Venue Location Recommender

Final Report

Introduction/Business Problem

- When someone is looking to open a venue such as a shop, restaurant or bar, how do they decide where to open it?
- One option is to look for areas that have fewer of that type of venue nearby (lower density)
- On its own this strategy is likely to be insufficient because an area may have lower density of this type of venue because there is limited demand for it
- In order to make a better decision they also need to able to identify areas where demand for the venue is likely to be high
- If we can quantify both the density of similar venues around a location and the likely demand, good potential locations could characterised as having low density compared to demand
- Quantifying density of venues around a location is a fairly trivial task, but how can we estimate demand?
- One solution is to assume that demand is related to the density of other venues around the location that are not of the same type as the venue being opened. For example if you are seeking to open a café, demand may be indicated by the range and number of non-café venues such as work places, shops and other restaurants nearby.

Data used

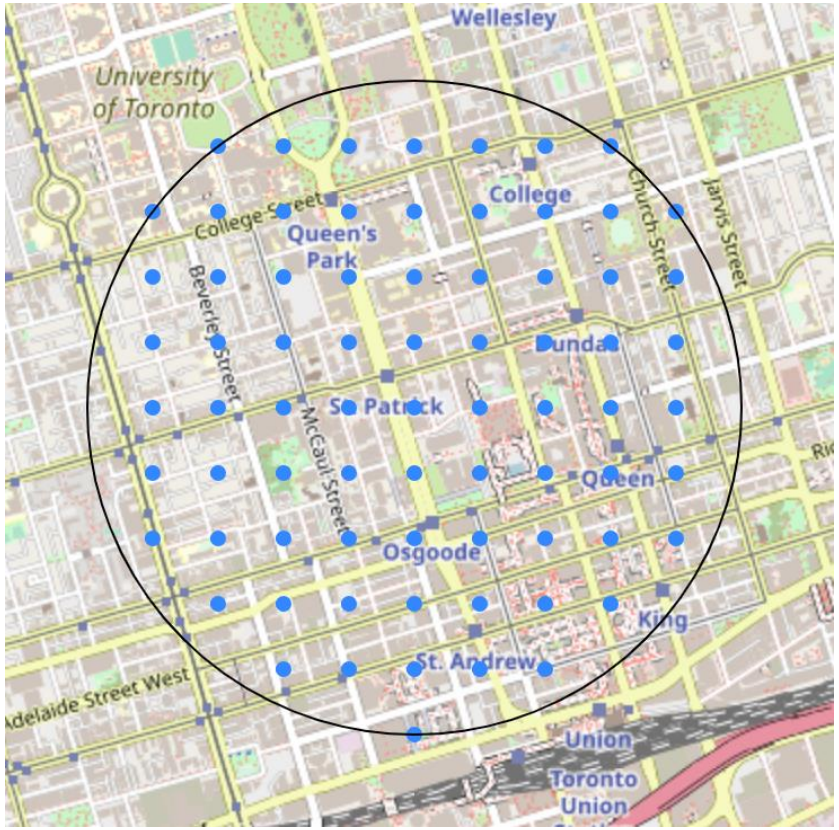
- Location data from the Nominatim Geocoder in GeoPy:
  - For this project I chose Toronto to be the location of interest
- Category data from Foursquare to define the type of venue we wish to open
  - For this project I chose to get recommendations for locations to open a café
- Data on venues was pulled using the Foursquare API search Endpoint including:
  - Venue Name
  - Venue Location (longitude and latitude)
  - Distance of Venue from the input location
  - Venue Category

Methodology: model assumptions

- The idea behind this model is that current venue density indicates where demand is
- We can build a model that predicts current venue density based on the other venues around the point that are not of the type we want to open
- We can then compare the densities predicted by this model to the actual densities we see at each location
- If the predicted density is high compared to the actual density then this may indicate that that location is currently under served for the venue type and so may be a good location to open a new venue

