A New Pizza Restaurant for Toronto

Introduction

An owner of a chain of pizza restaurants is looking to open a new location in one of Toronto's neighbourhoods. In order to inform this decision, it is intended to explore two key aspects of the current status of restaurants in Toronto using publicly available data and venue information from Foursquare. These aspects are:
1. The density of all types of restaurants in each postcode and
2. For each of the pizza restaurants listed on Foursquare to visualise their location, rating and price tier.

Using Folium to map the density within a postcode and to map the restaurant location, rating and price tier it is intended to visually explore where the opportunities to disrupt the market might be optimum. The approach used will facilitate two models of disruption that ask the following questions:
1. Greenfield - where is their low density of restaurants and an opportunity to site a new business?
2. Disrupt - where is there a set of poorly performing establishments that could be disrupted?

The approach should also indicate where there is good service, and disruption may be difficult. The use of visualising the data will provide understanding about how geographical aspects of the city such as the waterfront and airports impact upon the opportunity.

Data

The key outputs of the work are two maps of Toronto. One shows the density of all restaurants in a 500m radius of each postcode. This uses the postcode page on Wikipedia: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
Foursquare is then used in search mode to return all restaurants within the radius of the postcode. Folium is then used to visualise the density of restaurants.

The data for all restaurants is then selected down to just pizza restaurants. This list is used to return the price tier, rating and likes for all of the pizza places. This data is analysed to see how high rating and highly liked pizza places are clustered. A scatter plot is used to show that the relationship between ratings and likes is highly non-linear. Therefore, a cluster approach is used to sort the pizza places into 10 categories. This allows the visualisation of the distribution of like restaurants.

Methodology

The Wikipedia page is initially scraped for the postcodes and their latitude and longitude. The location of the postcodes is then visualised using folium to show their distribution. Foursquare is used to return up to 100 restaurants within 500m of the postcodes. The maximum returned was 50 so there was adequate representation of the density around each postcode. The density was then visualised as coloured markers in Folium.

A new data frame is then generated that dropped duplicates (as some postcodes are within 500m of each other). This data frame then selects just the pizza restaurants. The data frame is then extended using Foursquare to include the longitude and latitude of each pizza restaurant as well as the price tier, rating and likes.

The result is a detailed frame of 101 pizza restaurants across Toronto. This data is then used to produce a second interactive map using Folium. The markers now represent pizza restaurant locations and are colour coded for rating. The exception is that restaurants in price tier 2 are shown in blue. This interactive map provides a way of exploring where the Greenfield and Disrupt models might work.

A scatter plot of rating against price tier alongside a scatter plot of likes against ratings showed the following features:

- Likes rise rapidly as ratings rise above 6. This indicates a non-linear relationship that is based on general indifference until a certain level of quality is attained.
- Only price tiers 1, 2 and 3 are represented. As price increases ratings increase with much less spread. At lower pricing there is much greater spread in ratings. As price increases so does ratings and the spread decreases.

Since the relationship is non-linear between likes and ratings the pizza places were clustered using a k=10, k-cluster approach to see which restaurants are similar. This was also plotted using coloured markers on a map.

Results

The maps indicated the following:

- The density of restaurants is highest in downtown and along the waterfront. These also tend to be the more expensive restaurants.
- There are concentrations in certain postcodes that possibly relate to public transport or freeway exits/entrances.
- Quality appears to be more important than price. There were many restaurants with no ratings, and it can be seen that below ratings of six there are far fewer likes.
- Away from the waterfront quality and density decreases.
- The clustering was run for different k and it was found that 10 captured the highly non-linear region above rating = 6. It also demonstrates the lack of variation in land.

Discussion

It would appear that both business models could be supported by the data. There are areas of low density of restaurants that may represent opportunities. There are also areas that are low or no ratings. These may require more research.

Downtown and along the waterfront appear very competitive as represented by the non-linearity in the likes vs ratings scatter plot.

The use of clustering to investigate the non-linearity is not in line with using it as a purely machine learning approach. However, the approach could be extended to help predict likes from ratings such that it could form part of a campaign approach during launch.

Conclusion

That data analysis provides support for an approach to exploring the greenfield or disruptive approaches. The missing component is the cost of property. The next step would be to overlay property prices on the data to see how this affects the landscape.

The data analysis acts as a guide which, along with the property prices would indicate where to start in exploring the neighbourhods in more detail.