Predicting Customer Traffic for Coffee Shops

Edwin Sutrisno

June 21, 2020

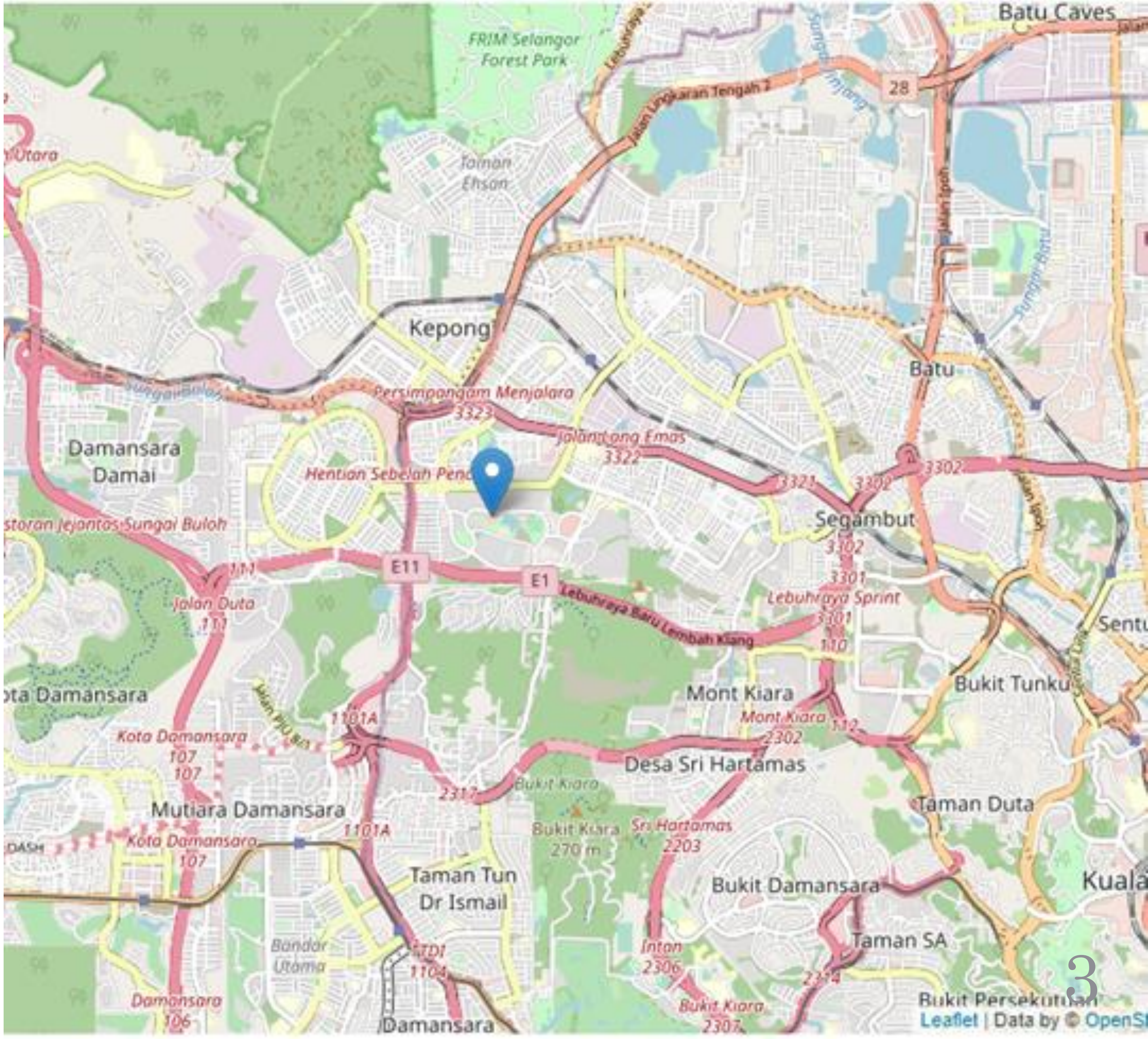
Business Background

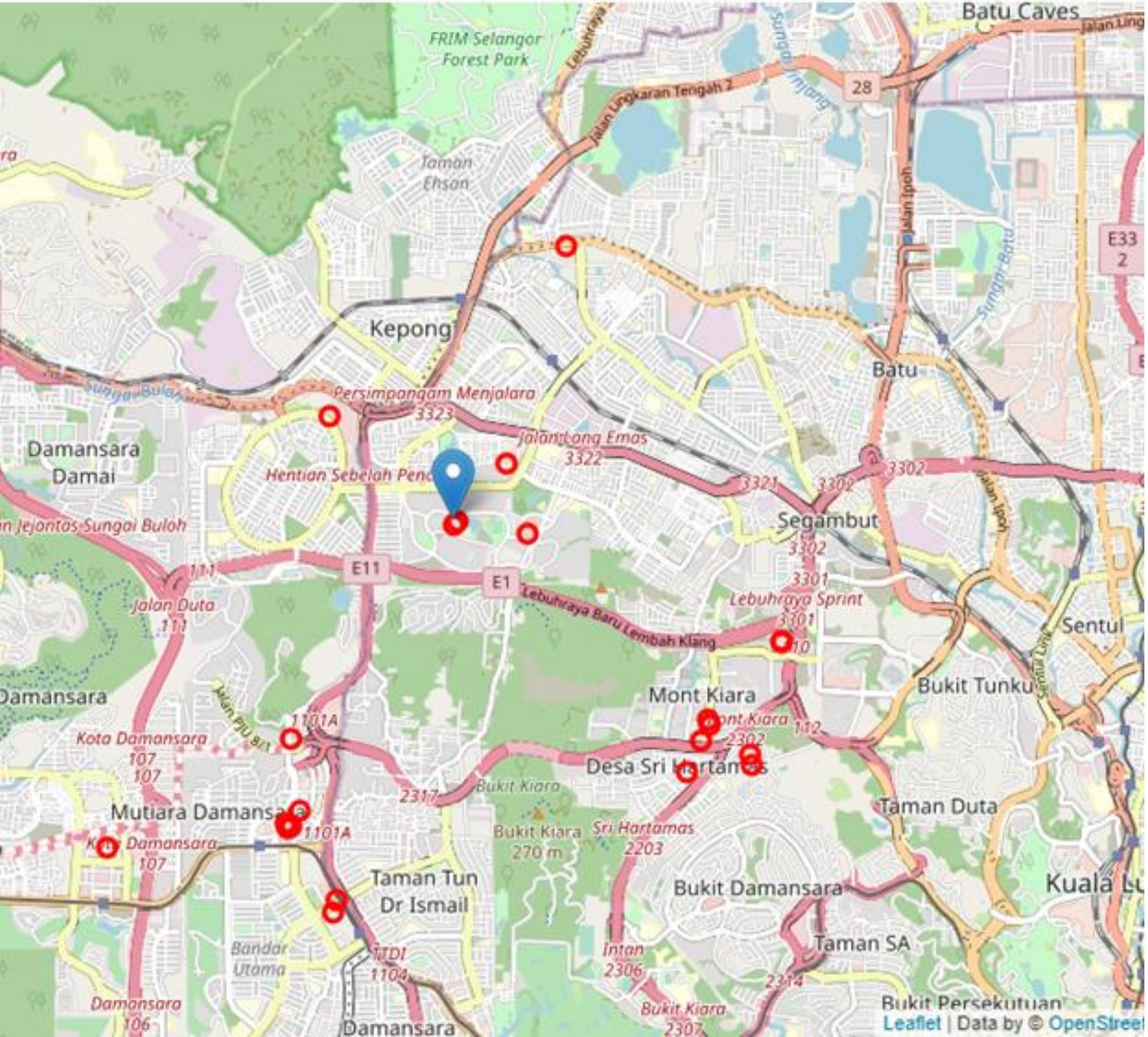
- Location is a known important factor for a business to attract customers
- Successful coffee shops offer some insights on the characteristics of their locations
- By building a predictive model, we can predict if a location will be suitable to establish a new coffee shop
- Predictive model will give the factors that are important in getting high customer traffic to the business

This Photo by Unknown Author is licensed under CC BY-SA

Data Source

- Foursquare provides free open-source data about venues such as coffee shops
- The location in study is an area in the West of Kuala Lumpur, Malaysia