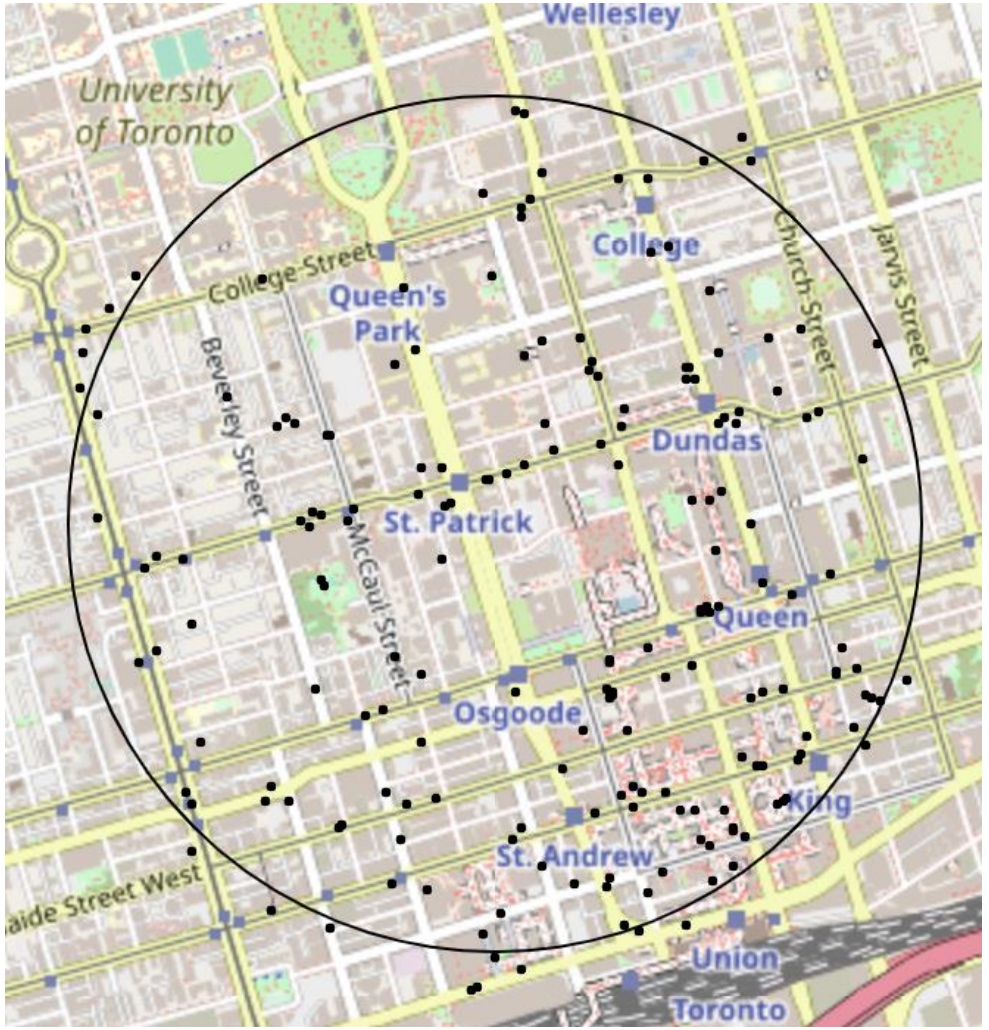
Methodology: defining the search location

- Longitude and Latitude for Toronto were pulled using the Nominatim Geocoder in GeoPy
- I decided to investigate the best locations to open a café within 1km of this location, to within 200m
- I then developed a function grid which would return the set of points in a grid 200m apart within 1km of the centre of Toronto
- These points were plotted in Folium as blue circles, see figure to the right



Methodology: extracting target venues from Foursquare

- I developed the function getNearbyVenues to return venue data from FourSquare
- Using this function I pulled data on cafés within 200m of each point in the grid and their distance from that point in the grid
- These venues were added to a set target venues and plotted using Folium

