Regression Analysis to Determine Neighborhoods Underserved by Restaurants

IBM Applied data science capstone project

Adam Epstein

2020-03-28

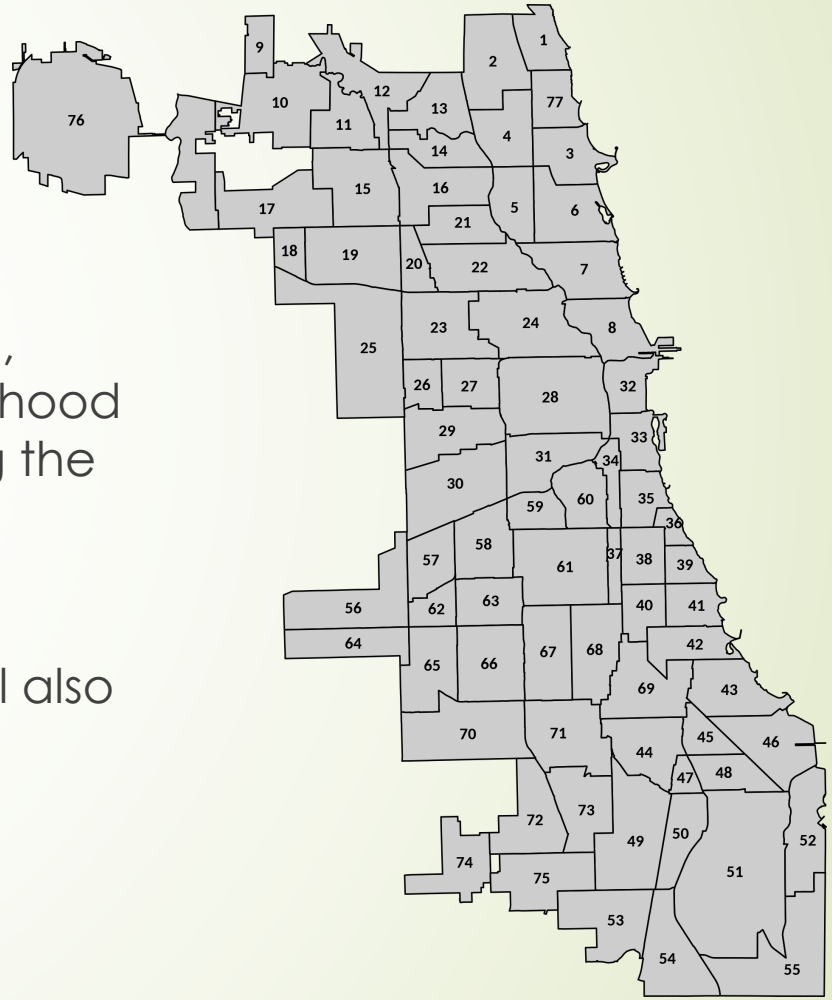
Problem

- Up to 60% of new restaurants close within three years of opening[1]
- Selecting the right location is critical for a new restaurant
- This analysis will investigate the potential of Foursquare data to identify neighbourhoods for a new restaurant

1. https://daniels.du.edu/assets/research-hg-parsa-part-1-2015.pdf

Data

- 77 Chicago neighbourhoods will be analyzed
- Using the coordinates of each neighbourhood, businesses within 2 kilometres of the neighbourhood centre will be identified and categorized using the Foursquare API
- Population density of each neighbourhood will also be used



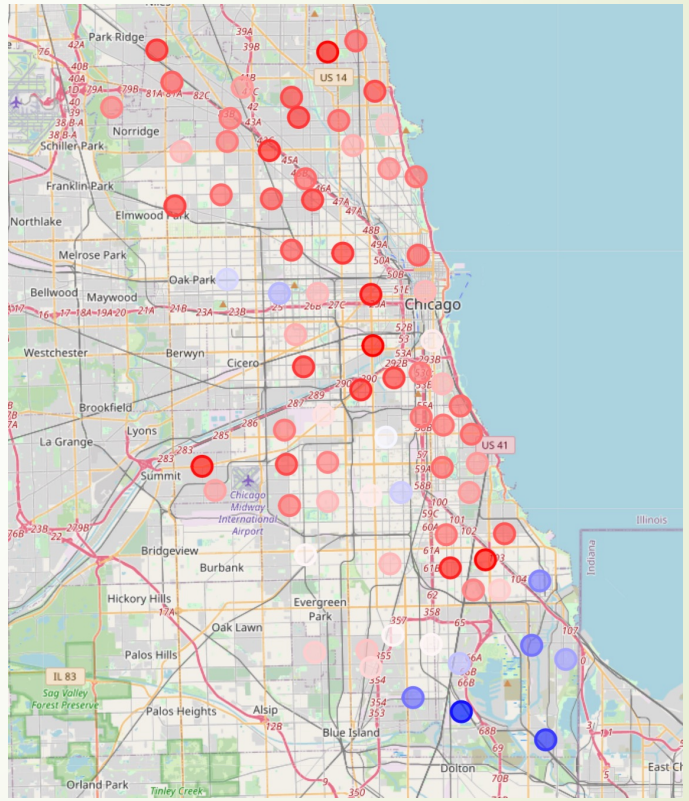
https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Methodology

