Title: A model to identify new neighborhoods that could support a new vegan / vegetarian restaurant

By Marina Santiago, January 2, 2021

Introduction / Business Problem

Vegetarian and vegan restaurants are becoming more and more popular, especially in high population cities like New York City, Los Angeles, and Chicago. Some of these restaurants have been very successful. For example, the Chicago Deli, in Chicago, has multiple locations. However, others have not. The New York City vegetarian restaurant, Nix, closed after four years. Is there a way to predict which neighborhoods are most likely to be able to support a vegetarian / vegan restaurant?

Because there are fewer vegetarians in smaller cities, opening vegetarian/vegan restaurants in smaller cities can be an even riskier proposition. However, if successful, a vegetarian / vegan restaurant can have high monetary rewards. We can decrease this risk by identifying neighborhoods that are most similar to neighborhoods in big cities that already support vegetarian/vegan restaurants.

Many types of people would be interested in this type of analysis. These include people interested in starting a restaurant, investors, developers, and even city officials looking to improve different parts of their city in order to attract more people. It could even be used by everyday people looking to move to a new neighborhood if they want to live in the type of area that might have a vegan / vegetarian restaurant in the future.

Data

For this project, I plan to train a machine learning model to identify neighborhoods that are likely to support a new vegetarian / vegan restaurant.

To train this model, I will use neighborhood data from NYC, LA, and Chicago, which all have many vegetarian vegan restaurants. I think it's important to use training data from a few different cities to control for any particular demographics or characteristics unique to one city.

This neighborhood data will consist of:

- Venue category data from Foursquare. I will determine what types of venues are within a half mile radius of the center of the neighborhood. I chose a half mile radius because that is a walkable distance for most people in most types of weather conditions.
- For each city, I have identified sources of data that can provide me with the following demographic metrics for each neighborhood. I chose these metrics because they roughly capture characteristics of a neighborhood without making the model overly complex.

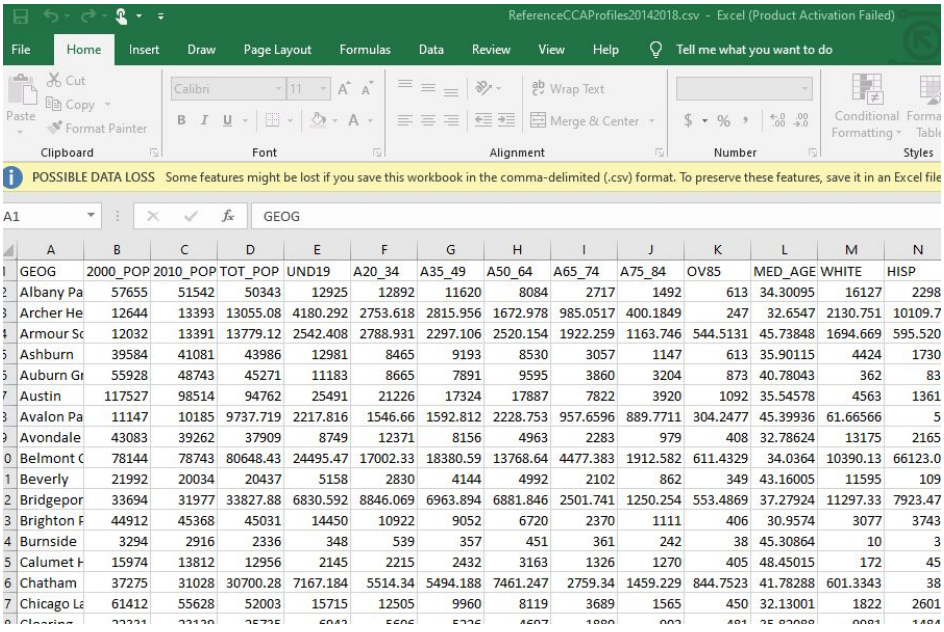
    - Median Income - I will use this metric as a proxy for how wealthy the neighborhood is
    - Percent over 65 years old - This is a proxy for the age of the community
    - Percent white - This roughly describes the racial makeup of the neighborhood

- Income and demographic data can be found at the links below. Screenshots of samples of the data are included below each link.