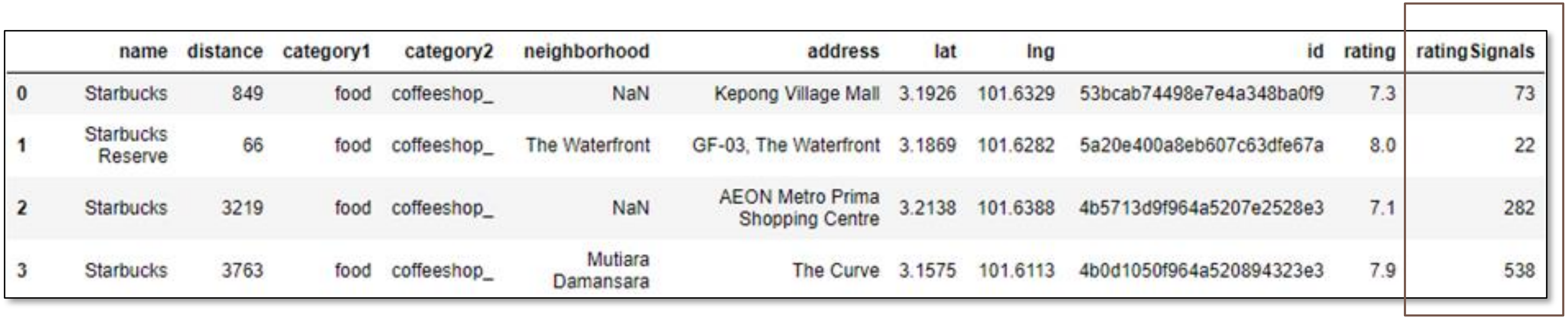


Coffee Shop Location Search

- The two most popular coffee shop chains in the city are Starbucks and the Coffee Bean & Tea Leaf
- We search for these two chains to collect data for training
- Use the Foursquare's "search" API to identify their ID's and geo-locations (red circles)
- Search Radius = 4 km

Customer Traffic Data

- The search API only provides the locations and unique ID for each venue
- Customer traffic data is obtained by querying for the venue's details using the "venue" API on Foursquare
  - https://api.foursquare.com/v2/venues/{venueID}
- There is not an exact data for customer traffic but the returned venue details has a "ratingSignals" field which can be used as a substitute



	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

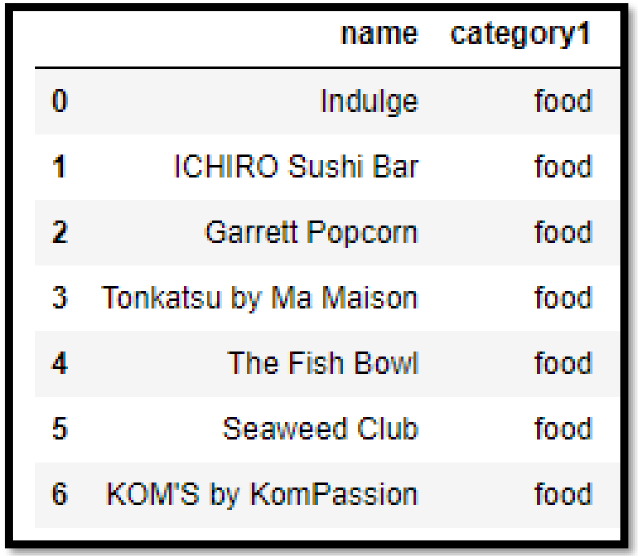
Data Cleaning

- Venues returned by Foursquare are not all accurate
- Manual search on Google Maps found that some shops did not exist
- Some shops were duplicates
- Data had to be manually cleaned because there was not a hard-rule to identify the false data
- The cleaned data set has 22 coffee shops