- Train a regression model using business categories other than restaurants as the input and number of restaurants as the output
- Based on the predicted number of restaurants, identify neighbourhoods with less restaurants then expected.
- Three different regression models will be tested. Due to low number of training samples, regularization will be important to prevent overfitting

Results – Exploratory Data Analysis

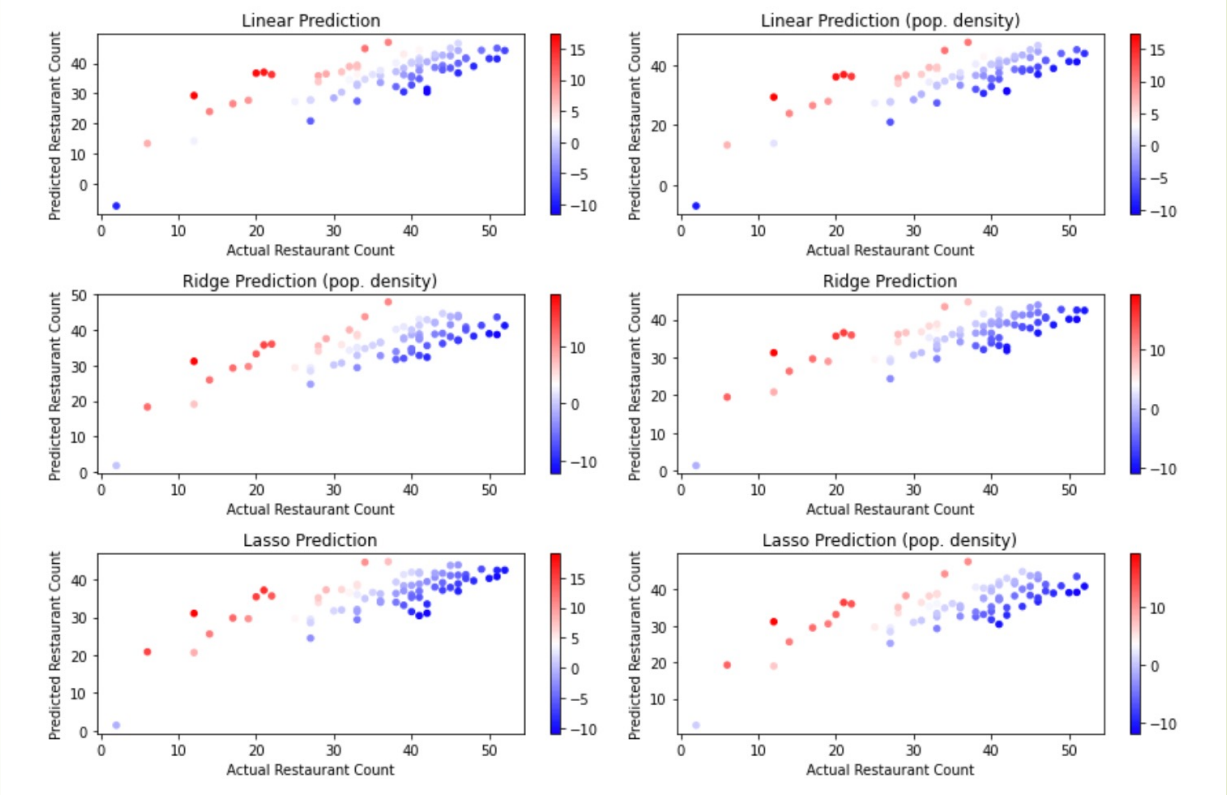
- As part of exploratory data analysis, a heat map was generated based on restaurant counts by neighbourhoods
- Neighbourhoods with low restaurant counts may be a good starting point
- More detailed analysis may show that low number of other businesses and low population density show that these neighbourhoods have the expected number of restaurants