    - NYC: http://app.coredata.nyc
        - Example income data:

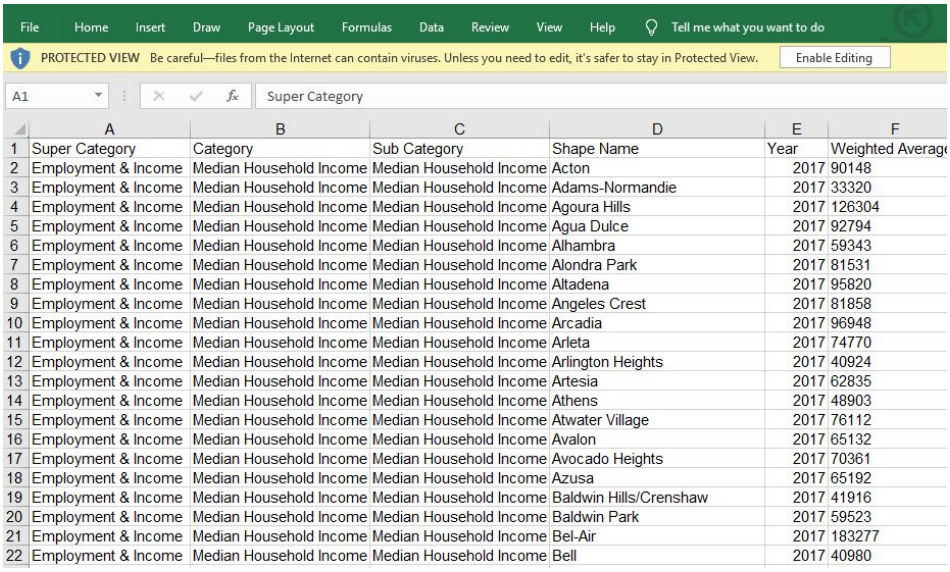
Sub-Borough Area	2000	2005	2006	2007	2008	2009	2010	2011	2012
Astoria	$57,619	$46,660	$53,078	$56,937	$61,424	$55,715	$53,730	$52,458	$57,107
Bay Ridge	$67,577	$60,180	$69,438	$60,340	$64,696	$62,182	$55,857	$63,490	$57,710
Bayside/Little Neck	$88,792	$81,950	$85,040	$85,921	$87,299	$84,651	$81,411	$86,495	$79,404
Bedford Stuyvesant	$35,831	$38,140	$34,784	$38,432	$45,537	$35,423	$42,712	$36,308	$41,960
Bensonhurst	$53,433	$49,670	$48,794	$50,860	$45,604	$45,307	$44,166	$49,929	$55,084
Borough Park	$49,818	$40,450	$50,465	$46,605	$47,353	$48,129	$41,737	$38,599	$41,644
Brooklyn Heights/Fort Greene	$64,591	$63,290	$69,261	$78,502	$73,358	$77,704	$82,910	$69,671	$83,410
Brownsville/Ocean Hill	$33,474	$27,480	$28,957	$31,006	$32,260	$30,354	$30,300	$28,933	$31,233