Features for Traffic Modeling

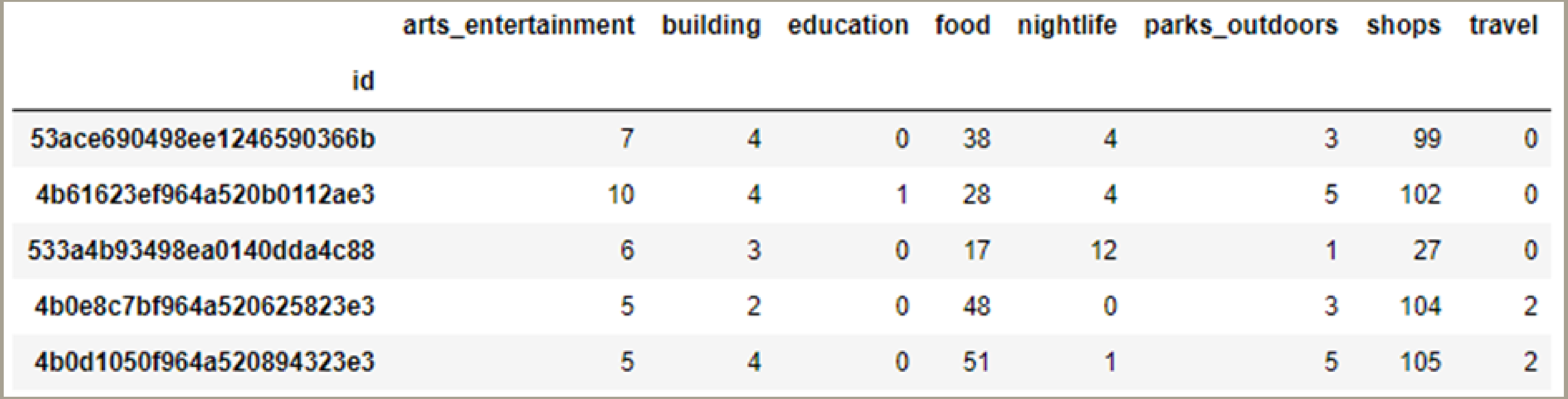
- Features for modeling are the counts of different businesses that are nearby a coffee shop location
- Use the "explore" API to search for nearby venues within 120 m radius and loop the search for all 22 coffee shops in the data set

Neighbors Example



```
                    name  category1
0                Indulge       food
1       ICHIRO Sushi Bar       food
2        Garrett Popcorn       food
3  Tonkatsu by Ma Maison       food
4          The Fish Bowl       food
5           Seaweed Club       food
6    KOM'S by KomPassion       food
```

All Neighbor Categories	
1. Arts & Entertainment	5. Nightlife
2. Building	6. Parks & Outdoors
3. Education	7. Shops
4. Food	8. Travel



	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

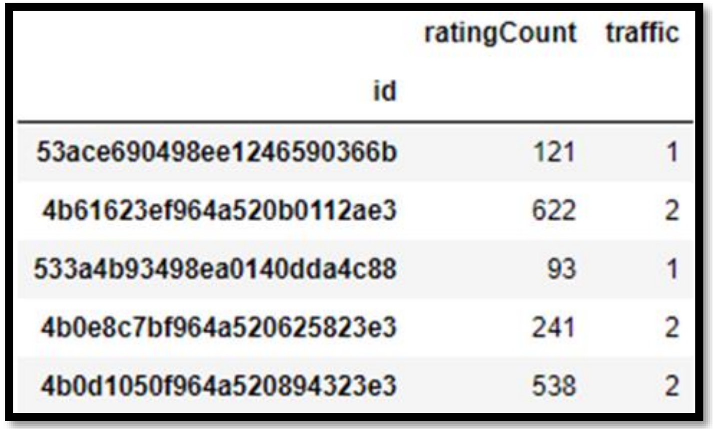
Features Table

A table of count of neighbors in each category where each row is a coffee shop

Target Variable Preprocessing

- The rating count data represents customer traffic however there will be some issues to use the raw values directly in the model:
  - rating count is highly inaccurate because not all customers reported a rating, so its raw values may be perceived as detailed variations by a model
  - rating count does not have an upper limit and it continues to increase over time, so after some level a higher number is no longer relevant

- To address the issues above, the rating count is transformed into 3 levels of customer traffic:
  - rating count < 50 is Low Traffic (score 0)
  - rating count 51-149 is Medium Traffic (score 1)
  - rating count 150+ is High Traffic (score 2)



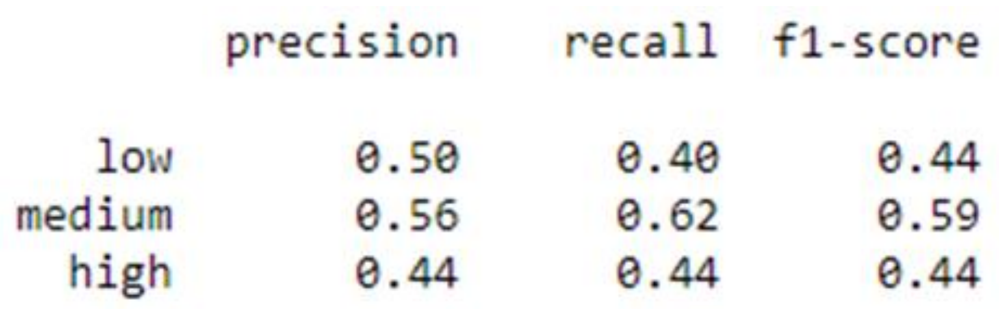
	ratingCount	traffic
id		
53ace690498ee1246590366b	121	1
4b61623ef964a520b0112ae3	622	2
533a4b93498ea0140dda4c88	93	1
4b0e8c7bf964a520625823e3	241	2
4b0d1050f964a520894323e3	538	2

