Best neighborhoods in NY to open a new pizza place

Introduction:

I tried to find the best neighborhoods to open a pizza place in New York. Finding the neighborhoods with the best potential could be helpful for people looking to open a new pizzeria who want to find the most promising place to start their new business. This is based on the clear premise that the location influences the success a restaurant can have, and a careful analysis of existing businesses in the different city areas can help an investor make the right choice when choosing the place to start his new restaurant.

Data:

To segment the city of NY in different areas (neighborhoods), a json file provided in the IBM Data Science Professional Certificate course on coursera.com will be used.

The ML model used (k nearest neighbors regressor) requires external data to be trained (and tested on). For this purpose, I used data on the city of Chicago: this choice is due to the fact that a big city was needed (the more the data, the better the model), and that the city had to be possibly similar (culturally) to NY, which made a city in the US the best choice. While the two cities are certainly not the same, the model showed to work well on the city of NY when applied. (A later exam of the results showed that Pizza places tend to be slightly more successful in respect to other businesses in NY than in Chicago: this doesn't seem to be a problem for the analysis since it tends to act in a conservative way, recommending less areas). To segment the city of Chicago into different neighborhoods, data from wikipedia.com was used (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago). The python library geocoder was used to get the coordinates of the listed neighborhoods.

The machine learning model was trained to predict the frequency of pizza places based on the frequency of other types of businesses in the area: to do so, the Foursquare API was used to get lists of businesses around the examined location. Of the returned information, I was interested in the business type.

Methodology:

First, the above described data was retrieved, cleaned and organized. Two datasets, one for the city of New York and the other of Chicago, were created, containing the name and coordinates of the neighborhoods. Using the coordinates, a call to the Fourquare API retrieved venues nearby, and saved them. Since the model I used was only concerned with the venue type, all other information was removed from the dataset. The final datasets showed, for each neighborhood, the frequency of the different types of location present in both cities.

The ML algorithm used is K nearest neighbors regressor. The model was trained on the city of Chicago, using as predictors the frequency of businesses other than Pizza places, to predict the frequency of pizza places. The only hyperparameter of the model, the integer K, was tuned evaluating the

performance of the model on a separate validation set. Once the best value was found, the entire data was used to obtain a yet better model.

The third step was applying the model to the data on the city of New York, to predict the amount of Pizza places in each neighborhood. The prediction was the compared to the actual frequency, to see if the predicted potential was significantly higher than the actual value. Such a case is taken to indicate that a new pizzeria could be successful in the area, since the potential has not been reached.

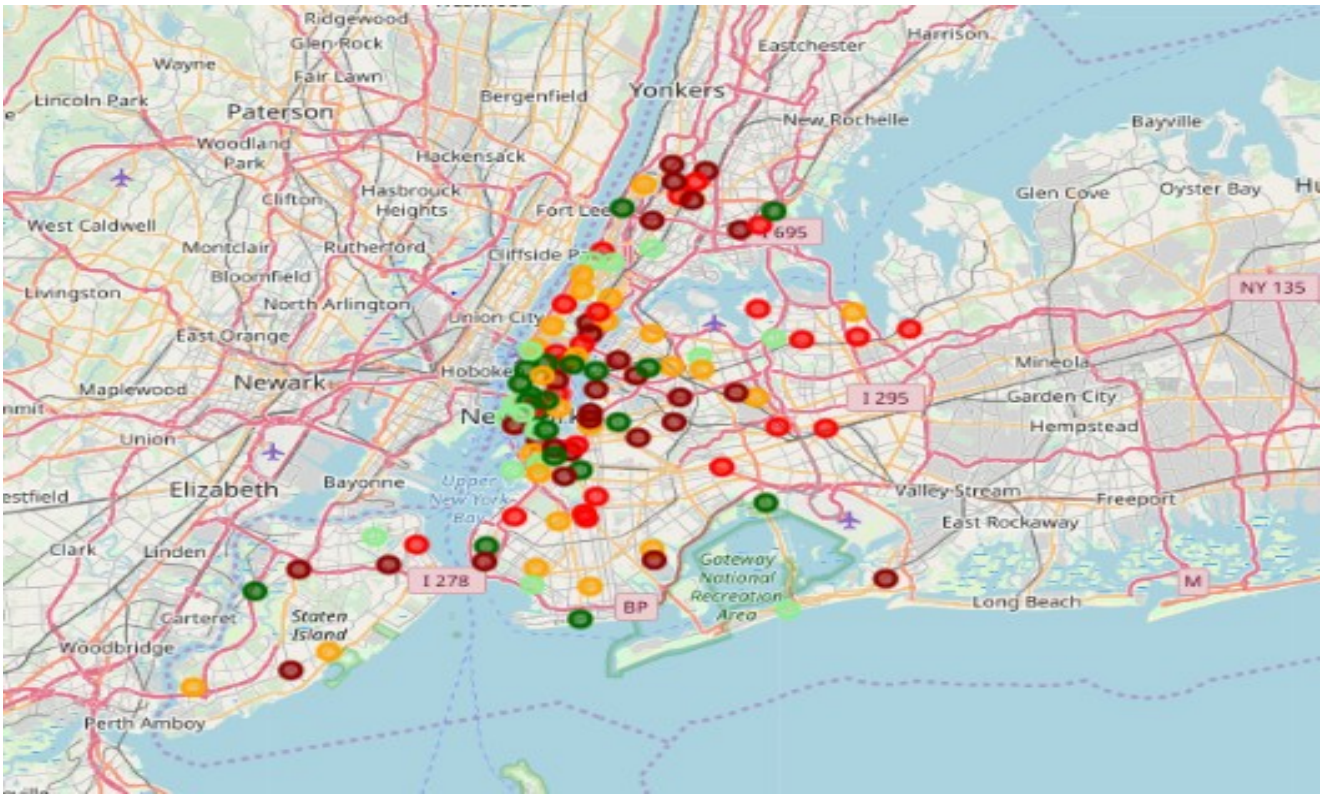
Lastly, the results were plotted on a Folium map to show visually which neighborhoods seem to have the best potential.