    - Chicago: https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data
        - Example demographic data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	GEOG	2000_POP	2010_POP	TOT_POP	UND19	A20_34	A35_49	A50_64	A65_74	A75_84	OV85	MED_AGE	WHITE	HISP
2	Albany Pa	57655	51542	50343	12925	12892	11620	8084	2717	1492	613	34.30095	16127	2298
3	Archer He	12644	13393	13055.08	4180.292	2753.618	2815.956	1672.978	985.0517	400.1849	247	32.6547	2130.751	10109.7
4	Armour Sc	12032	13391	13779.12	2542.408	2788.931	2297.106	2520.154	1922.259	1163.746	544.5131	45.73848	1694.669	595.520
5	Ashburn	39584	41081	43986	12981	8465	9193	8530	3057	1147	613	35.90115	4424	1730
6	Auburn Gr	55928	48743	45271	11183	8665	7891	9595	3860	3204	873	40.78043	362	83
7	Austin	117527	98514	94762	25491	21226	17324	17887	7822	3920	1092	35.54578	4563	1361
8	Avalon Pa	11147	10185	9737.719	2217.816	1546.66	1592.812	2228.753	957.6596	889.7711	304.2477	45.39936	61.66566	5
9	Avondale	43083	39262	37909	8749	12371	8156	4963	2283	979	408	32.78624	13175	2165
10	Belmont C	78144	78743	80648.43	24495.47	17002.33	18380.59	13768.64	4477.383	1912.582	611.4329	34.0364	10390.13	66123.0
11	Beverly	21992	20034	20437	5158	2830	4144	4992	2102	862	349	43.16005	11595	109
12	Bridgepor	33694	31977	33827.88	6830.592	8846.069	6963.894	6881.846	2501.741	1250.254	553.4869	37.27924	11297.33	7923.47
13	Brighton P	44912	45368	45031	14450	10922	9052	6720	2370	1111	406	30.9574	3077	3743
14	Burnside	3294	2916	2336	348	539	357	451	361	242	38	45.30864	10	3
15	Calumet H	15974	13812	12956	2145	2215	2432	3163	1326	1270	405	48.45015	172	45
16	Chatham	37275	31028	30700.28	7167.184	5514.34	5494.188	7461.247	2759.34	1459.229	844.7523	41.78288	601.3343	38
17	Chicago La	61412	55628	52003	15715	12505	9960	8119	3689	1565	450	32.13001	1822	2601

- LA: http://la.myneighborhooddata.org/data/
    - Sample income data



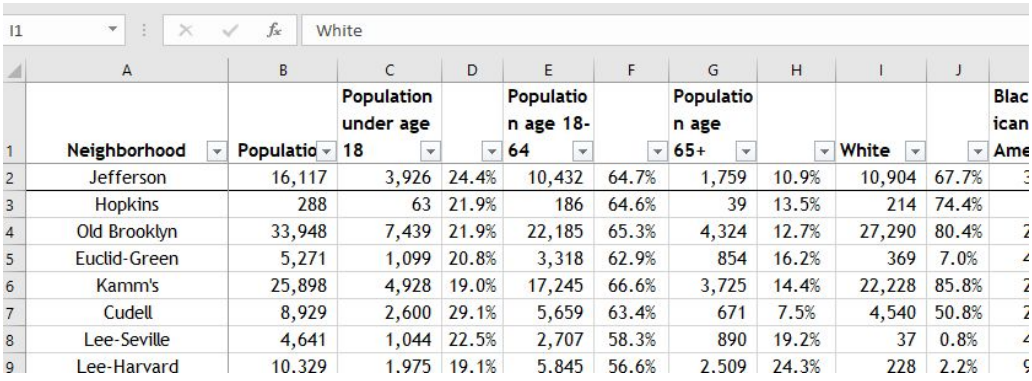
	A	B	C	D	E	F
1	Super Category	Category	Sub Category	Shape Name	Year	Weighted Averag
2	Employment & Income	Median Household Income	Median Household Income	Acton	2017	90148
3	Employment & Income	Median Household Income	Median Household Income	Adams-Normandie	2017	33320
4	Employment & Income	Median Household Income	Median Household Income	Agoura Hills	2017	126304
5	Employment & Income	Median Household Income	Median Household Income	Agua Dulce	2017	92794
6	Employment & Income	Median Household Income	Median Household Income	Alhambra	2017	59343
7	Employment & Income	Median Household Income	Median Household Income	Alondra Park	2017	81531
8	Employment & Income	Median Household Income	Median Household Income	Altadena	2017	95820
9	Employment & Income	Median Household Income	Median Household Income	Angeles Crest	2017	81858
10	Employment & Income	Median Household Income	Median Household Income	Arcadia	2017	96948
11	Employment & Income	Median Household Income	Median Household Income	Arleta	2017	74770
12	Employment & Income	Median Household Income	Median Household Income	Arlington Heights	2017	40924
13	Employment & Income	Median Household Income	Median Household Income	Artesia	2017	62835
14	Employment & Income	Median Household Income	Median Household Income	Athens	2017	48903
15	Employment & Income	Median Household Income	Median Household Income	Atwater Village	2017	76112
16	Employment & Income	Median Household Income	Median Household Income	Avalon	2017	65132
17	Employment & Income	Median Household Income	Median Household Income	Avocado Heights	2017	70361
18	Employment & Income	Median Household Income	Median Household Income	Azusa	2017	65192
19	Employment & Income	Median Household Income	Median Household Income	Baldwin Hills/Crenshaw	2017	41916
20	Employment & Income	Median Household Income	Median Household Income	Baldwin Park	2017	59523
21	Employment & Income	Median Household Income	Median Household Income	Bel-Air	2017	183277
22	Employment & Income	Median Household Income	Median Household Income	Bell	2017	40980