Decision Tree Classification Model

- A decision tree is chosen for the classification model because it offers the ability to show the split of the features which can tell the important features to predict

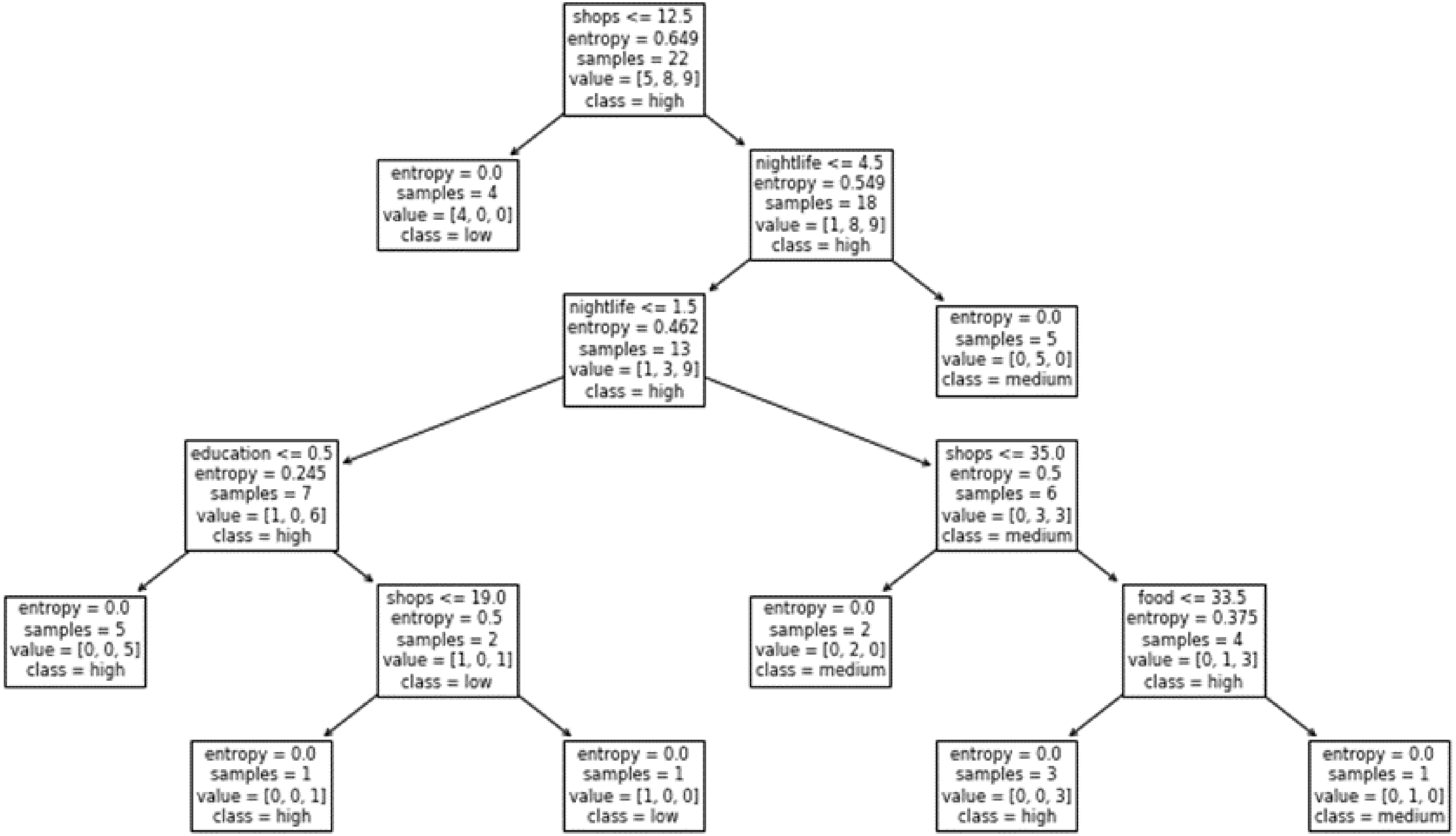
- To maximize the available data (22 samples) a leave-one-out cross validation approach was used to generate the test accuracy

