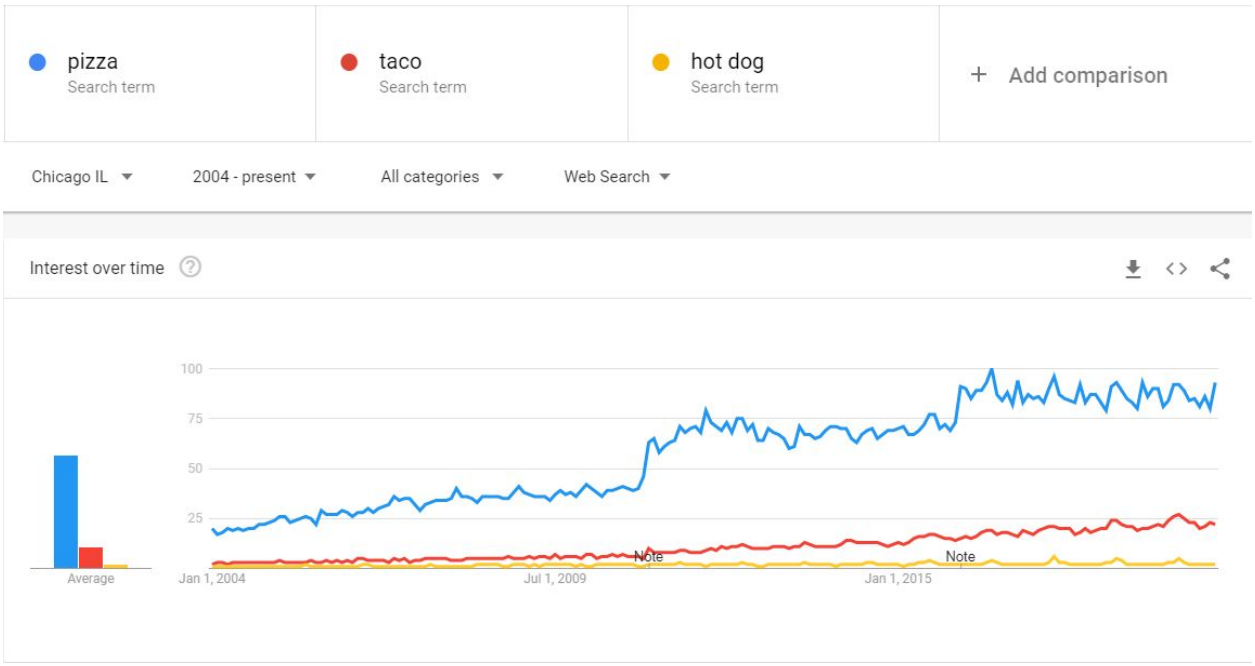
IBM Data Science Capstone Project

Using Foursquare Data to Determine Best Neighborhood in Chicago to Open New Pizza Place

Introduction

Background

According to Google Trends, Chicago is the #5 city when it comes to search popularity for the term “pizza” dating back to 2004. In fact, Chicago’s interest in pizza far outweighs other popular dishes such as tacos and hot dogs during that same time frame:



Chicago pizza even made national news in 2019 when the local police department made a Twitter post that provoked a response from the New York Police Department. This debate has been ongoing for years between residents of the two cities, neither of which are willing to give in.

So, it’s fair to say that Chicagoans are passionate about pizza. Tourists flock to try one of the large chain deep dish joints, while locals tend to order tavern style thin crust. Regardless of style, it’s important that all neighborhoods in Chicago have appropriate access to the dish.

Problem Statement & Research Question

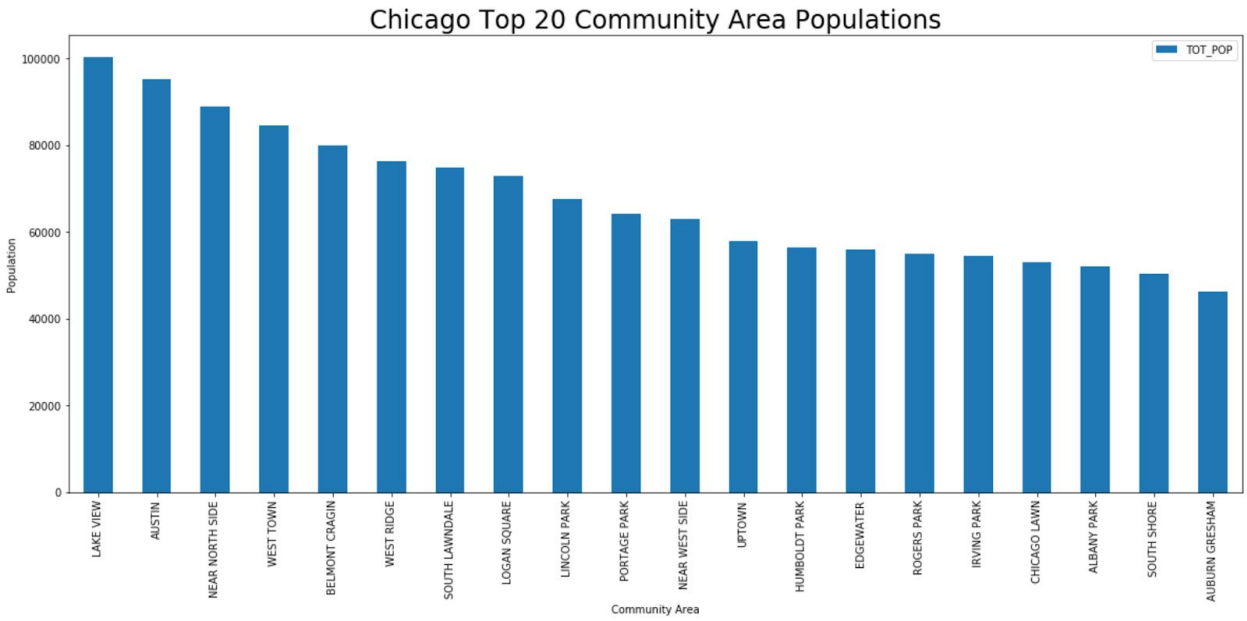
Are there neighborhoods in Chicago that are currently underserved by pizza places? It's unlikely that a "pizza desert" exists, but there may be neighborhoods with relatively few options when compared with others of similar composition. This knowledge would present an opportunity for small business owners to locate into a market that almost certainly would welcome additional options.