Then, I will use data from a smaller city (Cleveland, OH) to identify new neighborhoods where a new vegetarian / vegan restaurant is likely to do well. I will verify the model's accuracy by determining whether it predicted neighborhoods that already contain vegan / vegetarian restaurants. Once validated, this model could in theory be applied to any city.

The neighborhood data I will use can be found at:

- The same type of venue data from Foursquare as described above
- The same type of demographic data described above

- Additional data sources
    - Income and demographic data from: https://www.communitysolutions.com/resources/community-fact-sheets/cleveland-neighborhoods-and-wards/
    - Sample demographics data



	A	B	C	D	E	F	G	H	I	J	
1	Neighborhood	Populatio	Population under age 18		Populatio n age 18-64		Populatio n age 65+		White		Blac ican Ame
2	Jefferson	16,117	3,926	24.4%	10,432	64.7%	1,759	10.9%	10,904	67.7%	3
3	Hopkins	288	63	21.9%	186	64.6%	39	13.5%	214	74.4%	
4	Old Brooklyn	33,948	7,439	21.9%	22,185	65.3%	4,324	12.7%	27,290	80.4%	2
5	Euclid-Green	5,271	1,099	20.8%	3,318	62.9%	854	16.2%	369	7.0%	4
6	Kamm's	25,898	4,928	19.0%	17,245	66.6%	3,725	14.4%	22,228	85.8%	2
7	Cudell	8,929	2,600	29.1%	5,659	63.4%	671	7.5%	4,540	50.8%	2
8	Lee-Seville	4,641	1,044	22.5%	2,707	58.3%	890	19.2%	37	0.8%	4
9	Lee-Harvard	10,329	1,975	19.1%	5,845	56.6%	2,509	24.3%	228	2.2%	9

Methodology

Collecting demographic data

The first step in my methodology was processing the data from many different data sources into one cohesive table. The neighborhood demographic data for Chicago was already in the same .csv file, but the neighborhood demographic data for New York City and Los Angeles had to be downloaded in separate .csv or excel files and recombined into the same table. Furthermore, I had to calculate the percent white and percent over 65 and make sure that the data was all the correct data type.

Finding longitudes and latitudes

Once that was done, I was able to use geocoders from geopy to find the longitude and latitude for almost every neighborhood in each city. Some of the neighborhoods in my list did not come back with longitudes and latitudes even after removing the timeout error, but the number of missing neighborhoods was relatively small, so I just dropped those rows. Then, I concatenated all the tables from the different cities to create a large table of neighborhoods, their demographic data, and their longitudes and latitudes.

Getting and processing Foursquare venue data

To get the Foursquare venue categories data, I used my Foursquare account and the "getNearbyVenues" function we created in class. However, I increased the radius from 500m to 800m (roughly a half mile). I did this because I wanted to make sure I fully captured the types of venues in an area, and a half mile walk is a reasonable distance for most people to walk in most weather conditions. Then, I used one hot encoding and grouped the data by neighborhood to create a new data frame with venue frequency that I could analyze.

After this step, there were 469 different venue types. I was concerned that rare categories would skew the results and that if these categories were too specific, it would be hard to generalize to different cities. Therefore, I had to decrease the number of venue types. First I created a histogram of total venue frequency across all neighborhoods, but because the frequency is so small for many of them, the data was hard to interpret (Figure 1). So, I performed a Log10 transformation of the data to better visualize the distribution (Figure 2). At around -1.5, there is a natural dip in the data, but I wanted fewer categories, so I chose a cutoff of -0.5, where there is also a natural dip, and it will also decrease the number of venue categories by roughly 50%. This cutoff corresponds to a total venue frequency over all neighborhoods of greater than or equal to 0.316.

