Identifying the best expansion opportunities for NY based Patsy's pizzeria

Farukh Zaripov

March 11, 2019

Contents

Introduction

My Capstone Project is focused on developing ML models to recommend the best possible LA locations as expansion opportunities for Patsy's pizzeria, one of the most iconic pizzerias of New York city.

Since its origination in 1933, Patsy’s was able to build a loyal client base and became a regular hang out place for many celebrities including Frank Sinatra, Dean Martin and Tony Bennett. Despite its popularity, Patsy’s still operates at a single location at 2287 First Avenue New York, NY 10035.

The ML model will allow comparing various neighborhoods (on zip code level) based on common sets of features and recommend the best match to current successful location.

Business problem and targeted audience.

The model would help resolving a common problem of identifying the optimal operating location to expand a successful business model.

While large corporations employ a team of data engineers and scientists to address the problem, there is a large underserved market of small businesses that are deprived of those analytical capabilities.

Although, the focus of the project is Patsy's Pizzeria, the model could be leveraged by other businesses to address common expansion challenges.

Data sources.

Open source data and feature engineering.

I used the following sources collect additional insights about every geographic area on zip code level. The collected data was transformed, merged and aggregated to produce additional data feeds.

Lat/long for all US zip codes (https://gist.github.com/erichurst/7882666)

- The data was used primarily to generate the map with the recommended new locations.

Database of active US businesses based on zip codes.

(https://www2.census.gov/programs-surveys/cbp/datasets/2016/zbp16detail.zip?#)

The file provides a breakdown of US businesses based on number of employees

- Total Number of Establishments
- Number of Establishments: 1-4 Employee Size Class
- Number of Establishments: 5-9 Employee Size Class
- Number of Establishments: 10-19 Employee Size Class
- Number of Establishments: 20-49 Employee Size Class
- Number of Establishments: 50-99 Employee Size Class
- Number of Establishments: 100-249 Employee Size Class
- Number of Establishments: 250-499 Employee Size Class
- Number of Establishments: 500-999 Employee Size Class
- Number of Establishments: 1,000 or More Employee Size Class

IRS tax return for every US zip codes (https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi). The file provides number of returns, which approximates the number of households and groups total tax returns into the following bins based on the income:

- under 25,000 USD >
- between 25,000 and 50,000 USD
- between 50,000 and 75,000 USD
- between 75,000 and 100,000 USD
- between 100,000 and 200,000 USD
- over 200,000 USD

Methodology used to achieve the project goals.

Identifying analytic approach.

Out of four available patterns (descriptive, diagnostic, predictive and prescriptive) prescriptive approach seem to be most fitting to achieve the project goals.

In particular, I focused on unsupervised learning algorithms to identify some similarities between different geo locations and place them into appropriate clusters.

Agglomerative Hierarchical Clustering algorithm seemed to be the best suited for the job. The selection choice was driven by the following advantages the algorithm provides for this particular case.

- No need to worry about identifying clusters and let the model to do the job
- It allows visualizing the distances between different zip codes and helps to immediately identify the closest zip codes to the original location based on their feature similarities.

In our case, I used the current location Patsy's NY zip code 10035 and tried algorithmically identifying similar locations in LA, California area.

Data understanding and preparation

Since the source data files contained both tabular and non-tabular data formats , I had to combine all of the sources into a single file outside of the Jupiter using Excel pivot table techniques and data manipulation using SQL based engines. Once the file was prepared, I loaded it into Watson Studio project.

The following is the final combined version of a file that was used to build the model.

File layout (19 data elements)