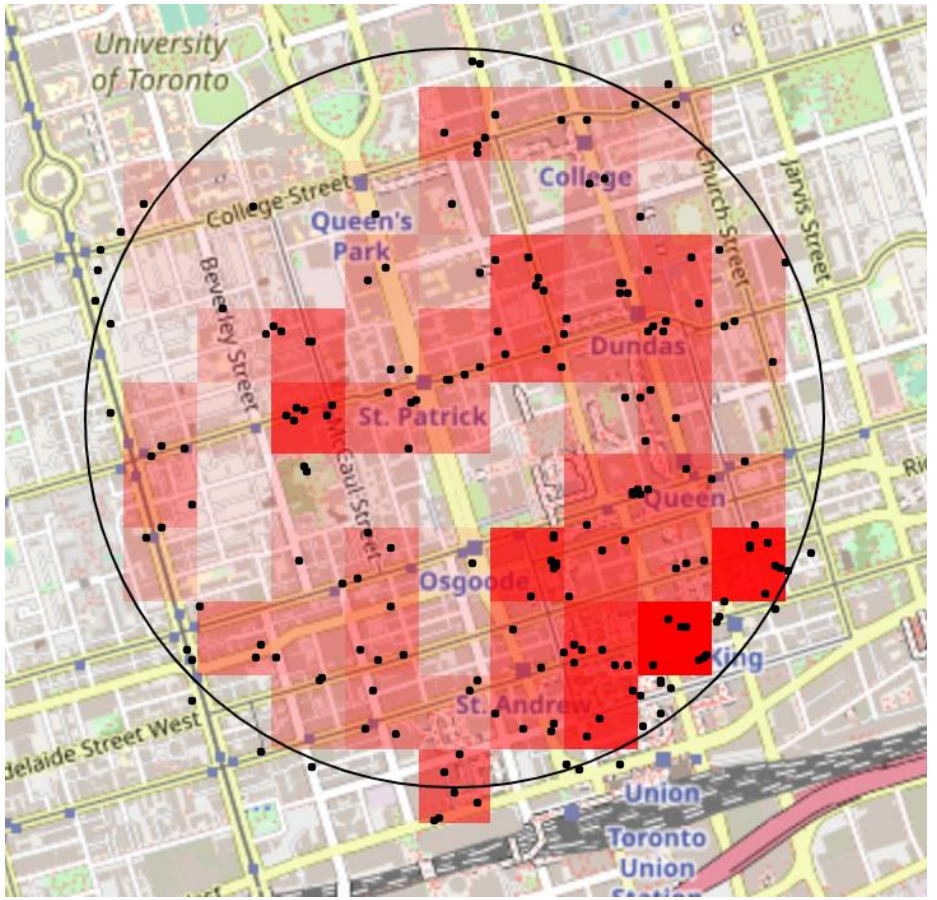


Methodology: Measuring density of target venues

- I developed a function venueDensity which was used to calculate the density of target venues around each point in the grid
- The density D for the point p in the grid was defined as follows:

$$D(p) = \sum_{v} r - d$$

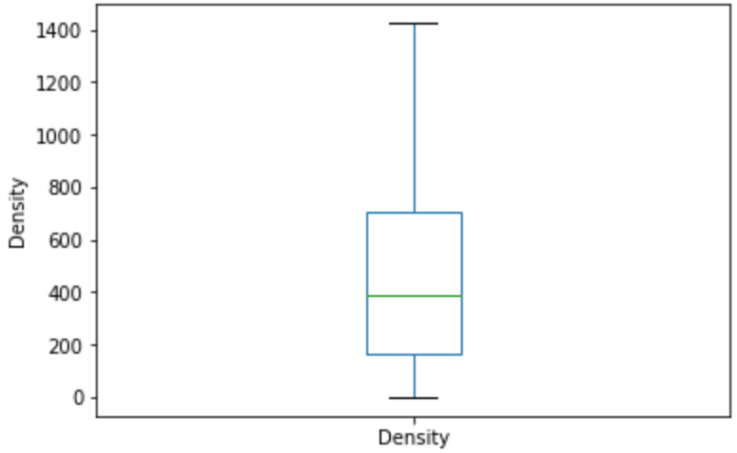
- So if r = 200m a venue that is 0m from the point will add 200 to D, while a venue which is 200m from the point will add 0 to D
- These densities were visualised as a heatmap over the Toronto map using Folium

