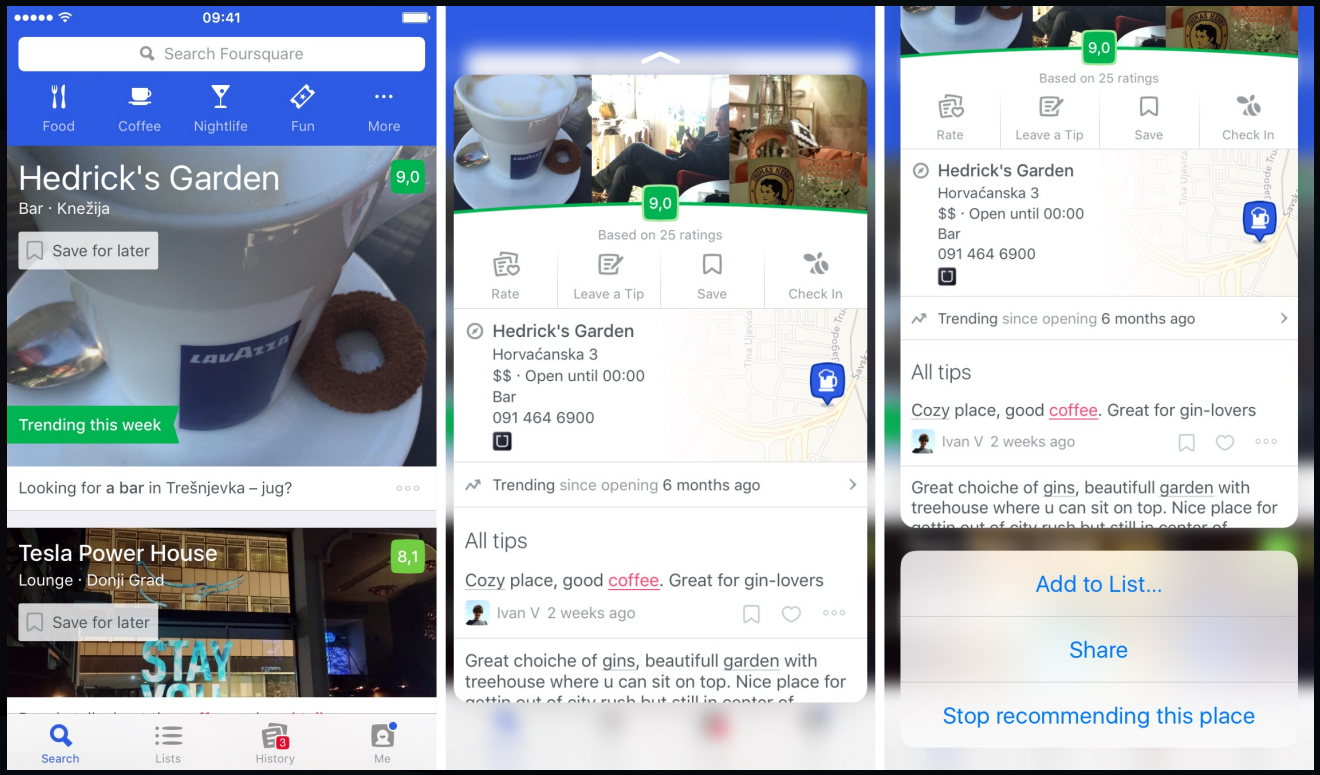


Ideal Locations for New Restaurants in Brandon, FL

Predicting Foursquare “Likes” ratio based on local demographics

Introduction

- Tampa, FL – most residents live in suburbs, difficult for restaurants to pick locations due to sprawling population with inconsistent demographics
- Foursquare's user-driven evaluations help attract even more customers to restaurants
- Predicting locations where a new restaurant is likely to earn positive feedback can help sustain business
- Model based on local demographics may explain enough variation to produce useful insights

Data Sources

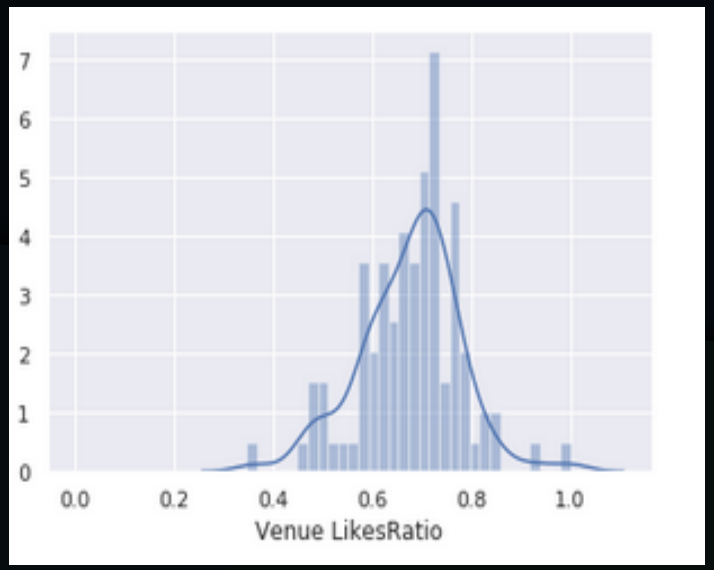
- Foursquare Venues
  - Locations
  - Category
  - Likes
  - Signals
  - Rating
- US Census Bureau
  - Census tracts' estimated counts of subpopulations
  - Shapefiles that outline tracts using coordinates