- ZIPCODE
- LAT
- LONG
- IND_INCOME_25
- IND_INCOME_25_50
- IND_INCOME_50_75

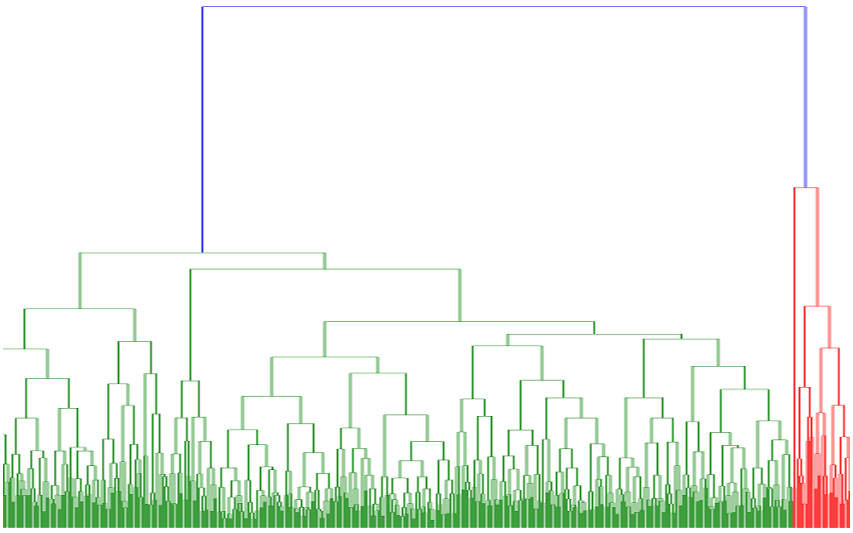
- IND_INCOME_75_100
- IND_INCOME_100_200
- IND_INCOME_OVER_200
- BUS_TOTAL
- BUS_EMPL_1_4
- BUS_EMPL_5_9
- BUS_EMPL_10_19
- BUS_EMPL_20_49
- BUS_EMPL_50_99
- BUS_EMPL_100_249
- BUS_EMPL_250_499
- BUS_EMPL_500_999
- BUS_EMPL_OVER_1000

File size: 1483 records (zip codes). The filed consisted of all available LA zip codes and a single zip code of the current NY location

ML model.

As mentioned previously, I leveraged Hierarchical Clustering algorithm to identify the new potential locations for Patsy's Pizzeria. The same ML model was used three times against decreasing subset of data. After each run, I would focus on the cluster that contain the actual NY location and run it again for the subset of the data.

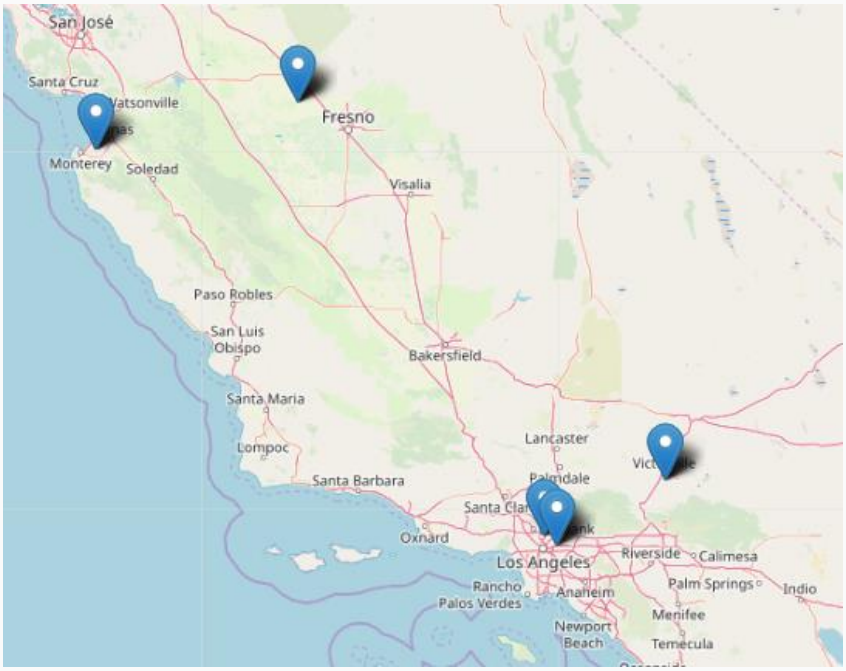
Original diagram of the hierarchal clustering model