A first consideration: to use the model, I had to make sure it would work well on the city of New York when trained on the city of Chicago: this seems to be the case, since the average difference between the predicted frequency of pizza places in a neighborhood and the actual value is small. Hence, big differences in certain neighborhoods can be taken to represent a not-fully-used potential for pizza businesses.

Furthermore, this kind of analysis works on the assumption that neighborhoods follow some kind of pattern in the success of different kinds of venues. Hence, the success of pizza places in a specific neighborhood can be predicted by the amount of pizza places in neighborhoods that have similar patterns of businesses.

Results:

The results are a map of the city of New York showing visually the best areas, and a more telling and detailed table showing the difference between the expected absolute amount and frequency of pizzerias and the actual values.



Neighborhood	Latitude	Longitude	counts	Pizza Place	predicted	Frequency difference	Absolute difference
Greenwich Village	40.726933	-73.999914	100	0.010000	0.053854	0.043854	4.385410
Sunnyside Gardens	40.745652	-73.918193	100	0.030000	0.068807	0.038807	3.880698
Washington Heights	40.851903	-73.936900	85	0.023529	0.068738	0.045209	3.842752
Hunters Point	40.743414	-73.953868	75	0.013333	0.057851	0.044518	3.338857
East Williamsburg	40.708492	-73.938858	69	0.000000	0.045870	0.045870	3.165022
Chelsea	40.594726	-74.189560	105	0.009524	0.038717	0.029193	3.065236
...							
Sunnyside	40.612760	-74.097126	45	0.088889	0.000000	-0.088889	-4.000000
Lenox Hill	40.768113	-73.958860	100	0.050000	0.000000	-0.050000	-5.000000
Belmont	40.857277	-73.888452	98	0.091837	0.033983	-0.057853	-5.669620
Greenpoint	40.730201	-73.954241	100	0.060000	0.000000	-0.060000	-6.000000

While the table can be used to look at more precise results, the map is good at showing the most favorable regions and groups of neighborhoods, adding context to otherwise useful but separated data.

Discussion:

The resulting table for the best locations can be found in the Jupyter Notebook I published. Due to the inaccuracy of the model and scarsity of data, the results cannot be taken too seriously, and ought to be used in integration with other kinds of analysis. However, the model seems to show that neighborhoods such as Greenwich Village, Sunnyside Gardens, Washington Heights, Hunters Point are promising locations, whether Greenpoint, Belmont and Lenox Hill should be advised against.

Conclusion:

As has been hinted before, the analysis is clearly not complete and should be integrated with other information, such as the demographics of the different neighborhoods, the housing market and other data. An on-the-site evaluation of the situation cannot be replaced by a model such as mine, which can nonetheless act as a useful tool to be considered when deciding for the best location.