Figure 1. Histogram of venue frequency data

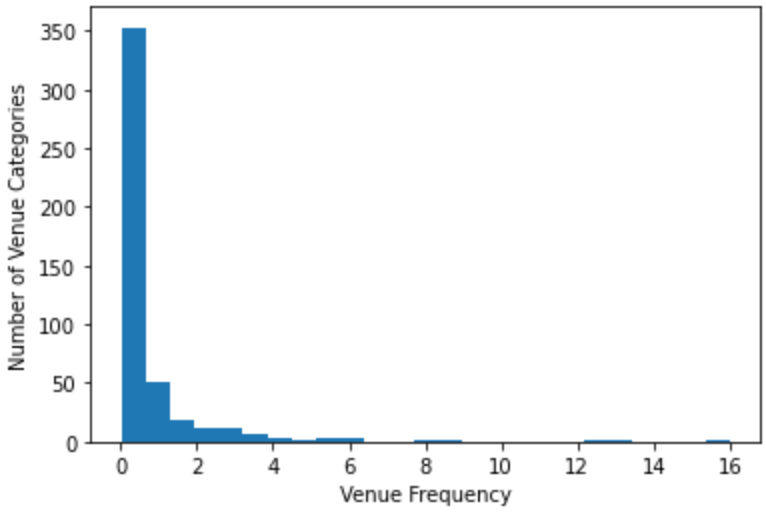
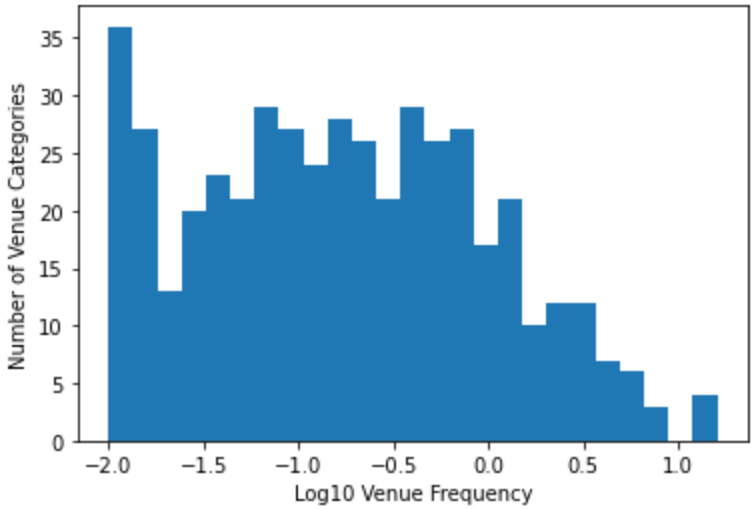


Figure 2. Histogram of Log10 transformed venue frequency data



After dropping columns with less than the frequency cutoff described above, I was left with 186 columns, a much more reasonable number to analyze. However, I decided to combine the columns such that they would be easier to generalize across different cities. For example, a city like New York City or Los Angeles is likely to have a lot of bodegas, but smaller cities that we can to use this data with won't, so we can combine bodegas with convenience stores. I created a file with the venue categories I wanted to combine, which resulted in 24 separate categories. I used this file to create a new dataframe with these features. For each new feature, I summed the venue frequencies of all the venue categories that went into the new category.

The new venue categories included:

- Stores
- Offices
- Indoor Entertainment
- Outdoor Entertainment
- Health
- Transportation
- Sports related venues
- Restaurants
- Bars / Brewery
- Automotive
- Cafe / Bakery / Dessert Shop
- Food / Groceries
- High School
- ATM
- Currency Exchange
- Flea Market
- Hotel
- Intersection
- Rental Car Location
- Salon / Barbershop
- Shopping Mall
- Storage Facility
- Waste Facility
- And finally our target variable: Vegetarian / Vegan Restaurant

Creating the training set

Now that the venue data was simplified, I added back in the demographic categories described earlier, making sure that all the neighborhoods were present, and that they were in the same order before adding the data back in. At this point, the model was ready to be built.