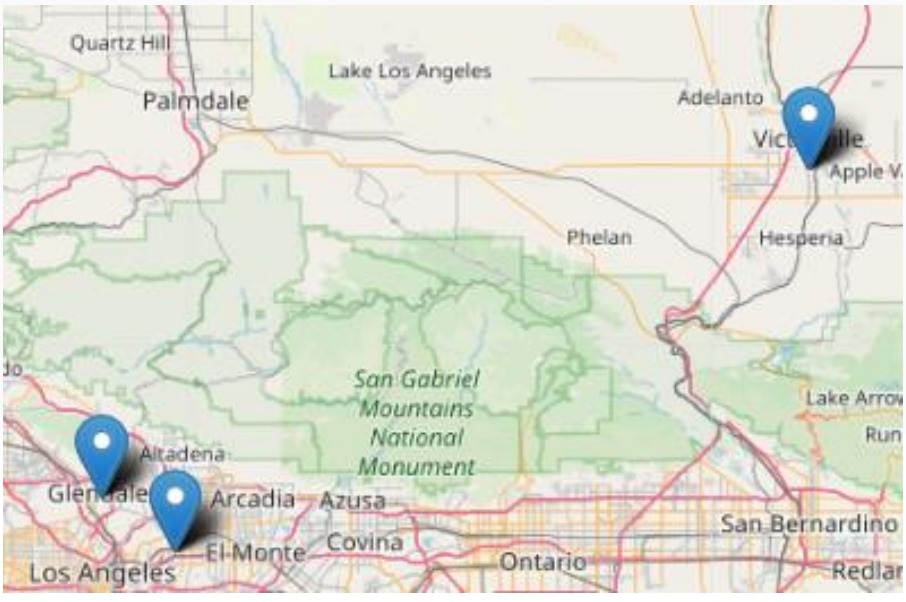


I was able to gradually reduce the larger set of data to more manageable subsets and ended up with only 5 new locations with the shortest Euclidian distance to the original location.

Results and recommendations.

My extensive analysis revealed the following five California locations with the corresponding features to the that current Patsy's location at 2287 First Avenue New York, NY 10035.





As shown in the table below the features of the five targeted locations correlate **very closely** to the current location of the Patcy's.

	ZIPCODE	10035	91205	91803	92395	93637	93955
1	USD_25	7040	7350	5910	7660	6120	5470
2	USD_25_50	3890	4240	3770	4110	4520	4470
3	USD_50_75	1950	1980	1900	2040	2110	2210
4	USD_75_100	910	1090	1180	1100	1090	1150
5	USD_100_200	1020	1220	1560	1160	1210	1270
6	USD_OVER_200	320	240	310	220	270	200
7	BUS_TOTAL	3918	4488	4682	3608	3790	3546
8	EMPL_1_4	2256	2832	2954	1790	1738	1828
9	EMPL_5_9	648	696	708	780	876	770
10	EMPL_10_19	414	468	504	516	540	426
11	EMPL_20_49	348	348	342	354	408	342
12	EMPL_50_99	150	102	108	84	138	132
13	EMPL_100_249	78	30	48	54	60	36
14	BUS_EMPL_250_499	12	12	18	18	18	6
15	BUS_EMPL_500_999	6	0	0	12	6	6
16	BUS_EMPL_OVER_1000	6	0	0	0	6	0
	Correlation Coefficient		0.997493	0.977515	0.994565	0.987003	0.977373

- Columns represent zip codes:
    - 10035 for actual NY location
    - five additional zip codes that were recommended by ML model.
- Rows 1-6 represent household baskets by different income sizes
- Rows 7-16 represent total companies broken down by their size
- The last row calculates Correlation Coefficient between five targeted and the existing locations. **The correlation is very high.**

Venue and location data using FourSquare

In conclusion, I leveraged FourSquare API to provide additional insights on potential new California locations with focus on the top most popular venues for each zip code.

targeted zip codes	1st Most	2nd Most	3rd Most	4th Most	5th Most	6th Most	7th Most	8th Most	9th Most	10th Most
91205	Fast Food Restaurant	Bakery	Thai Restaurant	Convenience Store	Pizza Place	Mediterranean Restaurant	Sandwich Place	Park	Asian Restaurant	Donut Shop
91803	Asian Restaurant	Mexican Restaurant	Pizza Place	Fried Chicken Joint	Italian Restaurant	Szechuan Restaurant	Spa	Fast Food Restaurant	Vietnamese Restaurant	Hot Dog Joint
92395	Convenience Store	Café	Construction & Landscaping	Karaoke Bar	Pizza Place	Electronics Store	Donut Shop	Farm	Fast Food Restaurant	Fried Chicken Joint
93637	Mexican Restaurant	Gym / Fitness Center	Farm	Gift Shop	Wings Joint	Department Store	Donut Shop	Electronics Store	Fast Food Restaurant	Fried Chicken Joint
93955	Business Service	American Restaurant	Trail	Korean Restaurant	Basketball Court	Golf Course	Greek Restaurant	Grocery Store	Gun Range	Cuban Restaurant

As per the table above, every one of the new locations seem to be very accepting of the diverse world cuisine. That provides additional reason to believe that Patcy's pizzeria has a great chance of succeeding at any of the five new locations.