Data

Chicago Community Area Demographics

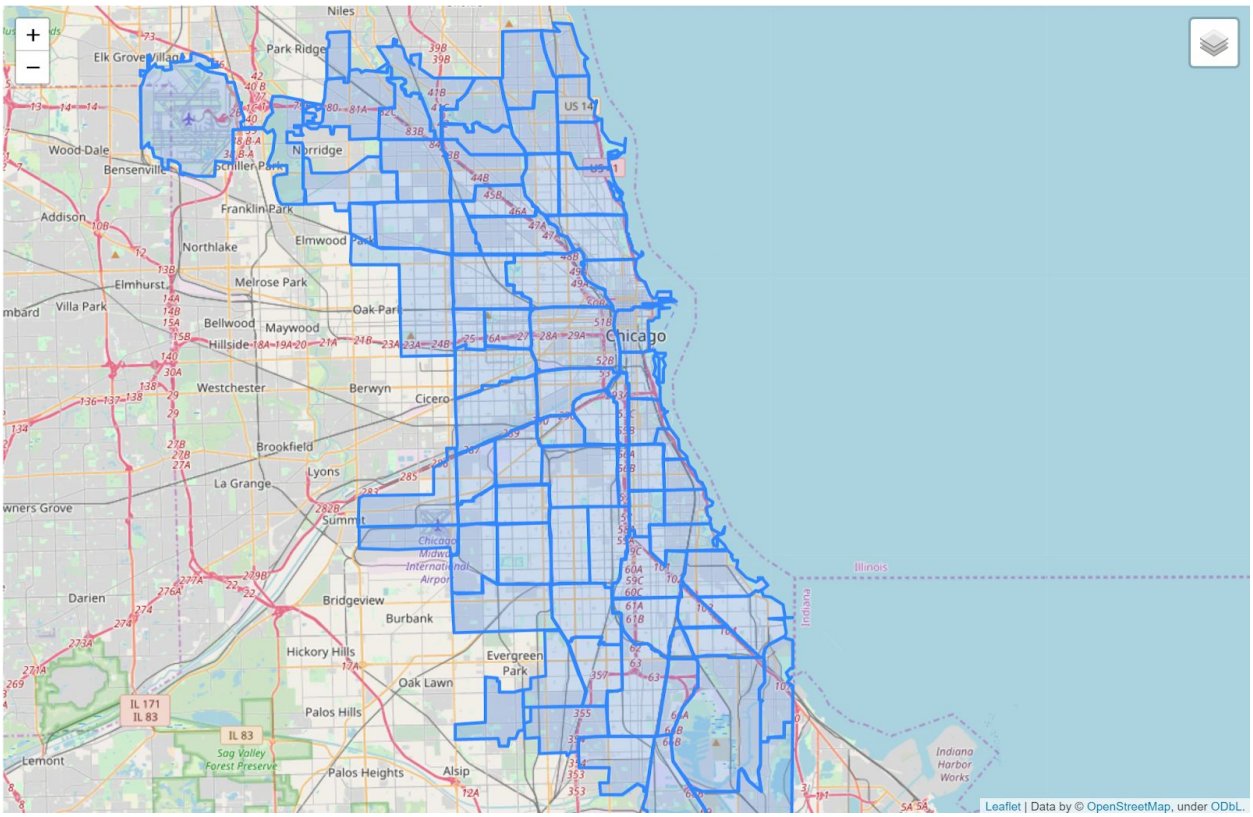
The state of Illinois maintains data on the demographic characteristics of the 77 official community areas (neighborhoods) in Chicago. This data spans 230 different features across the 77 neighborhoods, and can be found at the following URL:
https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data

As an example of the type of data available, here are the top 20 neighborhoods in Chicago ranked by total population:



Chicago Data Portal - Neighborhood Boundaries

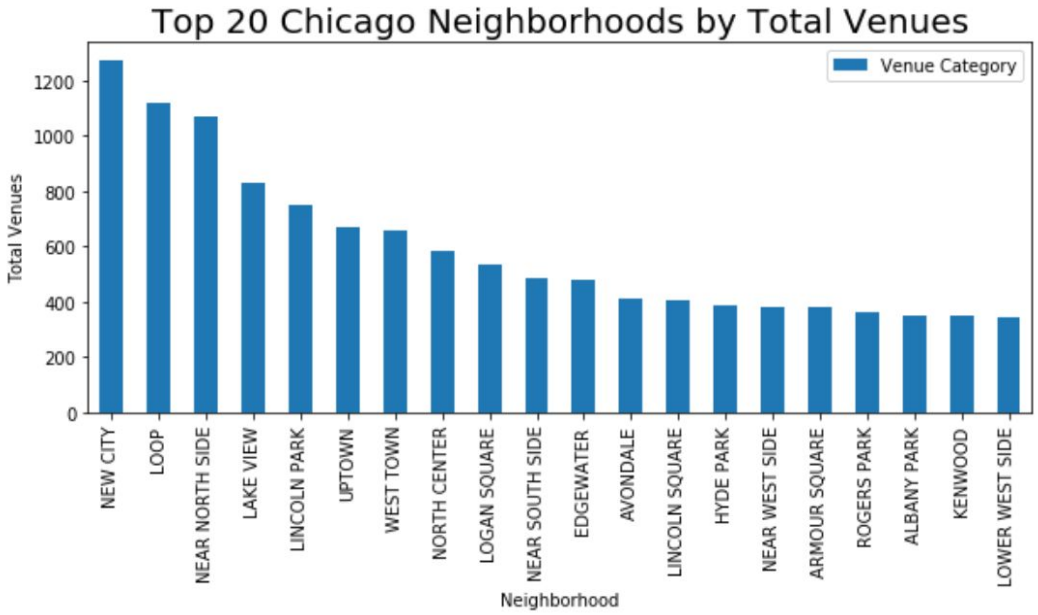
The city's data portal provides a neighborhood boundary geojson file that I'll be using to visually represent neighborhood level data, and can be found at the following URL: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9



