Locating the Best Neighborhood for a Company's Headquarters

Mark Sproull

April 14, 2019

1. Introduction

1.1 Scenario

A fast-growing web content management company began with no central headquarters and all employees working remotely but had grown to a size for which a physical headquarters location made sense.

1.2 Business Problem

Although the company's management believed a brick-and-mortar headquarters would beneficial, they were keenly aware that most companies in their industry were located in the major metropolitan areas of New York and San Francisco; cities with high real estate prices and high cost of living. The company did not have the resources to lease space in one of these cities and needed a way to find a location outside of the major cities that would, nevertheless, enable them to retain their current workforce and attract excellent new talent.

1.3 Strategy:

Management decided to focus on emulating the highly-rated working conditions at top-tier competitors in their industry, but to do so in a city and neighborhood that would be affordable for their young, growing company. They began with the following assumptions:

- Renovating a building to modern standards (if necessary) in a less expensive city would be more cost-effective than leasing space in a high cost-of-living city.
- Aside from the physical building and company policies, the quality of a work location is, to a large extent, determined by the quality of the neighborhood in which it is located.
- For a work location, the quality of a neighborhood is determined by the assortment of service-oriented businesses, recreational, and cultural opportunities it offers.
- Their top-tier competitors are located in neighborhoods that now represent ideal mixes of these venues that attract top-notch talent.

1.4 Groundwork

They engaged a data science consultant to advise them on the cities and neighborhoods that would offer the best environment for the headquarters. The consultant's first recommendation was that they not use cost as the only quality-of-life criterion for picking candidate cities, but also include criteria such as population and commuting time.

1.5 Process:

Foursquare venue data from the neighborhoods where their top-tier competitors operate was used to train a machine learning model, allowing the model to recognize neighborhoods that attract top talent in the field. The model was then used to identify candidate neighborhoods in other US cities that offer a similar mix of venues.

2. Data

- **Names of Top-Tier competitors:**

  US Companies appearing in the "EContent 100" for 2018.

- **Top-Tier competitor neighborhood data:**

  Foursquare data for venues within 500 meters in all directions from the street address of the company headquarters.

- **Negative examples for training the model**

  Foursquare data from locations in the competitors' cities, but from neighborhoods where there are no competitor companies.

- **Candidate Cities List :**

  Cities were chosen from the 2018 U.S. News & World Report Best Places Rankings, filtered as follows:

  - Population size < 1,000,000
  - Average Rent < $1000.00
  - Median Home Price < $390,000
  - Commute Time < 20 minutes

- **Candidate city neighborhood data**:

  Full city records, by neighborhood, of Foursquare data from the candidate city list above.

3. Methodology

The question of whether, or to what degree, one neighborhood matches an established ideal is a classification problem. The single constraint presented for the problem was that the solution must use Foursquare venue data for neighborhoods, leading to the assumption that a dataset consisting of venue data can be used as sort of "fingerprint" to identify a neighborhood. Further, it was assumed that training a machine learning model to recognize an ideal neighborhood would allow it to recognize similar neighborhoods in other cities. The specific machine learning model that was ultimately chosen to address this question will be discussed in a later section. The initial steps in preparing data for training and using the model are identical regardless of model used.

1. Identify top-tier competitors

The report on top-tier companies identifies the following 81 companies, which have US Headquarters:

Acquia	Cision	Haivision	Smartling
Acquire Media	CognitiveScale	Linkedin	Snap
Acroilinx	Crafter Software	Lionbridge	Spotify
Act-On Software	Crownpeak	Lucidworks	Sysmos
Adobe	Csoft	MarketMuse	Taboola
Akamai	Curata	Marklogic	Talkwalker
Amazon.com	DNN	Moat	Techsmith
Apple	Ebsco	Netflix	Transperfect
Appnexus	Ephox	Newscred	Twitter
Aptara	Episerver	Nielsen	Viglink
Aria Systems	e-Spirit	Onespot	Vimeo
Attivio	Evergage	Oracle	Webtrends
Atypon Systems	Facebook	Perfect Sense	Welocalize
Automattic	Frame.io	Proquest	Widen
Bloomreach	Google	Quark	Wistia
Brightcove	Hortonworks	Reprints Desk	Wochit
Buffer.	Hubspot	Revizzit	Youappi
Ceros	IBM	Salesforce	Youvisit
Clarabridge	Impelsys	SAS	Zumobi
Cloudera	Ingeniux	Sizmek	
Cloudwords	Lingotek	Skimlinks	

Companies with headquarters outside the United States were excluded from the analysis, as the candidate cities were exclusively within the US.