- The overall accuracy = 54%



```
        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
```



shops <= 12.5
entropy = 0.649
samples = 22
value = [5, 8, 9]
class = high

entropy = 0.0
samples = 4
value = [4, 0, 0]
class = low

nightlife <= 4.5
entropy = 0.549
samples = 18
value = [1, 8, 9]
class = high

nightlife <= 1.5
entropy = 0.462
samples = 13
value = [1, 3, 9]
class = high

entropy = 0.0
samples = 5
value = [0, 5, 0]
class = medium

education <= 0.5
entropy = 0.245
samples = 7
value = [1, 0, 6]
class = high

shops <= 35.0
entropy = 0.5
samples = 6
value = [0, 3, 3]
class = medium

entropy = 0.0
samples = 5
value = [0, 0, 5]
class = high

shops <= 19.0
entropy = 0.5
samples = 2
value = [1, 0, 1]
class = low

entropy = 0.0
samples = 2
value = [0, 2, 0]
class = medium

food <= 33.5
entropy = 0.375
samples = 4
value = [0, 1, 3]
class = high

entropy = 0.0
samples = 1
value = [0, 0, 1]
class = high

entropy = 0.0
samples = 1
value = [1, 0, 0]
class = low

entropy = 0.0
samples = 3
value = [0, 0, 3]
class = high

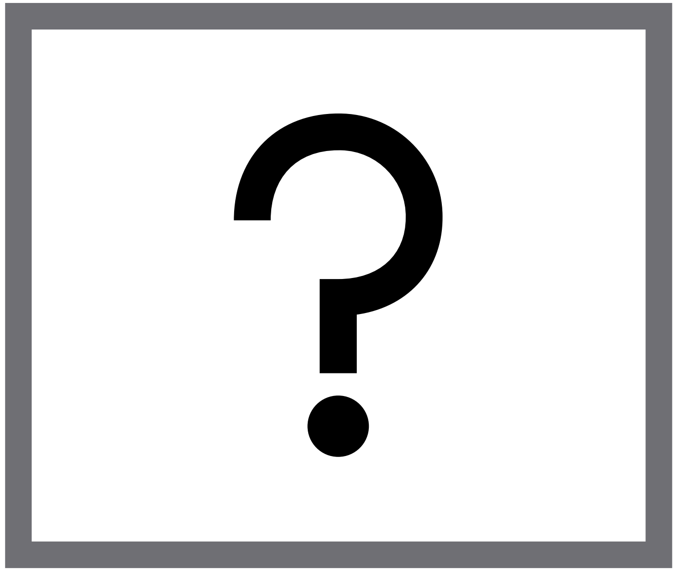
entropy = 0.0
samples = 1
value = [0, 1, 0]
class = medium

Result Discussion

- Decision Tree showed that the number of shops took the top position in predicting customer traffic, followed by nightlife.
- A leaf node with 5 samples of "high" (left most) is a pure node with the highest samples of high traffic compared to other leaf nodes. The path to get to this node is:
  - If the number of nearby shops >= 13, and
  - number of nearby nightlife venues >= 5, and
  - number of nearby education venues >= 1, then the customer traffic will be high
- It is quite easy to understand the above path in real-life situation. If a coffee shop is surrounded by many shops, near nightlife attractions and a college, then it can attract a wide spectrum of customers from daytime shoppers, night goers and students.

Future Application

- Use the model to predict the traffic outcome for a given new location
- OR
- Use Foursquare to search all locations in the city for shops, nightlife, and education and apply a k-means clustering to find clusters which satisfy the criteria and place a new coffee shop there



Conclusion

- A decision tree classifier model was built to predict customer traffic using open-source data collected from Starbucks and The Coffee Bean & Tea Leaf locations around west of Kuala Lumpur area
- The model gave a reasonable quality prediction despite the limited amount of data and data quality issues in the Foursquare open-source database
- Among the key factors that drove customer traffic are the number of neighboring shops, nightlife attractions, and educational places
- Further improvement to the model can be done by collecting more data from additional sources such as Google Maps and conducting more rigorous data cleaning