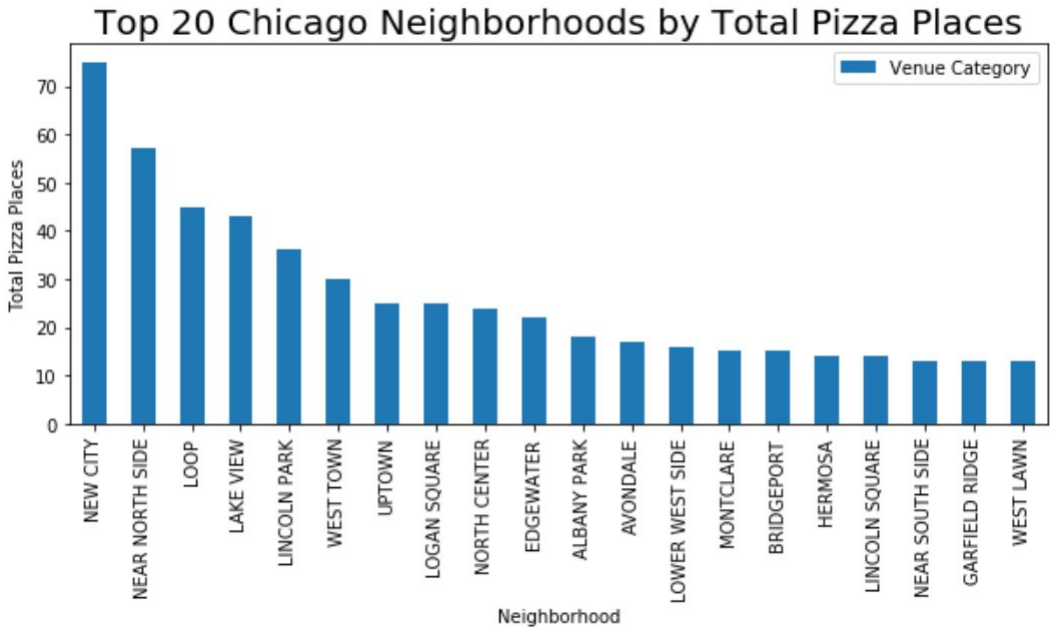
Foursquare API

Finally, I collected venue information for a 1 square mile radius in all of the neighborhoods. Ultimately, I obtained details for 23933 venues in Chicago. 880 of the venues were pizza places, which was the 3rd highest frequency of any venue category in the data collected.

Here are the top 20 neighborhoods by total venues collected:



Here are the top 20 neighborhoods by total pizza places. You'll notice that a lot of the same neighborhoods are appearing here. Because of the discrepancy in total venues collected by neighborhood, I'm ultimately going to be using Pizza Places as a percentage of total venues as my target variable.



To acquire propensity by venue type, I one hot encoded the neighborhood data, and then grouped by neighborhood and found the mean for each venue category. That gave me a venue data set of 484 features across the 77 neighborhoods.

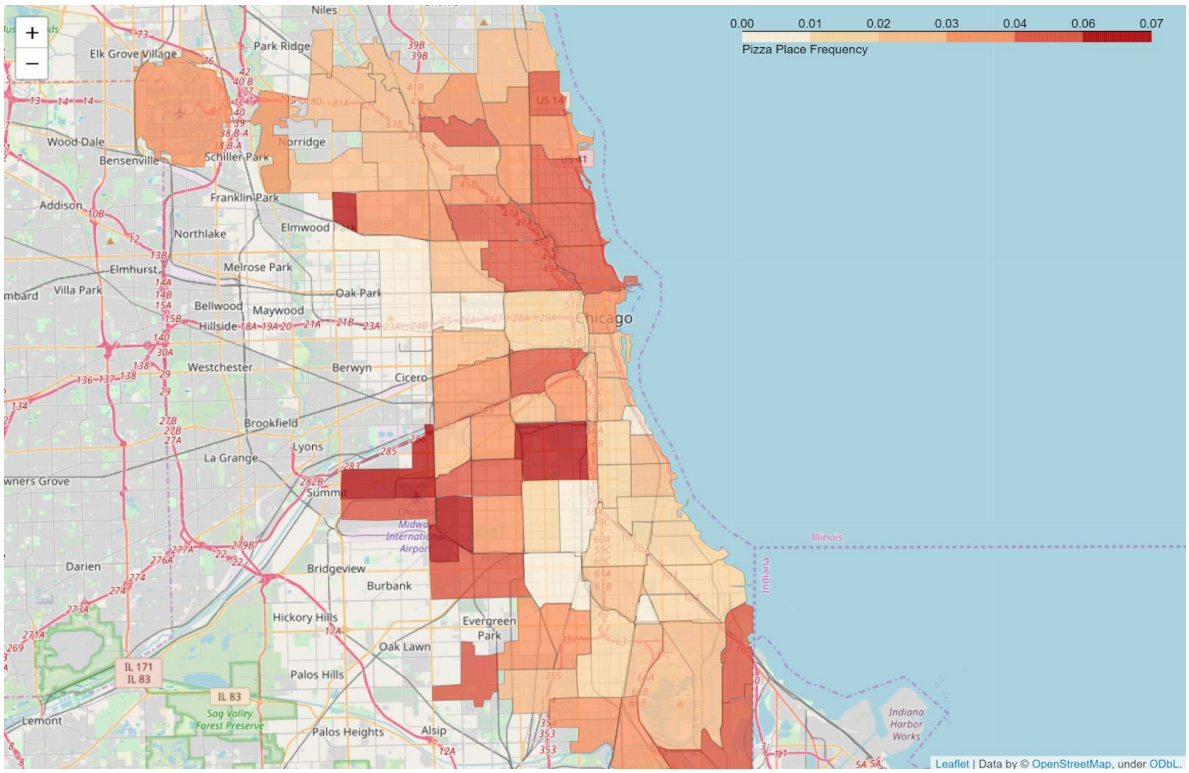
Methodology

The overarching goal is to find a regression method that best explains the variation of why some neighborhoods have a higher propensity of pizza places compared to others. Once this is established, we can use the same regression method to determine predicted pizza place propensity values for each neighborhood.