Building the model

I decided to use a decision tree classifier for this model because they are good for classification and prediction (the goal of this project), and they are easily interpretable. The training data consists of all the data described. I split the data into training and testing sets. I chose a test size of 50% because the frequency of Vegan / Vegetarian restaurants is so low that the model does not learn appropriately unless there is a similar number of restaurants in the training and testing set. I also used the ‘entropy’ criterion for the model.

Results

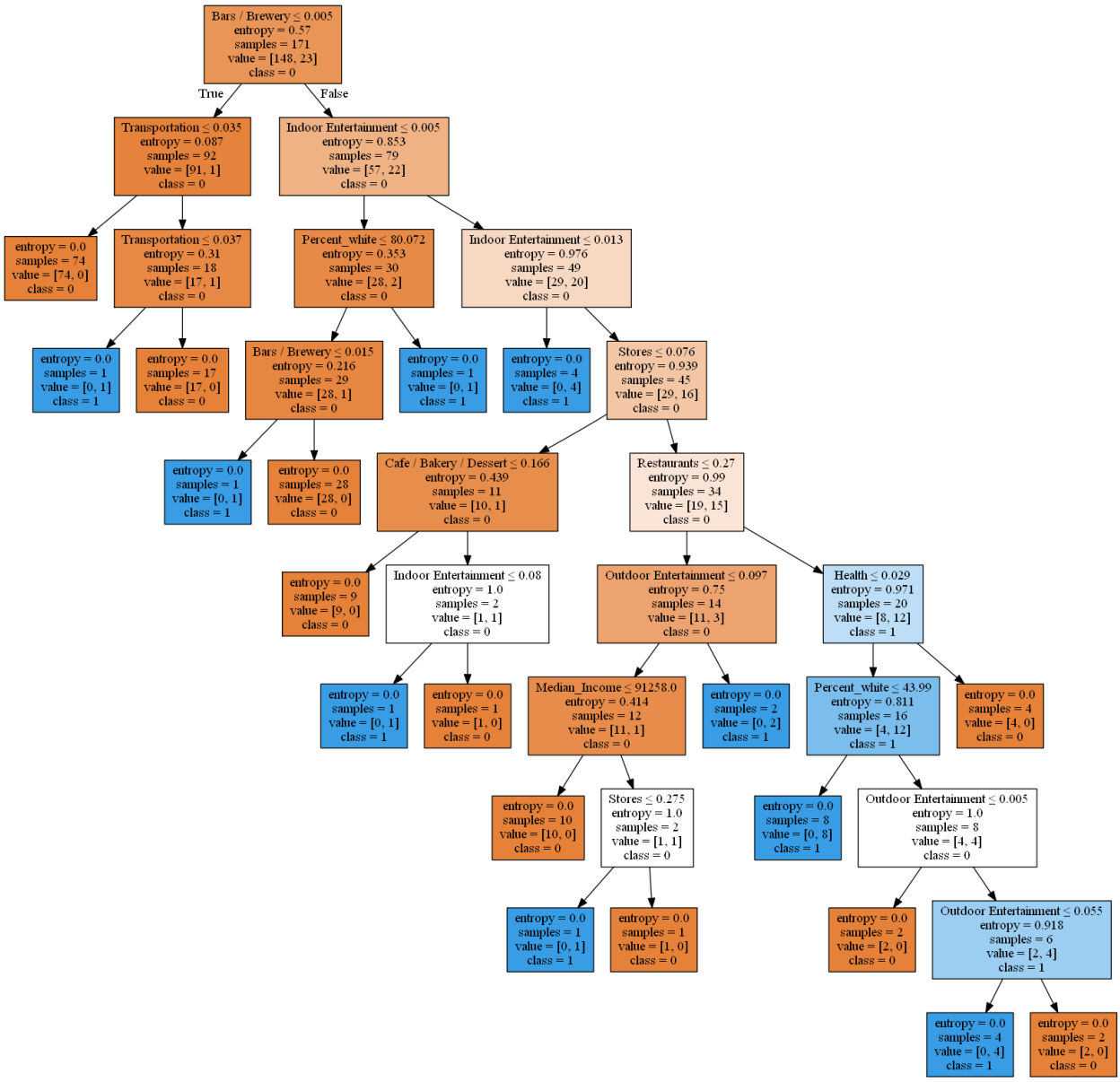
Using the training data, I ran the model a few times, and I found that the accuracy score was always roughly 81-82%. This is a reasonably good model. The top 5 features that were important for this decision tree are shown below (Figure 3)