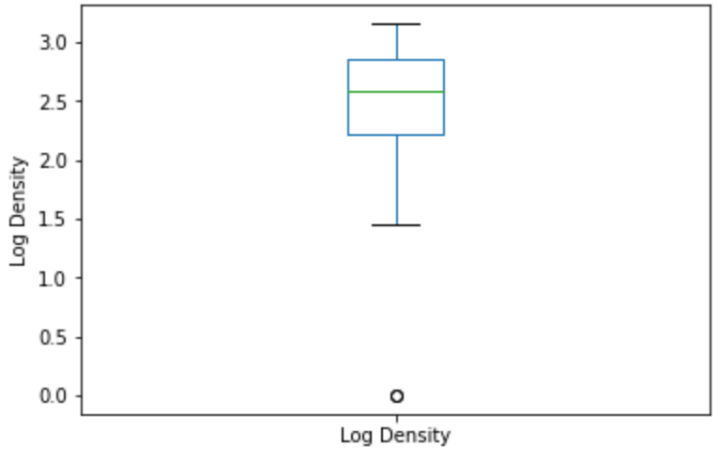


Methodology: statistical properties of our density measure

- Densities were stored in a dataframe and their properties investigated using describe() and boxplots
- Density was skewed so I performed a log transformation on it to make the distribution closer to normal which is optimal for linear regression

	Density	Log Density
count	74.000000	74.000000
mean	457.202703	2.462652
std	331.438251	0.573261
min	0.000000	0.000000
25%	163.750000	2.216648
50%	387.500000	2.589388
75%	704.000000	2.848184
max	1425.000000	3.154120



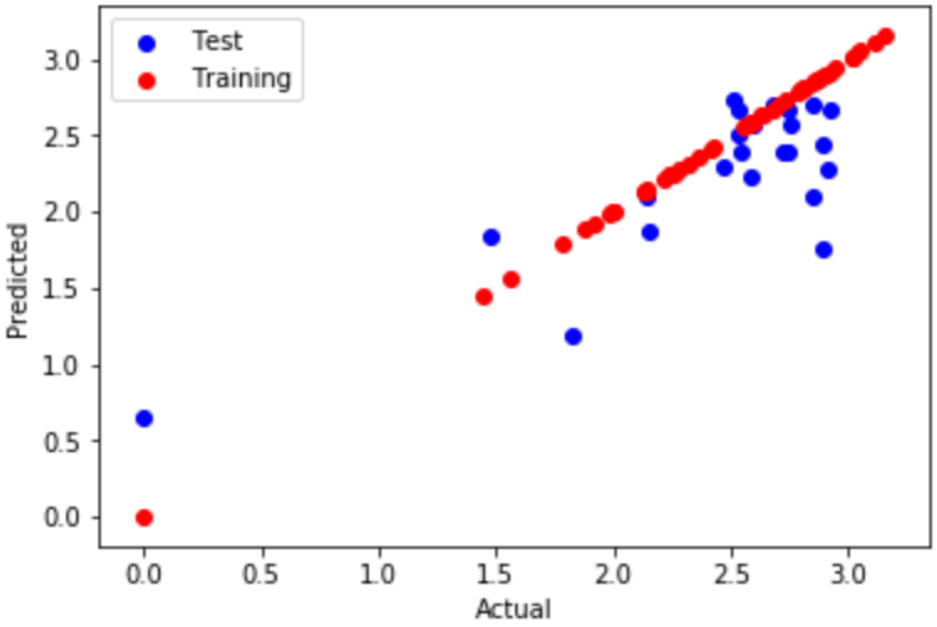


Methodology: Feature engineering

- The features used to predict density of target venues around a given point are the densities of non-target venues around the same point
- The data on these other venues was pulled from Foursquare using the same getNearbyVenues function but without specifying a category ID
- Venues are added to the list of otherVenues if their ID is not in the set of IDs for the target venue and if they have been given a category by Foursquare (not all venues have a category)
- For each venue we calculate the density measure in the same way as we did for the target venues, then sum these up by category for each point using the pandas dataframe pivot_table function
- If a point has no venues in a particular category nearby, the cell for that category is given value 0
- Using the describe() method on the dataframe we see that the densities are skewed so applied a log transformation on them