Using the predicted values, I will then compare with the actual values and find the neighborhoods with the greatest difference between predicted and actual. For example, if the predicted value is 5% but the venues in that neighborhood are only 3% pizza places, there is a significant opportunity for business owners.

Target Variable

As mentioned, I am using the Pizza Place propensity by neighborhood as the target variable in this project. We can visualize each neighborhood's pizza propensity on a choropleth maps:



It appears that some of the southwest neighborhoods have the highest propensity of pizza places in the city, while the near north neighborhoods all appear to have between 4% and 6% of their total venues being pizza places. It's interesting that there are some southern neighborhoods with very low pizza propensity neighboring those with high propensity. These neighborhoods may potentially be an opportunity area, or there could be a good reason for the disparity.

The top 5 neighborhoods for pizza places are as follows:

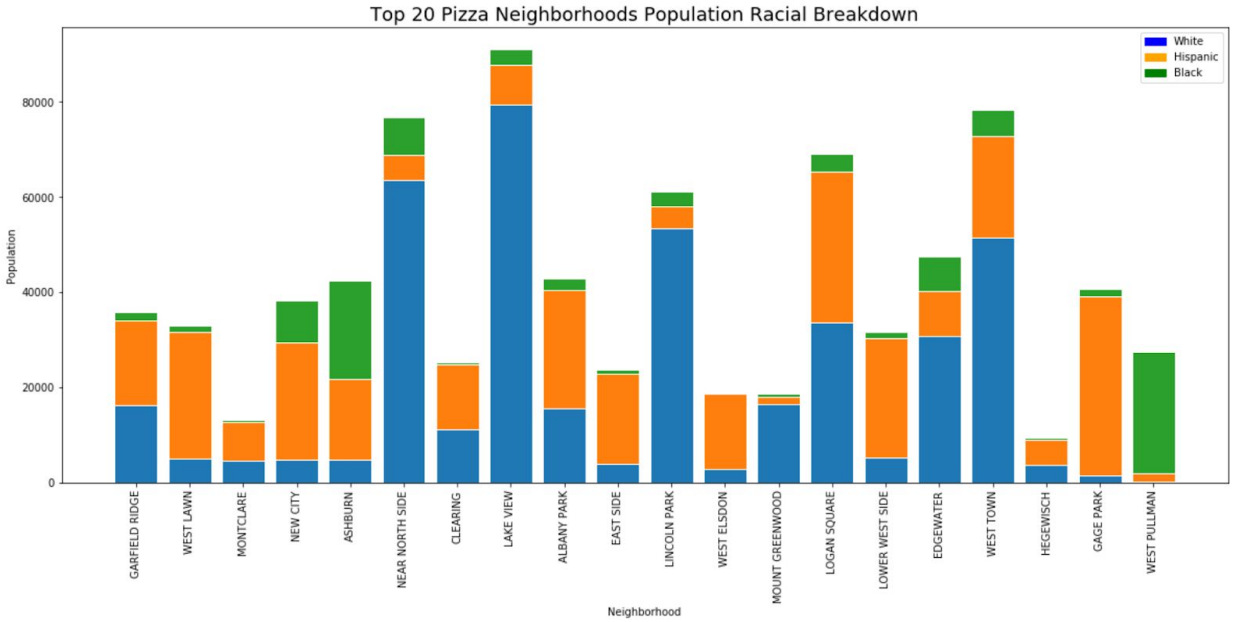
Neighborhood	% Venues Pizza Place
Garfield Ridge	6.7%
West Lawn	6.3%
Montclare	6.2%
New City	5.9%
Ashburn	5.5%

Feature Set

With 230 demographic features and 484 venue propensity features, it would be difficult to model without doing some initial feature selection. I used univariate regression testing to automatically eliminate 95% of the features, and only keep those which scored best when regressed against the target variable. I then reviewed correlations and eliminated redundant features from the final set:

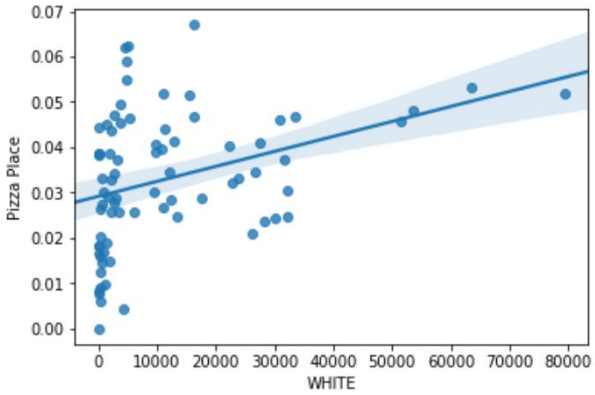
Demographic Variables	Venue Propensity Variables
WHITE	ATM
HISP	Bakery
BLACK	Bar
CARPOOL	Cosmetics Shop
INC_100_150K	Elementary School
MEDINC	Fast Food Restaurant
OWN_OCC_HU	Fried Chicken Joint
HV_150_300K	Gas Station
AVG_VMT	Gym
VACperc	High School
HCOV75K_LT20PCT	Hotel
in_lbr_frc_pct	Italian Restaurant
	Mexican Restaurant
	Park
	Southern / Soul Food Restaurant

