Optimal Venue Location for Chicago

1. Introduction

1.1 Background

This project is an attempt at determining optimal venue locations for Chicago, by approximating the demand for venues in general with the median income of the neighborhood and by determining existing supply of relevant venues with Foursquare API. Business owners seeking to establish new venues ideally want high demand for the service, and low existing supply.

1.2 Problem

Searching existing literature on the subject, several projects explore venue recommendation from the customer point of view using social media and other methods, and indeed the Foursquare application also includes a recommendation feature. However there is less research on optimizing venue location from the business owner point of view using Foursquare data, traditionally it has been handled by analyzing expensive demographic data [1]. A possible approach is to explore existing association rules between venues and extrapolating that to new venue locations, however venue associations can also arise from co-locality (such as a city or shopping center) without necessarily implying a relationship between the venues. If a customer arrives from a more suburban area, he might visit several venues during his visit simply because they happen to be in the same place, not necessarily because they have anything other than location in common. Compared to market basket analysis, because of distances involved the customer is more restricted in his choices when selecting venues. A more comprehensive data mining approach found chain specific differences between typical customer flow for chains like McDonald's and Starbucks, as well as effects relating to local architecture and city planning [2]. This project focuses on maximizing potential business revenue for venue founders, by identifying high income areas with low existing competition. The assumption is, that rather than relying on typical existing consumption patterns in other localities for venue placement decisions the new venue can realize the potential demand of the area through marketing efforts and by providing a superior service. It is possibly geared more towards the needs of small business owners needing a competitive edge or niche, than to the operations of larger chains.

2. Data Acquisition and Processing

2.1 Data Sources

The primary data sources used in the project were Foursquare venue data, census income data from the DataUSA API as well as city locations given by the Geopy API. Census tract boundaries for Chicago in 2010 were obtained from the Data.gov website. Census tracts were found to provide an appropriate granularity between the county level and census blocks. According to the U.S. Census Bureau, they are

“ Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants.” [3]

2.2 Data Processing

The census tract boundaries were obtained in a JSON file which needed to be preprocessed as the latitudes and longitudes were given in polygon format. For this project, it was considered appropriate to display the data on folium maps using city markers, rather than colored choropleth maps that can accommodate polygons better. The potential user needs to have a clear grasp of terrain features and even street names, which a colored map would obscure. The central coordinates of the polygons were approximated by using the maximum and minimum values and treating them as a squares. Using the squares, the radius parameter to be sent to Foursquare for each census tract was approximated from the latitude and longitude data using the Haversine formula, which is used for approximate sphere calculations in geographical applications.

Because the only interest of the project was competing venue count there was no need to process the Foursquare data extensively, and the data from DataUSA needed simple unpacking. However the call limit of Foursquare was exceeded due to the number of venues fetched. The city center was obtained using a call to the Python Geopy API.

3. Methodology

3.1 Data Exploration

Potential venue owners are enabled to compare the potential demand of an area in Chicago, displayed as a map of average income broken down by census tract, to the level of existing competition for that venue type in the area as determined by venue counts in Foursquare. The object is to visually determine unexploited market niches, where potential demand exists in the form of higher than average income combined less than average existing supply for the venue in question. In order to avoid clutter the visualizations are performed on separate maps.

3.2 Visualization

For the visualization of the data the Folium library was used. Income data was normalized by dividing it with the highest available income area, since relative area differences were considered more important than absolute income levels and can also be more easily visualized. Cutoff values for income mapping were 20%, 40% and 60% of the maximum value, since income is known to follow a Pareto like distribution rather than a normal distribution. The color scheme of the income mapping was also adopted for the competitor mapping in order to facilitate comparisons. By experimentation cutoff values for the competitor count were determined as 0, 3 and 5, however it may be appropriate to alter this depending on venue type.

3.3 Regression Analysis

As a validation of the methodology, a simple linear regression analysis is performed on the data in order to determine whether the venue type being researched has some degree of correlation with income. If there is no correlation of with income a different approach to venue location optimization may be more appropriate for such a venue type. Hypothetically speaking a venue like “Arcade” might have zero or negative correlation with high income areas.

3.4 Clustering

The project also examines the data with a clustering approach, using DBSCAN as the algorithm as the clusters can be irregular. While it identifies some areas on the outskirts of the high competition premium areas correctly as optimal locations, other areas in lower income segments are mislabeled as optimal. A possible remedy might be a larger range of venue counts, increasing the scale of the competition and improving the clustering results, however the call quota of Foursquare imposed limits on the venue counts in this project. It’s not obvious this would have solved the problem since the used competing venue count range provided relatively comprehensive coverage of the locations with the example venue type, only relatively few locations would have exceeded the range.

4. Discussion

Unlike a typical machine learning application, the focus of the project was to find outliers and market niches for business owners to exploit. Therefore, as most statistics and methods such as clustering are focused on generalizing and eliminating outliers, it was not obvious what methodology to use. The clustering approach used provided semi-adequate results, however a different solution might be to apply Deep Learning Anomaly Detection. This method outperforms regular machine learning approaches and is optimized for anomalous outlier detection [4], which are in this project treated as potential market niches. A potential drawback however, is the generally black box nature of such methods.

5. Results

For the income mapping, it was found that suburbs in the north and southwest had higher than average income, along with an area in the northwest adjacent to the central business district of Chicago. These areas would be prime interest to venue owners according to the approach developed in this project, however in the competitor mapping section it was found that venues did not always concentrate in maximum income areas. Reasons for this could include specific ethnic areas, or that some venue types are more attractive to lower income customers. Such associations are however in this project considered less reliable than the aim of maximizing potential business revenue by considering quantitative supply and demand factors. While a given area might have an dominant ethnicity and also a certain venue pattern, it is not necessarily true that another area with the same ethnicity will have identical or even similar venues. Another consideration for a business owner is the longevity of venue concentrations not related to income, as obviously the purchasing power is lower over time. The approach in this project has been, instead of relying on existing associations of any kind, to identify and exploit new market niches. This approach has also anecdotally been used by many big businesses when they were formed.

6. Conclusion

This project has developed an approach for determining the optimal value of a location, by focusing on supply and demand factors with the object of maximizing potential revenue for the business owner. However a more comprehensive approach might be to classify areas as unsuitable for certain venues, such as a high class restaurant in a purely suburban high income fringe area. Such areas however might be devoid of venues in general, so identification of such areas probably needs additional data.

Obviously other factors than the mean income of an area drive demand as well, including Foursquare recommendations and tips, but they are harder to quantify and are perhaps more likely to be short term trends. A static modeling of such demand factors is in this project considered less reliable than the assumption that the venue operator is capable of activating the needed demand in the right environment.

References

[1] Z. Yu, D. Zhang, D. Yang, “Where is the Largest Market: Ranking Areas by Popularity from Location Based Social Networks”, in 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing

[2] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, C. Mascolo, “Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement”, in Proceedings of the 19th ACM International Conference on Knowledge discovery and Data Mining, Chicago, USA, 2013

[3] U.S. Census Bureau Factfinder, https://factfinder.census.gov/help/en/census_tract.htm

[4] R. Chalapathy, S. Chawla, “Deep Learning For Anomaly Detection: A Survey” https://arxiv.org/abs/1901.03407