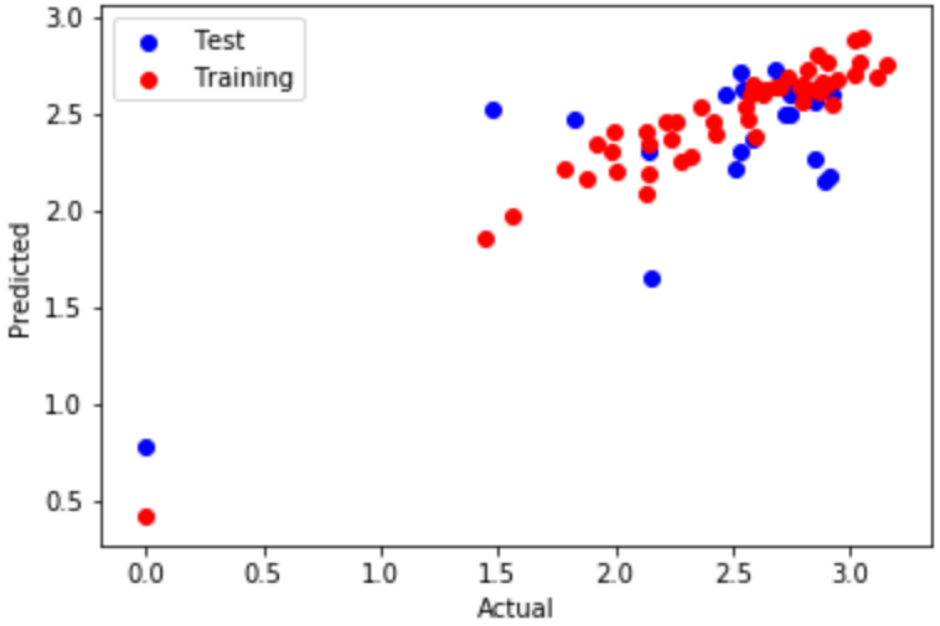
Methodology: model development

- Before training any models we first split the data set into a training dataset and a test dataset
- The training dataset is used to fit the models, the test dataset to evaluate them
- Models were evaluated using the R-squared statistic which provides a measure of goodness of fit
- The first model developed was a simple linear regression on all the features which had a fairly good R-squared of 0.55
- Models were also checked for over-fitting by comparing the predicted results against the actual for both test and training datasets on a scatter plot
- For a perfect model we would see all points being on the diagonal line from bottom-left to top-right where predicted = actual
- For the first model the results from the training data all lie on this line, but for the test data we see much more variance
- This means the model is over-fitted