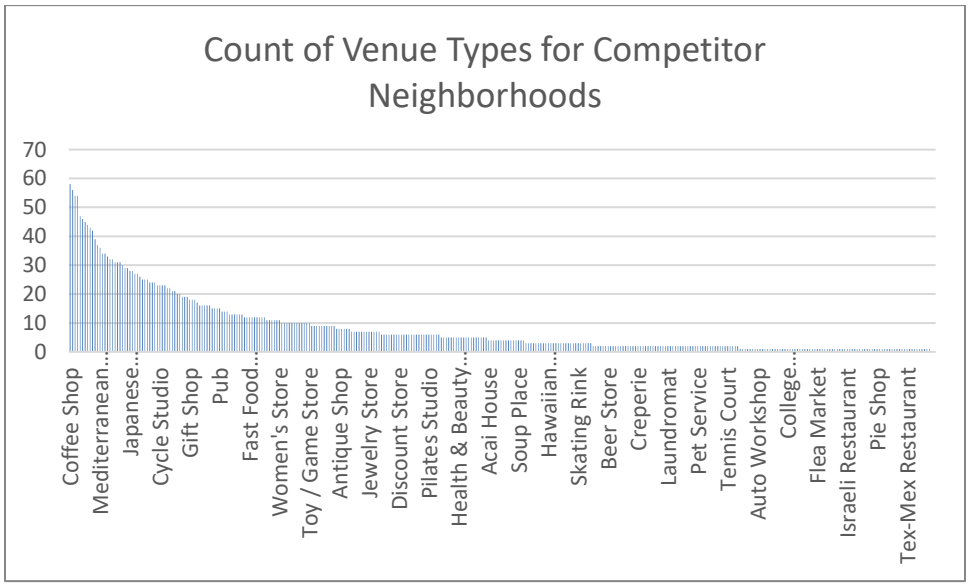
2. Gather Foursquare data for the top-tier competitors' neighborhoods

2.1 Competitor address and GPS coordinates

For each competitor, a Google search identified the street address of the headquarters facility. Nominatum was used to retrieve GPS coordinates for each competitor. For a small number of companies (Adobe, Aptera, Facebook, Hortonworks, Lingotek, Revizzit, Taboola, Youappi) Nominatum did not return results, and the GPS coordinates were retrieved directly from Google Maps.

2.2 Foursquare venue data for the competitors

For the 81 competitors, neighborhood venue data was retrieved from Foursquare, converted to one-hot representation and grouped by competitor name. The resulting dataset listed 347 distinct venues across the neighborhoods of the competitor companies. However, it was immediately apparent that the overall count of each venue type fell off rapidly across the neighborhoods:



It was observed that:

- only 11 venues appeared in more than half the neighborhoods
- The most frequent 50 venues appeared in only 18 (22%) of the neighborhoods
- The most frequent 100 venues appeared in only 9 (11%) of the neighborhoods.

It was decided to construct the training set with the top 25 most frequently occurring venues, which would include venues that appear in 35% of the competitors' neighborhoods while being a low enough number of features to minimize overhead in training the model.

An indicator column filled with 1s was added to the dataset to identify the records as positive examples.

3. Gather negative examples to be used in training the model

An equal number of negative examples were gathered using Google Maps by selecting locations in the competitors' cities that did not have a competitor headquarters and obtaining the GPS coordinates of those locations. These coordinates were then passed to the Foursquare API to gather venue data for the negative neighborhood examples.

The resulting dataset contained 343 venues across the selected neighborhoods. This dataset was converted to one-hot representation and grouped by location name, as was done with the competitor data. It was then filtered to the same 25 venues as the competitors' dataset, and an indicator column of zeros was added to identify these as negative examples.

4. Gather Foursquare neighborhood data for the top 10 candidate cities

4.1 Identify the cities, their neighborhood names and GPS coordinates

Using the filter listed above, the Best Places to Live Ranking gives the following results:

1. Wichita KS
2. Des Moines IA
3. Madison WI
4. Boise ID
5. Omaha NE
6. Greenville SC
7. Lancaster PA
8. Albany NY
9. Fort Myers FL
10. Winston Salem NC

A Google search of each city name was used to obtain a list of neighborhoods in the city, then Google maps was used to identify GPS coordinates of each neighborhood.

4.2 Gather and standardize Foursquare data for each neighborhood in the candidate cities

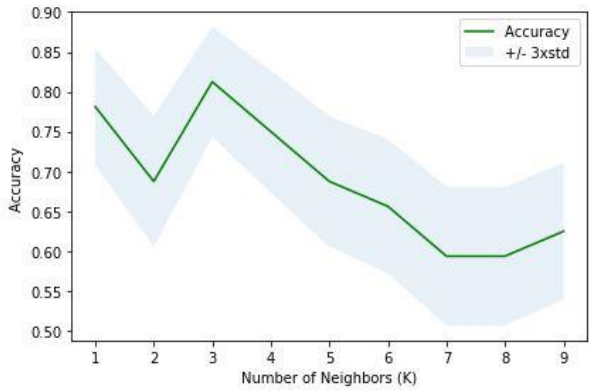
Candidate city neighborhood venue data was retrieved from Foursquare, converted to one-hot representation and grouped by neighborhood name. The resulting datasets listed varying numbers of venues across the neighborhoods of the candidate cities, with the low being 64 venues. No city result included all of the competitors' top 25 venues. Therefore, each city dataset was standardized to the same set of venues as the competitors dataset, by adding columns of zeros where venues were missing.

