Ideal Locations for New Restaurants in Brandon, FL

Predicting Foursquare “Likes” ratio based on local demographics

Brandon Snyder - 2019 July

Introduction

Tampa, FL has a reputation of being a city where most citizens live in the surrounding suburbs and commute into the city for work. Because of this, the population isn't as dense as other cities, which can make it difficult for restaurant businesses to pick locations for new venues. I'd like to explore whether the existing venues in a suburb to the east, Brandon, can be modeled based on their local area demographics to determine the best areas for a new venue to achieve a high ratio of "likes" or a high rating on Foursquare.

Being able to predict this information would help hedge against the usual risks that any small business faces when starting up. Sustaining a business with effective management is also important, but in order to even reach that point, a prospective founder needs to choose a location and actually open their venue. Mobile businesses such as food trucks are an alternative to making a solid decision about a specific location, but if that isn't a viable option, then it would be important to be well-informed about the potential customer base near any potential locations - otherwise, a struggling business may have to choose between relocating or closing shop altogether.

Common sense is that most restaurants tend to be more popular with certain demographics than others. Market research may be required to identify what those target demographics actually are, and even if a businessperson thinks they know best, it wouldn't take much time or effort to employ some data science to convert publicly-available information about local demographics and consumer sentiment about comparable venues into actionable insight on areas that are prime for a new venue.

Data

- Foursquare:
    - Venue locations (lat/long)
    - Venue category
    - Venue "likes", "signals" (total of "likes", "OKs", "dislikes"), and rating- only available in Venue Details, at 1 result per call. In order to avoid burning through alloted calls by repeatedly requesting the same data, I'll probably keep a set of finalized JSON files after figuring out the data I want, and load them into Python from storage.
- Census.gov
    - Demographic breakdown: CSV files of "SEX BY AGE" (B01001 and B01001A-I, ACS2017) per Census Tract for all of Florida (the Tampa Bay area includes a few counties so I'm not narrowing this data down too much yet)
        - All races/ethnicities https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001/0400000US12.14000
        - White alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001A/0400000US12.14000
        - Black alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001B/0400000US12.14000
        - Am. Indian alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001C/0400000US12.14000
        - Asian alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001D/0400000US12.1400
        - Hawaiian alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001E/0400000US12.14000
        - Other alone https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001F/0400000US12.14000
        - 2+ races https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001G/0400000US12.14000
        - White alone, non-Hispanic https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001H/0400000US12.14000
        - Hispanic https://factfinder.census.gov/bkmk/table/1.0/en/ACS/17_5YR/B01001I/0400000US12.14000
- Shapefiles for census tracts based on the same year as the demographic data (2017) https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2017&layergroup=Census+Tracts

Methodology

1. Get lat/long for center of Brandon, FL (suburb of Tampa)
2. Get venues within 7.5km radius of center of Brandon, FL (town boundaries identified earlier as roughly 15km x 15km)
    - Filter by category to restaurants (section = "food")
    - Take list of venueIDs and iterate thru to call for venue details of each, specifically seeking category, number of "likes", the rating, and number of “signals”; calculate “Likes Ratio” as “likes” / “signals”
    - Store this data to avoid repeating this call (it'll eat thru the allotment quickly)
    - Group venue categories into broader “cuisines” to simplify later analysis
3. Import shapefiles for census tracts
    - Get lat/long for center of each census tract
    - Find distance of each to center of Brandon, FL
    - Filter to census tracts within a radius slightly wider than venue radius used for Foursquare so tracts partially within venue radius can be included (difference in radii referred to as “radius margin”)
4. Import census tract demographic data from CSVs
    - Import per-race CSVs into individual dataframes; from 4th column on, only need every other column (estimates only, no need for margin of error; column headers can be confusing because they include an extra information, sometimes causing them to be misaligned)
    - Append each racial dataframe into a single master dataframe with an additional column to identify which racial dataframe the rows came from
    - Pivot to reduce data to 1 row per census tract; with age/sex/race and age/sex subpopulation columns
    - Add additional subpopulation columns aggregating by age/race, age, and race
5. Cluster census tracts based on subpopulation proportions
    - Divide every subpopulation count by that tract’s “Total:ALL” value to get proportion
    - Implement KMeans cluster method on proportion dataframe to label each tract with an integer “Cluster Label”, add this column to dataframe
    - Convert “Cluster Label” into a set of one-hot columns
6. Cross-join venue dataframes with census tract dataframes and aggregate demographic data for tracts within “radius_margin” of venue
    - Based on list identified from tracts’ geographic data, join relevant census tracts’ demographic data
    - Calculate distance from each venue to each census tract, then filter out any further than “radius_margin”
    - Pivot the dataframe to aggregate data per census tract as a sum of the number of nearby tracts, per cluster
    - Abandoned idea: calculate proximity of subpopulations as the weighted average value for each demographic subpopulation based on the distances of the census tracts from the venue times the proportion of the subpopulation in that tract vs all the tracts, divided by the number of census tracts ("how far away from the venue is this subpopulation?") *- see Discussion section*
7. Primary dataframe to be analyzed:
    - 1 row per venue; columns for category, lat, long, likes, rating, “likes ratio”, and each cluster
    - Additional dataframes can be generated from column subsets of joined venue and tract dataframes, such as:
        i. Venue cuisine one-hot columns
        ii. Aggregated age bracket proportions for tracts near each venue
        iii. Aggregated racial group proportions for tracts near each venue
8. Review possible dependent variables - price, likes, rating, and “likes ratio” - to determine which is most appropriate and viable to model; generate distribution graphs and descriptive statistics for each; conduct some brief background research before committing to any one variable to ensure it is not surprisingly more complex than it may seem (i.e. Foursquare’s “rating”)