Methodology: final model

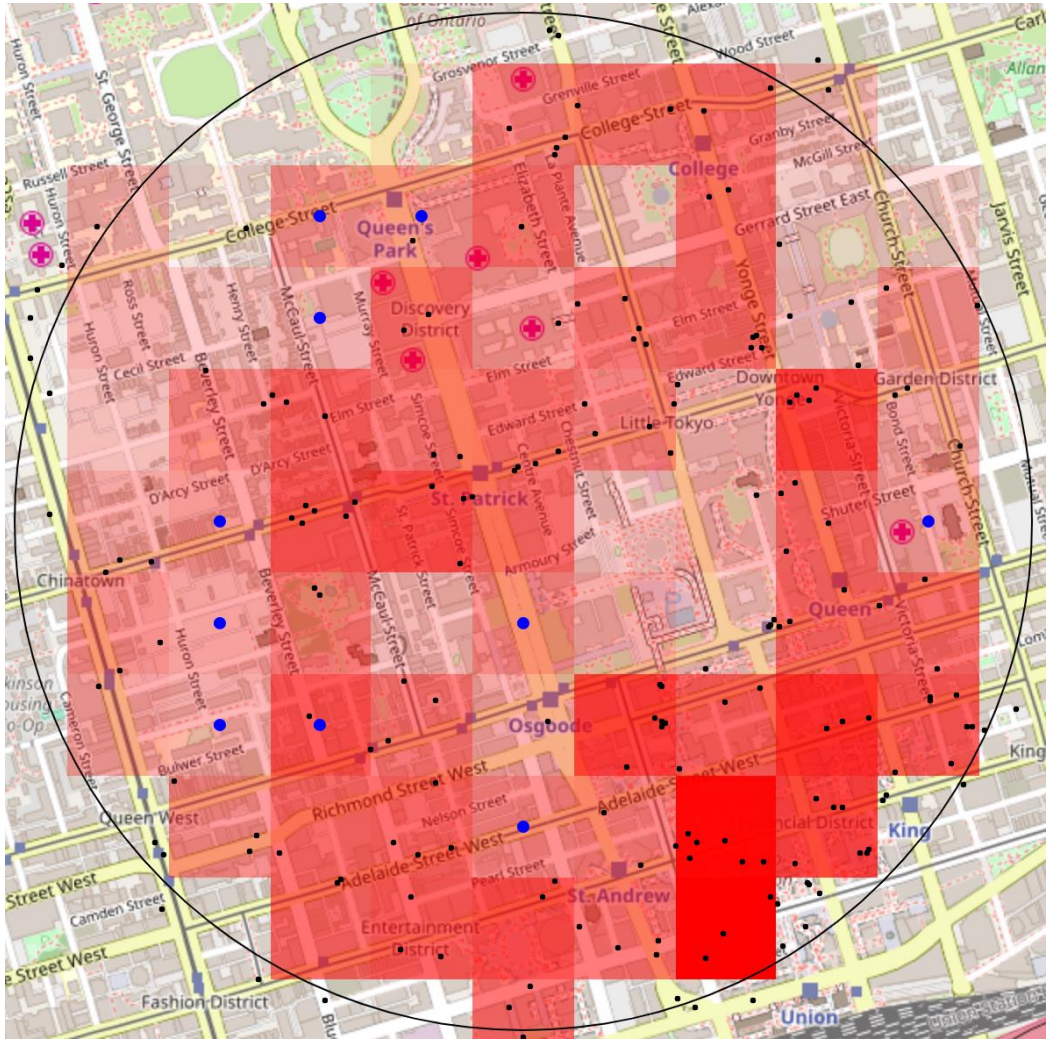
- The final model selected was developed using Lasso regression
- This method was used to improve prediction accuracy by selecting a subset of the model features to use in the final model rather than all of them
- This is done by selecting a parameter alpha > 0: the higher the alpha the fewer features that will be selected in the model
- To decide what value of alpha to select numerous values were tried with progressively finer adjustments (see appendix of the notebook for details)
- The final value for alpha selected was 0.0085
- This model had a slightly reduce R-squared value of 0.51, but was markedly less over-fitted.



Results: venue location recommendations

- To select locations to recommend for venues, the predicted density was compared to the actual density and the top 10 locations with the biggest difference were selected
- The map to the right plots these locations as blue dots, with the heatmap in red showing which areas had the highest predicted density

	Latitude	Longitude	Density	Predicted	Rank
26	43.652163	-79.387207	29.0	333.543761	304.543761
60	43.659363	-79.392165	65.0	297.595192	232.595192
15	43.650363	-79.392165	341.0	516.234263	175.234263
23	43.652163	-79.394645	98.0	257.924652	159.924652
14	43.650363	-79.394645	82.0	223.008651	141.008651
61	43.659363	-79.389686	161.0	293.443853	132.443853
39	43.653963	-79.377290	135.0	257.401395	122.401395
9	43.648563	-79.387207	231.0	344.683099	113.683099
32	43.653963	-79.394645	178.0	289.757553	111.757553
51	43.657563	-79.392165	94.0	204.515079	110.515079



Discussion

- The locations recommended seem to make sense: they are generally areas with low density of the target venues that are next to areas of high density
- Models may benefit from introduction of other data, but the risk of over-fitting is high
- Other modelling techniques and data transformations should be tested to see if they improve predictions and suffer less from over-fitting
- One cause of over-fitting may be that the Foursquare categories are too granular, so building models that group these categories into clusters may be advisable
- Other density functions could also be tested, e.g. just counting the number of each venue type around the point rather than considering distance

Conclusions

- These techniques did come up with plausible venue recommendations
- The hypothesis that we can predict locations to open venues with Foursquare data by comparing predicted density of venues to actual density is hard to prove
- The general methodology was promising, but could be refined further as discussed in the discussion section