City	Number of columns added
Wichita	3
Des Moines	5
Madison	3
Boise	9
Omaha	1
Greenville	6
Lancaster	9
Albany	6
Fort Myers	8
Winston-Salem	10

5. Train a machine learning model

The competitor and not-competitor datasets were combined and shuffled, split into training and test sets (test 20%), standardized, and processed using the following algorithms:

K Nearest Neighbor:



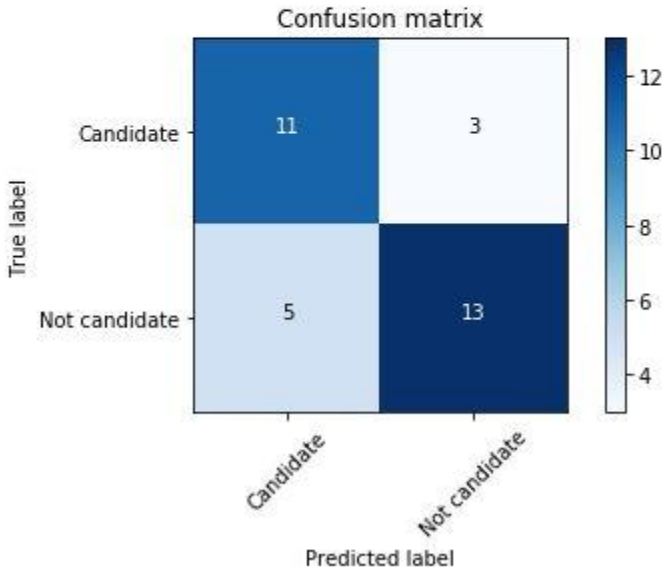

```
The best accuracy was  0.8125 with k= 3
```

Decision Tree:
Max depth of 5 gave the highest accuracy for this method, 0.75

Support Vector Machine:

	precision	recall	f1-score	support
0	0.69	0.79	0.73	14
1	0.81	0.72	0.76	18
avg / total	0.76	0.75	0.75	32

```
Confusion matrix, without normalization
[[11  3]
 [ 5 13]]
```



Random Forest:
In training, this method produced an accuracy of .81 and was, therefore, pursued.

Each city's data was run through the model. Results were combined. However the model produced positive results for 103 of the possible 437 neighborhoods, casting doubt on the value of a binary categorizer for this purpose.

Logistic Regression:

The probability function of logistic regression was identified as a way to rank the candidate neighborhoods highest to lowest, eliminating the problem of all results above 50% probability being reported as positive. Using the probability function, it was possible to sort the resulting output by the probability of a positive (i.e. "1") result, giving the following top 10 rankings:

City	Neighborhood	Probabilty
Madison	High Crossing	0.890728
Omaha	Aksarben Village	0.839298
Wichita	Sleepy Hollow	0.814578
Des Moines	East Village	0.810072
Wichita	A Price Woodard	0.807952
Des Moines	Downtown	0.749342
Greenville	Downtown	0.737687
Madison	Eken Park	0.728206
Madison	Capitol	0.724861
Omaha	Midtown	0.715809

The most recommended city is Madison, with 3 recommended neighborhoods. Wichita, Des Moines and Omaha have two each, and Greenville appears once in the top ten. It should be noted that 11th place was a tie between two Omaha neighborhoods, each with a probability of 0.714233. After 11th place, the probability fell below 70%.

4. Results

Based on the Foursquare training data provided to the classification models in this exercise, neighborhoods in four smaller US cities were rated as having the highest probability of matching the "fingerprint" of venues in neighborhoods with already-established top-tier companies in the Web Content industry. Wichita Kansas had both the highest-rated neighborhood and the largest number of neighborhoods in the results. Given that there are only 5 cities in the top ten, my recommendation would be to present all 5 cities and their neighborhoods to the client for evaluation.

5. Conclusion

The classification problem in this exercise proved to be a weeks-long effort due to the large amount of geographic data that needed to be gathered, and the restriction that it not be obtained commercially.

Throughout the exercise, my aim was to never second-guess the distinction between categories presented in the Foursquare data. (For example, should "Italian Restaurant" be folded into "Restaurant"?) It was tempting to consider collapsing some of the categories together as a means to provide more variety in the top 25 venues list. However, consideration of this question led me to resist that temptation, as none of the prior uses that had been demonstrated for this data had attempted to subjectively combine categories.

Spending the additional time working with the Random Forest classifier proved to be valuable, not only because we had not used Random Forest in the course, but also because it gave additional insight into the results of the Logistic Regression probability model. The two models agreed except where the probabilities approached 50%.

Given additional time and effort, a larger number of classification methods could have been explored, the number of venues in the training / test sets could have been increased, or a deep learning neural network could have been evaluated in order to provide better rankings. More powerful methods or larger numbers of features, however, might have required more compute power than the free tier of Watson Studio provides.