotting the precision and recall.



```
          precision    recall  f1-score

     low       0.50      0.40      0.44
  medium       0.56      0.62      0.59
    high       0.44      0.44      0.44
```

Figure 9 Cross-validation precision and recall scores

The f1-scores across 3 classes do not show any high imbalance of favoring one class, so we can assume that the classifier performed somewhat equally across different classes.

In addition to the accuracy, what is more interesting is the features that defined the tree which can give us an insight into what matters in having a successful coffee shop. We can obtain that information by probing into the tree diagram below.

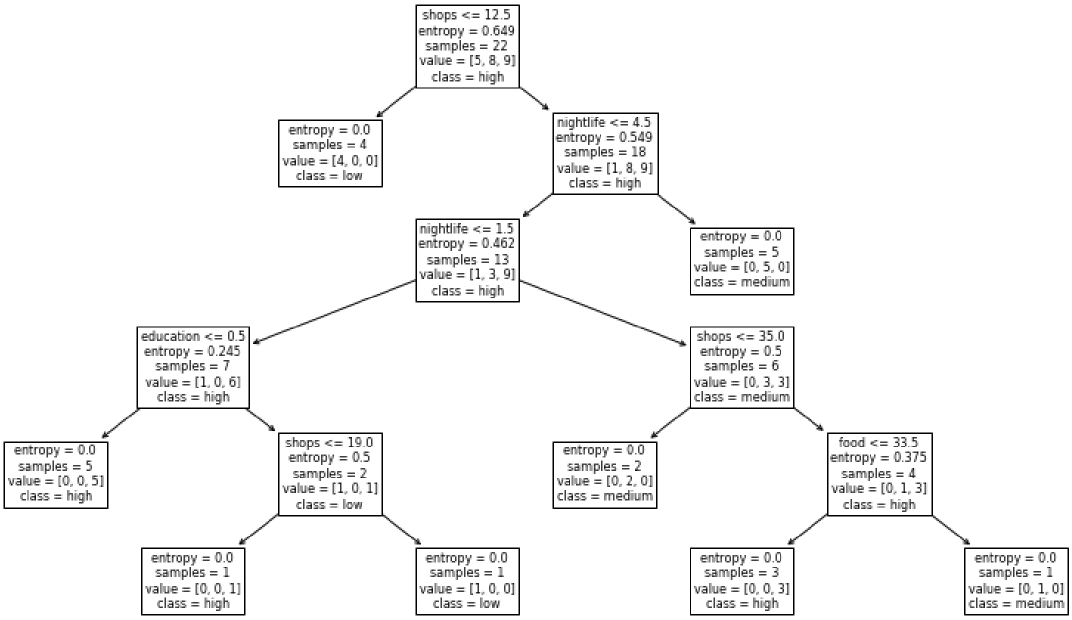


Figure 10 Final decision tree classifier for customer traffic prediction

5 Discussion

From the decision tree diagram, we found that the presence of shops took the top position in predicting if a coffee shop will get high business or not, and followed by nightlife. The bottom node with 5 samples of "high" (most left) is a pure node with highest number of high traffic. If we follow the path to get there, we come up with the following criteria:

1. If the number of nearby shops >= 13, and
2. number of nearby nightlife venues >= 5, and
3. number of nearby education venues >= 1, then the customer traffic will be high

It is quite easy to understand the above finding as we relate with real-life situation. If a coffee shop is surrounded by many shops, near nightlife attractions and a college, then it will attract a wide spectrum of customers from daytime shoppers, night goers and students.

So, how can we practically use this information to decide on a new shop location? Well, one way to do it is by using the Foursquare data to search for all locations of shops, nightlife, and education in each neighborhood. The data will be quite large to download but this should be computationally fast to process. Then, we can apply a k-means clustering to find clusters which satisfy our neighborhood requirements and place a coffee shop there.

We can further improve the accuracy by collecting more coffee shops data around KL, cleaning up the neighbors data from erroneous records, and using other sources such as Google Maps. We kept the scope of this study small and within the free services provided by Foursquare.

6 Conclusion

In this study, we have looked at some data from coffee shops in Kuala Lumpur to identify combination of factors that may help a business owner to decide where to place a new coffee shop so that it will receive high business traffic. The results were based on some limited data as provided by Foursquare in the form of location, customer ratings, and neighboring venues. We found several instances of inaccuracies in the Foursquare data. Despite the limitations with the data, the result of decision tree prediction gives a reasonable answer as to why certain coffee shops are more successful than others. Among the important factors are the presence of nearby shops, nightlife attractions, and education institutions.