9. Feature selection - examine relationships between possible features and intended dependent variable; the order I followed was based on trying to keep the model as simple as possible, as I didn't want to immediately jump to the full set of subpopulations (dozens of variables)
    - Aggregated subpopulations such as age brackets and racial groups
    - Venue "competition", i.e. nearby venues and any venues in the same category
    - Number of nearby census tracts per cluster
    - Because none were immediately obvious as a strong contender, I decided to try modeling with each set of potential features, with the "cuisine" one-hot variables also included in case that helped bring out a hidden trend
10. Use machine learning algorithms to develop model(s) for predicting the "likes ratio" of a venue based on different feature sets
    - Standardize data; double-check correlations in case they somehow changed due to this
    - Set up Features array and dependent variable array(s), including a classified version of the dependent variable to use with classifier algorithms
    - With each possible set of features, try various machine learning algorithms that support multiple features (multi-linear regression, logistic regression, decision tree, and support vector machine); monitor metrics to choose which model is most accurate
        i. I ended up settling for a classifier that was able to outperform pure chance (>50%) due to time constraints for this project
11. Once a model has been selected, identify "ideal markets" of census tracts where a restaurant could do well
    - Generate new potential venues as rows in a dataframe that matches the model's features; run these "venues" thru the model to predict values for the dependent variable ("Likes Ratio")
    - Analyze relationships between predictions and feature values for these potential venues, split by cuisine, and compare against map of clustered tracts to locate ideal locations

Results

- The census tracts of Brandon, FL can be clustered into 4 groups, referred to by the marker color used on a map of the clustered tracts:
    - Blue: in the western third of the overall area, along Interstate 75 (I75) where there are several commerce centers and apartment complexes; where black and young-adult residents are most concentrated
    - Green: covering most of the central and eastern thirds, mostly south of Florida state road 60 (SR60) where there are many neighborhoods of single-family homes; where white residents are most concentrated (although they are the majority in every cluster) and young-adult residents are least concentrated
    - Red: spread across the same area as Green, but more concentrated north of SR60 or immediately south of it, the original center of town where there's presently a mix of single-family homes and businesses; also forms a barrier between Blue and Green just east of I75; these tracts don't have much of a concentration of any age bracket or racial group
    - Cyan: a small number of tracts that are spread across the southwest portion of the overall area, closest to the commercial areas (including metropolitan Tampa to the west), with an additional tract in the far east where farmland is more common; residents in both sets of Cyan tracts are more likely to be Hispanic than other tracts, but have no distinctive age distribution
- Both support vector machine or logistic regression algorithms were able to model whether a restaurant classifies as a highly-"Liked" venue based on nearby tracts' clusters and venue cuisine. Both models' accuracy were around the equivalent of 5 / 8, after averaging across many shuffled train/test splits; not a whole lot better than a coin flip's 50/50 odds, but still better than an attempted decision tree, which was struggling to stay above that mark. The SVM (63.0 % +/- 7.8 %) was selected over the logistic regression (61.6 % +/- 8.6 %) for having a slightly higher average score, but the margin of error on either keeps them from being separated too much.
- Multi-linear regression models struggled to find a reliable pattern among the many features, usually returning a negative score close to 0 and occasionally slightly above 0, depending on the specific train/test split. When modeling with the dozens of dis-aggregated subpopulation features, this model was returning a massive negative variance score.