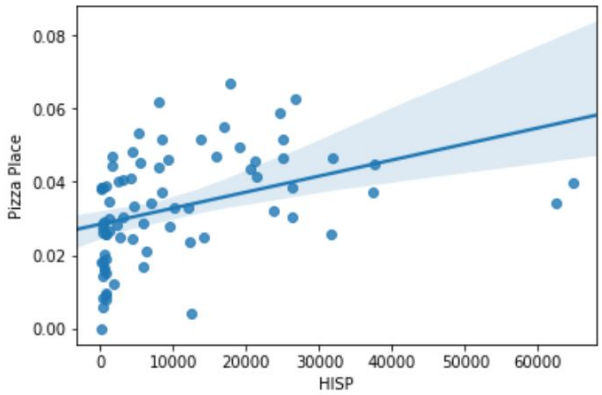
Three of the selected features (WHITE, HISP, BLACK) essentially explain the racial makeup of each neighborhood (although in a simplified manner - understanding there are more than three races of people present in Chicago). Just looking at the top 20 neighborhoods, we can look at racial breakdown:



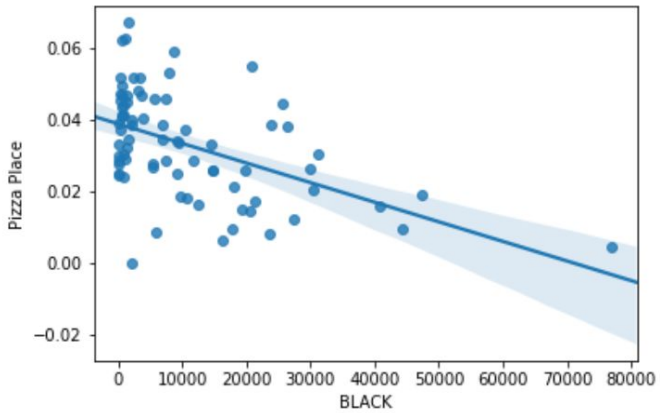
It's clear that these neighborhoods are almost all predominantly White or Hispanic, which also bears out in the regression plots to review relationships between the variables and the target:

White and Hispanic:



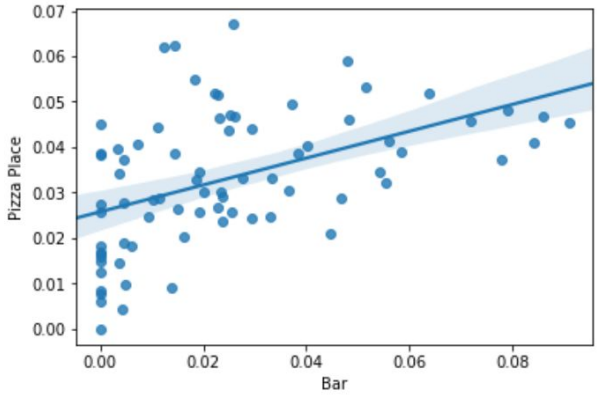


Black:

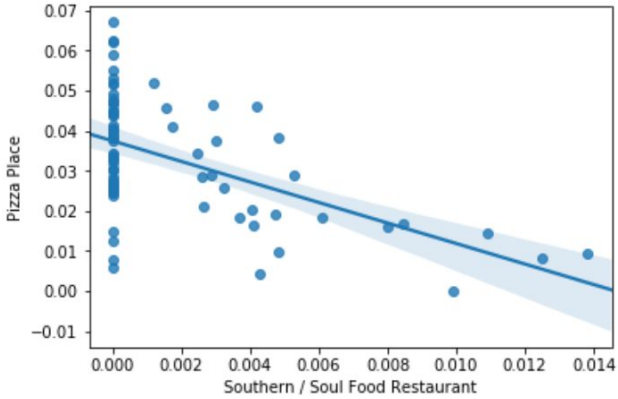


Additionally, there were clear relationships found between the selected venue propensities from the Foursquare variables (some examples below):

Bar:



Southern/Soul Food Restaurant:



Although this is just a sample of the selected features, all exhibit either a positive or negative relationship with the target variable. These will all be key when modeling and searching for neighborhoods that offer opportunity for additional pizza places.

Modeling

I decided to evaluate the 5 following regression methods:

1) Multiple Linear Regression
2) Ridge Regression
3) K Nearest Neighbors Regression
4) SVM Regression
5) Boosted Tree Regression

With each of these methods, I used a 5 fold cross validation to review test scores. The test scores are what I then used to evaluate each method. Additionally, I tested 3 different sets of features:

1) All features

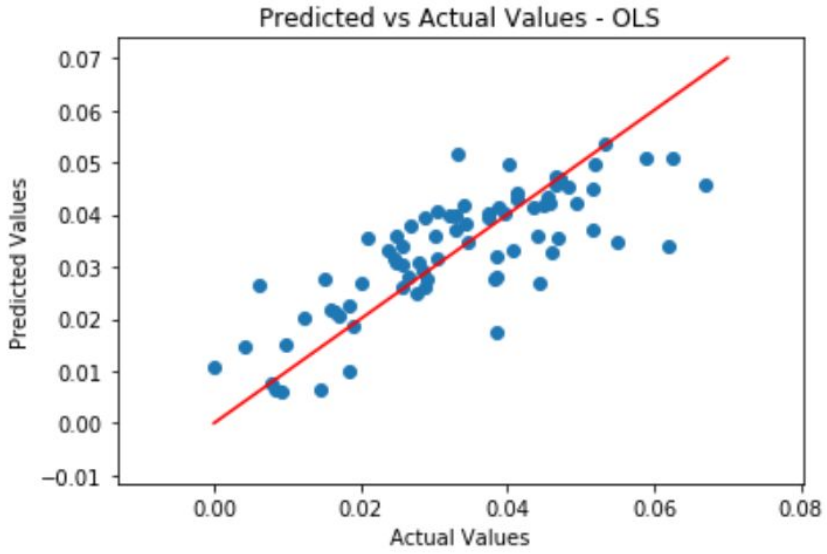
2) Only demographic features
3) Only Foursquare venue features