Heat map showing restaurant count by neighbourhood

Results – Linear Regression

- Plots of the expected versus actual restaurant counts show similar results for different models
- Adding in population density in the second column of graphs does not significantly change the predictions



Results – Linear Regression Coefficients

- Analysis of coefficients show that the linear regression model may be overfitting due to high coefficients
- The regularization models significantly reduced the coefficients with only a minor impact of R^2
- Population density did not improve any of the models tested

Model	Arts & Entertainment	College & University	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Pop. Density	R^2
Linear Prediction	18.3	0.0	-31.6	-136.0	-43.0	576.2	-51.7	-86.6	N/A	0.615
Ridge Prediction	21.4	1.7	-16.8	-102.6	-4.7	1.7	-37.0	-70.5	N/A	0.577
Lasso Prediction	0.0	0.0	-0.0	-110.6	-0.0	0.0	-35.4	-64.0	N/A	0.565
Linear Prediction (pop. density)	18.6	10.3	-29.2	-133.9	-43.8	566.0	-49.5	-84.1	1.6	0.616
Ridge Prediction (pop. density)	21.0	2.8	-11.4	-95.7	-5.1	1.7	-29.6	-62.7	6.8	0.576
Lasso Prediction (pop. density)	0.0	0.0	-0.0	-100.8	-0.0	0.0	-27.0	-55.6	6.9	0.562

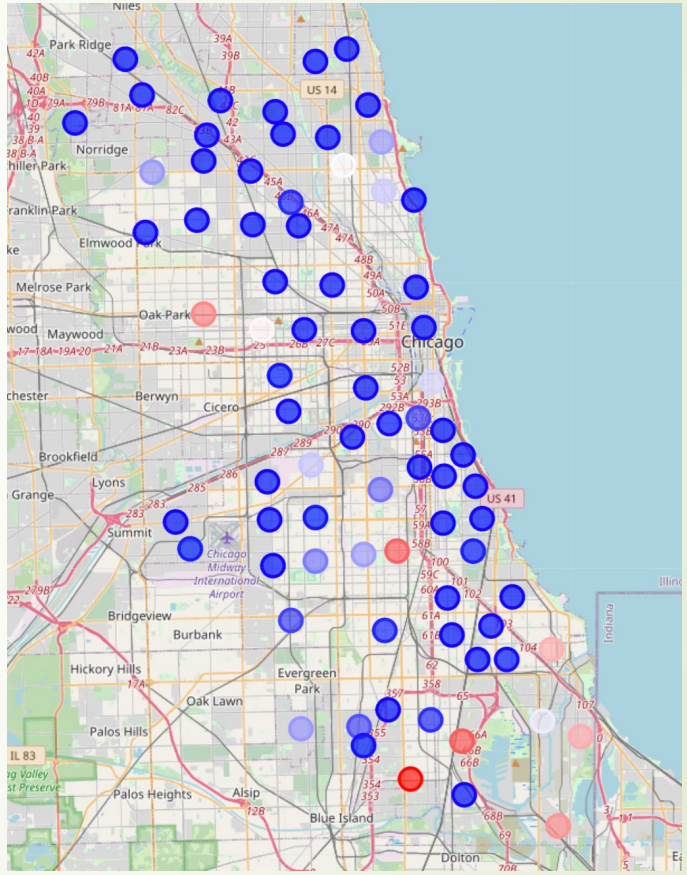
Target Neighbourhoods

- Nine neighbourhoods varied by more than 10 restaurants
- Seven of the nine are good candidates for a new restaurant
- Two of the nine appear to have significantly more restaurants than they can support

Neighbourhood	Actual Restaurant Count	Predicted Restaurant Count	Restaurant Shortage
West Pullman	12	31.2	19.3
Pullman	20	35.7	15.7
Englewood	21	36.5	15.6
Austin	22	35.9	13.9
Hegewisch	6	19.5	13.5
East Side	17	29.6	12.6
South Chicago	14	26.3	12.4
Rogers Park	42	31.9	-10.1
Avalon Park	51	40.1	-10.9

Target Neighbourhoods

- Good candidate neighbourhoods can be found throughout the city, with the most promising options in the south end.
- This analysis is meant to be a high level assessment. More detailed analysis incorporating additional demographics of the neighbourhood and the types of restaurants should be completed



Heat map showing restaurant shortage (red) or surplus (blue) by neighbourhood

Conclusions

- Foursquare venue data can be used to identify good locations for new restaurants. Seven neighbourhoods in Chicago were identified as good targets.
- This analysis can be extended to other cities and other business types
- The results would be improved by adding additional data and training with data in multiple cities.