Figure 3. Top 5 features and their importances

Feature	Importance
Bars / Brewery	0.290647
Indoor Entertainment	0.170038
Outdoor Entertainment	0.138880
Percent_white	0.095530
Transportation	0.081698

I visualized the decision tree (Figure 4) to better understand the model (see next page)

Figure 4. Decision Tree Classifier



To validate this model, I used data from a different city, Cleveland, OH to predict which neighborhoods could support a vegan / vegetarian restaurant. To do this, I processed the Cleveland demographic data in the same way I did the training data set, I found the longitude and latitude data using geopy, got the Foursquare venue categories data, and combined the categories to create a simplified dataset.

Then I used the model to predict which neighborhoods were likely to support a vegetarian / vegan restaurant. Of the 29 neighborhoods in Cleveland, this model predicts that 3 would support one.

These three include:

1. Buckeye-Woodhill
2. Detroit Shoreway
3. Euclid Green

I wanted to compare this list to the neighborhoods in Cleveland that already have a vegan / vegetarian restaurant. Only one neighborhood, Euclid Green, already has one. Because my model identified this neighborhood, it validates the model and suggests the other neighborhoods could support one as well!

Discussion

The model I created performed with 81-82% accuracy on the training set. This is remarkable and impressive considering that I used data from hundreds of neighborhoods in 3 different cities. I think that using this large and diverse training dataset helped increase the accuracy of the model. However, by incorporating data from more cities, the training set could be improved even more.

The model found that the top 5 most important features for the model. These features and their importances are shown in Figure 3. These features make sense because a restaurant is more likely to do well in areas with nearby entertainment and transportation infrastructure. Nearby bars also means that people will likely need to eat, so a restaurant is likely to be successful there. The fact that the model outputs features that you would expect to be important for a restaurant's success supports the model.

When testing the model on the Cleveland, Ohio data, I found that three neighborhoods (Buckeye-Woodhill, Detroit Shoreway, and Euclid-Green) were likely to support a vegetarian / vegan restaurant. According to livecleveland.org, the Buckeye-Woodhill neighborhood was once known as "Little Hungary" but now is known for its public art and urban farms/ the Detroit Shoreway neighborhood is one of Cleveland's most diverse neighborhoods along the north coast with many shops, restaurants, and pedestrian tunnels to help people walk around/ and the Euclid-Green area (according to niche.com) has a more urban/suburban mixed feel with a lot of bars and parks.

The only vegan / vegetarian restaurant in the Cleveland area is found in the Euclid-Green neighborhood. The fact that this model identified the only neighborhood with a vegan / vegetarian restaurant validates the model and suggests that the other Cleveland neighborhoods could also support new vegan / vegetarian restaurants. The one vegan / vegetarian restaurant in Cleveland, Tommy's Restaurant has been open since 1972, has 4.5 stars on TripAdvisor, and is ranked #43 of 1720 restaurants in Cleveland. They have also managed to stay open during the covid pandemic suggesting that they are doing well financially and that their community supports them.

Conclusion

In conclusion, I was able to create a model that predicts which neighborhoods can support a new vegetarian / vegan restaurant with more than 80% accuracy. This model could be improved by increasing the amount of training data, but even in its current state, it correctly identified the only neighborhood in Cleveland that currently has a vegan / vegetarian restaurant, validating the model. Furthermore, it identified two other neighborhoods that could potentially have a vegan / vegetarian restaurant. Further analysis is needed to confirm this model's recommendations, but based on the descriptions of these neighborhoods, they sound very promising.