Multiple Linear Regression

With multiple linear regression, the best results were obtained using only the features derived from the Foursquare data collection. However, OLS yielded disappointing results overall:

Method	Training Score	Testing Score
Multiple Linear Regression	0.64	0.24

So, this method was only truly able to explain 24% of variance when presented with unknown data. The training scores indicate better fit, but that's to be expected. One advantage of OLS is that it didn't visually show any apparent bias when comparing actual and predicted values of the target:



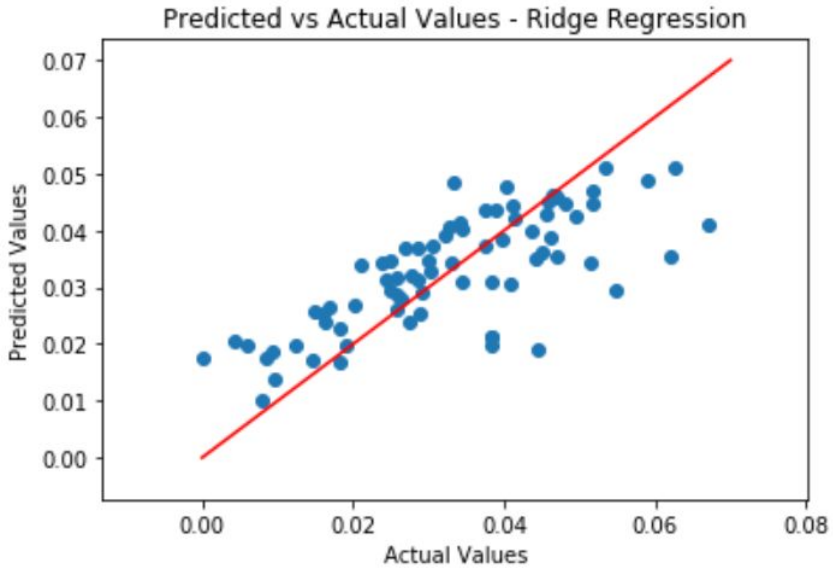
Results are well distributed both above and below the red line, which represents a perfectly fit model. So while we know that the method is not very good for predicting out of sample, it does seem to offer a good representation when looking in sample.

Ridge Regression

With ridge regression, the best results were again obtained using only the Foursquare venue data along with a light regularization factor of 0.005.

Method	Training Score	Testing Score
Ridge Regression	0.56	0.37

The regularization factor reduced in sample score when compared with OLS, but resulted in a 13% increase in out of sample testing scores. This is a significant improvement and the resulting scatter plot displays a slightly tighter distribution of results:



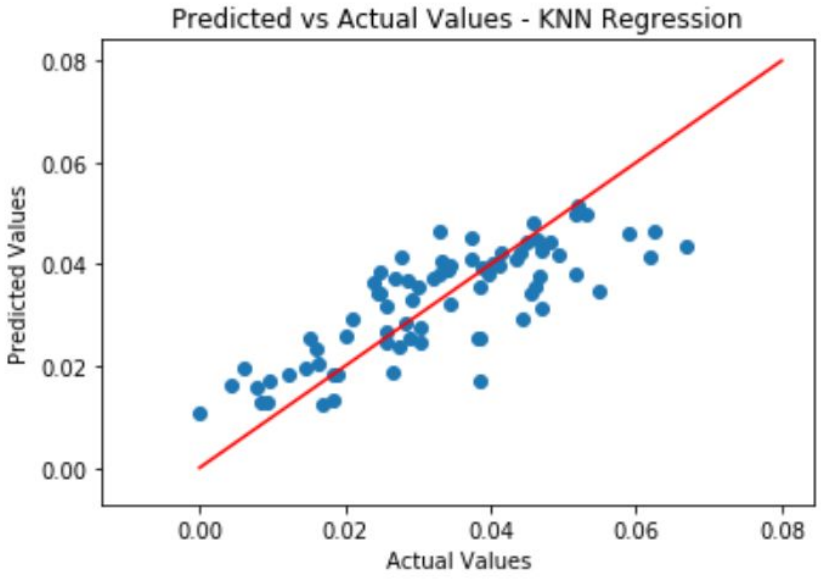
This method does seem to have a slight bias toward overpredicting lower values and underpredicting higher values.

KNN Regression

With KNN regression, the feature set that offered the best modeling results with the full feature set after normalization. Normalization is necessary for distance based methods, otherwise features with larger scales would be given greater weight when modeling. The best results were found using 4 nearest neighbors:

Method	Training Score	Testing Score
KNN Regression	0.62	0.41

KNN regression again offered an improvement in test score in comparison with the more sophisticated prior tested methods, which is a testament to simplicity. The results scatter plot again looks similar, but with a slight improvement in that its bias against predicting high values is less drastic:

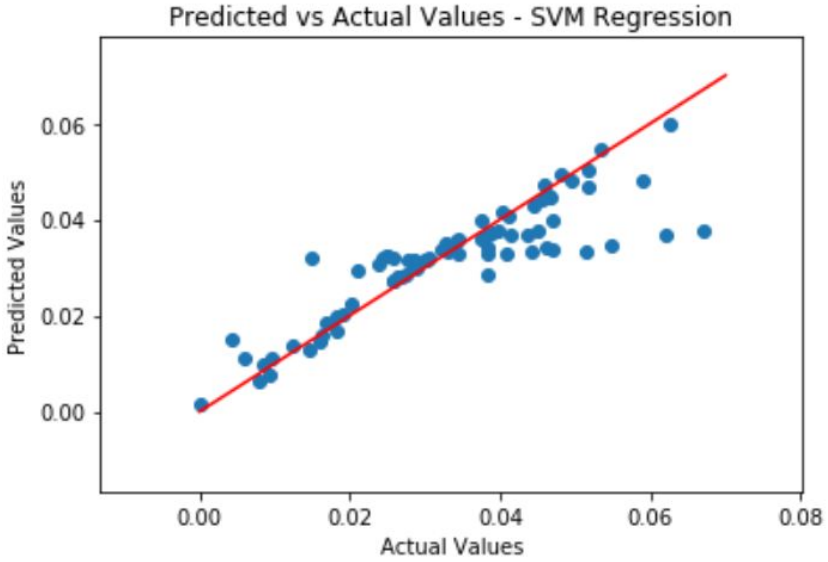


SVM Regression

The best results for SVM regression were found when using the venues data from Foursquare. Similar to KNN regression, I needed to use a normalized variant of the feature data because SVM's sensitivity to distance measures. An additional measure that was needed for SVM regression was using a normalized version of the target variable. I found that without normalizing the target, the model was incapable of predicting values other than the mean, which is no better than using a naive method.

Method	Training Score	Testing Score
SVM Regression	0.76	0.39

SVM offered a very tight prediction distribution in the low range, but again was unable to predict values at the higher end of the propensity range:

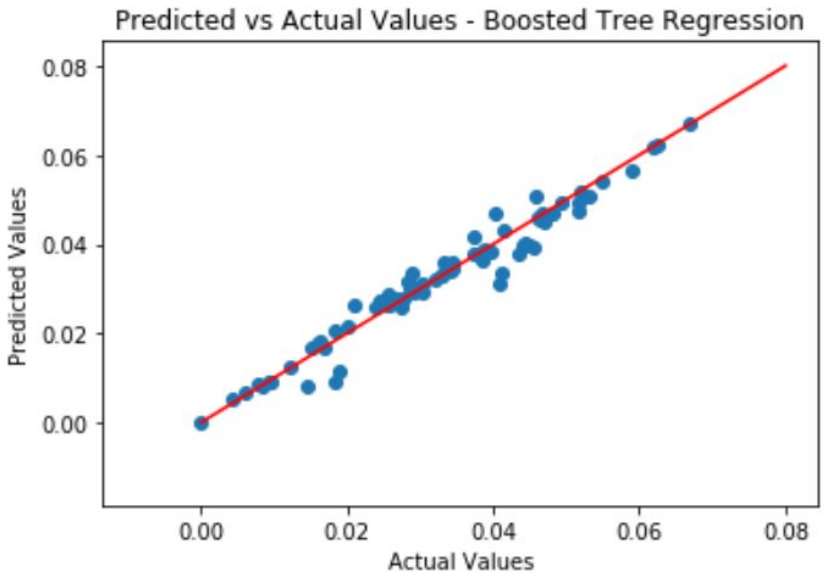


Boosted Tree Regression

The best results for boosted tree regression (using adaboost) came when using the full dataset of both demographic and Foursquare data. This intuitively makes sense because it gives the tree more options to choose from when developing branches and it can come up with its own hybrid set of features from both the demographic and venue variables.

Method	Training Score	Testing Score
Boosted Tree Regression	0.97	0.37

Ultimately, the boosted tree method was an overfit disaster. Out of sample predictions were slightly worse than other methods but it's clear from the highly divergent train scores during cross validation that the method was flawed for this specific application.



This method was certainly able to predict the higher values, but it seems like there is a heavy tradeoff in that it isn't at all able to be generalized.

Results

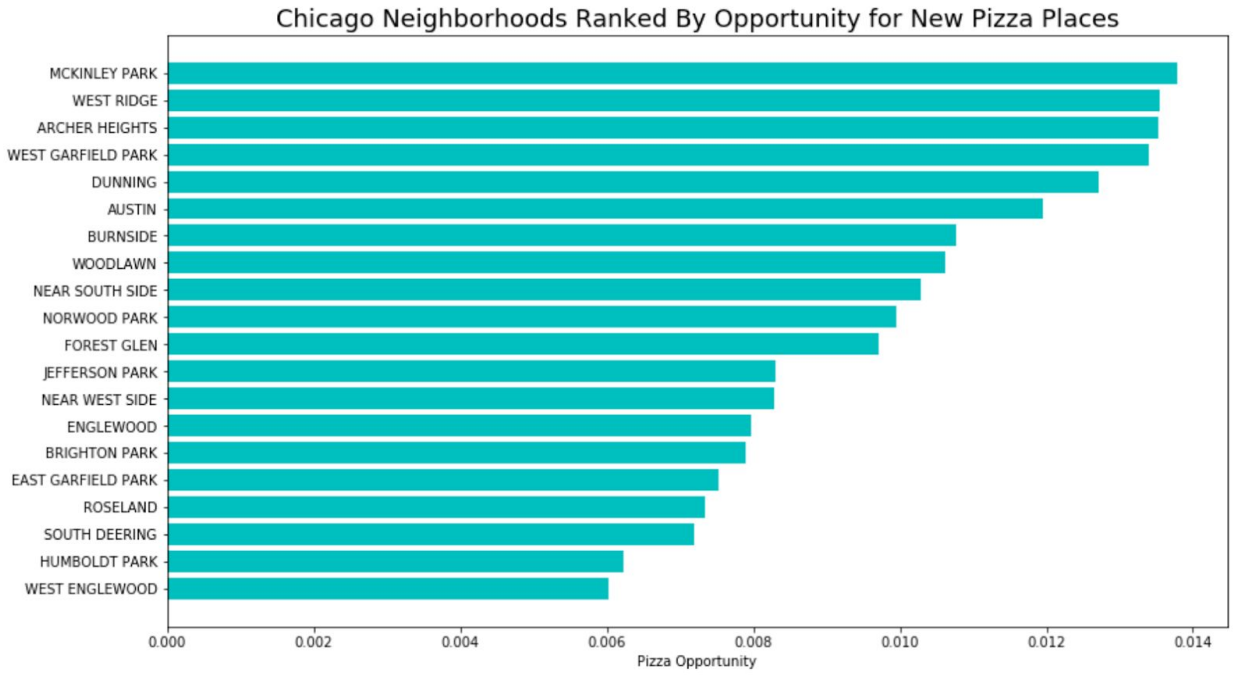
The final ranked results after all modeling was completed is as follows. Overall, KNN regression offered the highest test scores, and will be used to determine any actionable insights from the data.