Discussion & Recommendations

- Ambitious, ill-fated goals detracted from time that could have been spent on presentation
    - Developing a new geo-informational formula for “concentration factor” was determined to be unnecessary as it was not providing any more meaningful information than a simple Z-score of the subpopulation proportion amongst the broader area’s; instead, each venue’s local population demographics were determined as a simple sum of the census tracts whose centers were close enough to the venue to be within the “marginal radius” of the project
    - Foursquare’s “Rating” score for venues is a far more heavily engineered statistic than it’s name might imply, according to their own data science team on various public statements; unfortunately I did not consider this possibility until after wasting a couple of weeks on trying to model “Rating”, which resulted in a lower quality project overall due to the time constraint.
        - As an aside: given that Foursquare is not as popular with users as it once was, there’s a possibility that newer venues may be missing from the dataset, or that the venues’ Likes were mostly built up several years ago and do not necessarily reflect the opinions of current customers. Given how much work Foursquare has put into developing the Rating score, it seems reasonable that some form of time factor would be applied to the pure Likes as well, like an expiration period or some form of weighting, but I did not see it mentioned.
    - A simple formula for the strength of “competition” for a venue was potentially a viable predictor of success, but I wanted to push myself to use the census tract demographic data that I’ve used on projects in the past
- No demographic subpopulation had a high correlation with the “Likes Ratio” on its own; plotting the aggregated age brackets and racial groups against that statistic demonstrated that there were clearly other factors explaining the variation because no obvious patterns were visible
- Expanding the radius of the venue search might help build a more robust model by enlarging the sample size and introducing more diverse census tracts, but the scope of this project was kept intentionally small and familiar to help the author compare results within contextual knowledge of the target area
- According to predicted Likes Ratios for every feasible combination of nearby census tract cluster, the best locations in Brandon, FL for new venues are roughly the same for every cuisine
    - Multiple nearby blue tracts are a boon, while green and red tracts tend to be associated with lower Likes Ratios for a venue, and the rare, disconnected cyan tracts were generally not much of a factor.
    - The one exception to that general finding is Asian restaurants, which had a unique relationship with the rare cyan tracts as every predicted “Good” venue in this cuisine was near at least one such census tract; Asian restaurants were also far less dependent on blue census tracts than any other cuisine. Combined, these two trends seem to give prospective Asian restaurateurs license to venture out further from I-75, which is where the cyan tracts seemed to pop up more frequently.
        - Ironically, these tracts tended to have the least Asian residents of any cluster, which suggests that the “Asian” restaurants’ cooking is probably not as good as a home-cooked family recipe.
- Geographically, the general recommendation appears to confirm common-sense as the area that best fits that description is the corridor along I-75, which is home to the largest centers of commerce as well as many apartment complexes; the remainder of town to the east is primarily single-family homes in sprawling neighborhoods typical of American suburbia. Assuming that young adults would be more likely to appreciate a restaurant makes sense, as they are in a sweet spot as consumers who are earning an income (vs. minors and retiree age brackets) while not necessarily having to worry about feeding a complete family yet (vs. older adults). I am less confident about a racial “common sense” that could be compared to these findings, but it is also worth considering that the racial distribution may be inseparable from the age distribution, i.e. the areas with more young adults tend to be most diverse due to evolving social dynamics.

Conclusion

I sought to develop a model that would allow me to identify the best location for a new restaurant to set up shop, based on how local demographics predicted how Foursquare users evaluated existing venues. Several challenges arose during this project, from discovering the degree to which Foursquare's own data scientists have been studying and engineering their own evaluation statistics, to the feeble attempted design of a new method for representing the local population around a given point, to the terrible metrics on the initial models attempting to incorporate dozens of subpopulation disaggregations. By revisiting the information that had been gathered and implementing some different techniques to simplify the data, better models were able to be developed, finally arriving on a classifier that is able to predict whether a potential venue is going to be often "Liked" on Foursquare. As a result, a recommended area for new restaurants could be provided, with a caveat for a specific cuisine. However, the general recommendation seems easily replicable just from contextual knowledge of the broader area's economic layout; future analysis would likely start by examining whether economic data could help better explain the variance in Foursquare's Like Ratio, especially with the venues' price ranges as an additional variable. Additional data about local non-restaurants on Foursquare or restaurants who aren't on Foursquare could help as well. Attempting to disentangle the effects of the race and age distribution could also be insightful, if the scope of this project were to be expanded.