As discussed at the onset of the methodology section, I'll use predicted values from the selected regression method to compare with

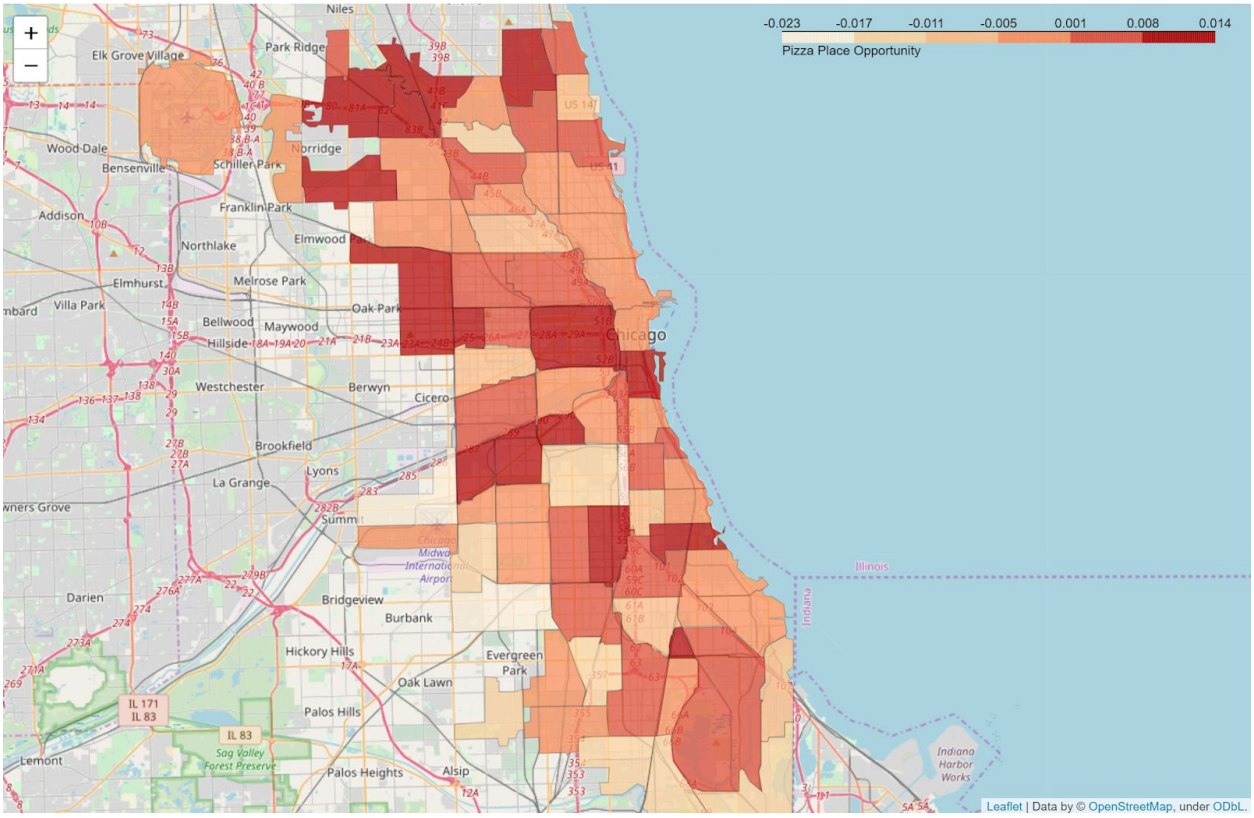
Method	Test Score	Comments
KNN Regression	0.41	Best fit, low bias in results
SVM Regression	0.39	Excellent fit on low values, but unable to predict higher values
Boosted Tree Regression	0.37	Heavily overfit to train data
Ridge Regression	0.37	Light regularization improved performance compared with OLS
Multiple Linear Regression	0.24	Poor performance overall

Discussion

Using the predicted values from KNN regression, I found the neighborhoods with the greatest discrepancy between predicted value and actual percentage of pizza places (represented below as “Pizza Opportunity”:



Neighborhoods such as McKinley Park, West Ridge, Archer Heights, and West Garfield Park scored very high on this measure. I can visualize this metric on a map as well to look for Pizza Opportunity hot zones:



It appears that neighborhoods to the immediate West and South of the downtown Loop district present an opportunity to business owners, although this may be a high rent area.

Additionally, a series of neighborhoods clustered together on the Southwest side (McKinley Park, Brighton Park, and Archer Heights) are all along a public transit corridor and all rank in the top 15 for pizza opportunity. They all neighbor one of the highest pizza place propensity neighborhoods in the city: New City. These three neighborhoods would all be excellent locations to open a new pizza place.

Conclusion

In summary, demographic and venue information collected from Foursquare were used to determine a regression model and predict which neighborhoods in Chicago presented most opportunity for additional pizza places. It was determined that KNN regression provided the best method of those tested. After comparing predicted values with actual values, it was determined that a number of neighborhoods presented an opportunity with the top 5 being:

1) McKinley Park
2) West Ridge
3) Archer Heights
4) West Garfield Park
5) Dunning

While these neighborhoods present the most opportunity, I think I speak for all Chicagoans by saying that no matter the neighborhood, we will welcome additional options!