Methodology

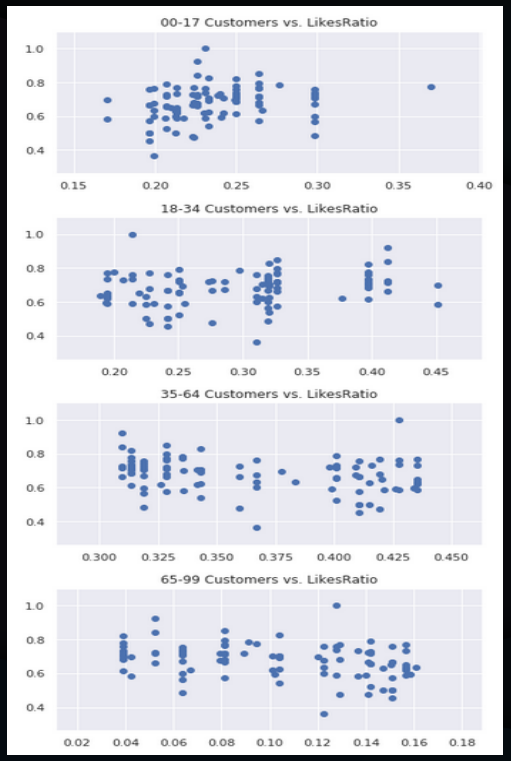
1. Get coordinates for center of suburb (Brandon, FL)
2. Use Foursquare API to get venue JSONs w/in given radius
3. Import shapefiles for census tracts, filter to local area
4. Import census tract demographic CSVs into dataframes
5. Cluster census tracts based on subpop. proportions
6. Cross-join venue data to tract data, aggregate per venue
7. Prep for analysis – main dataframe and subsets

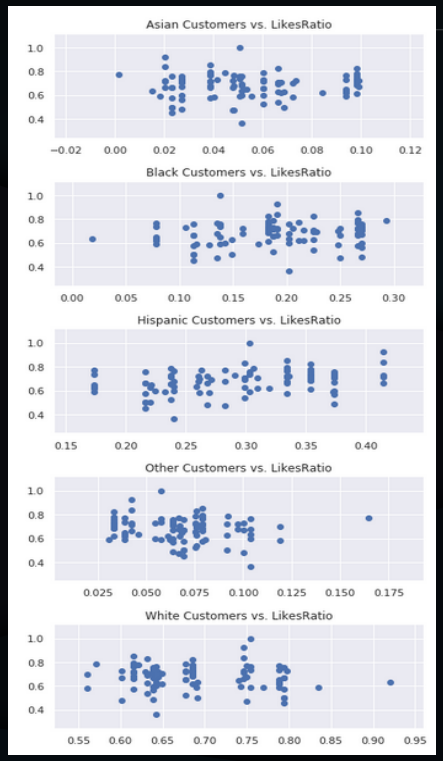
Methodology (cont.)

8. Review possible dependent variables
    - “likes” too biased toward older venues
    - “rating” heavily engineered by Foursquare
    - “likes ratio” most viable

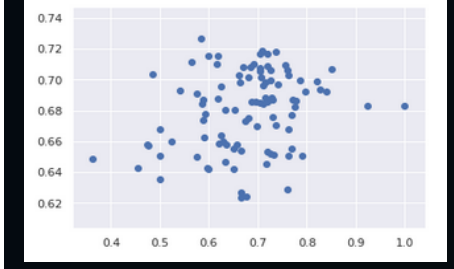


Methodology (cont.)

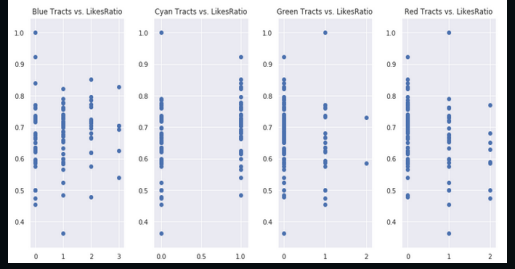




	Venue LikesRatio_x	Venue LikesRatio_y
Venue LikesRatio_x	1.000000	0.252085
Venue LikesRatio_y	0.252085	1.000000



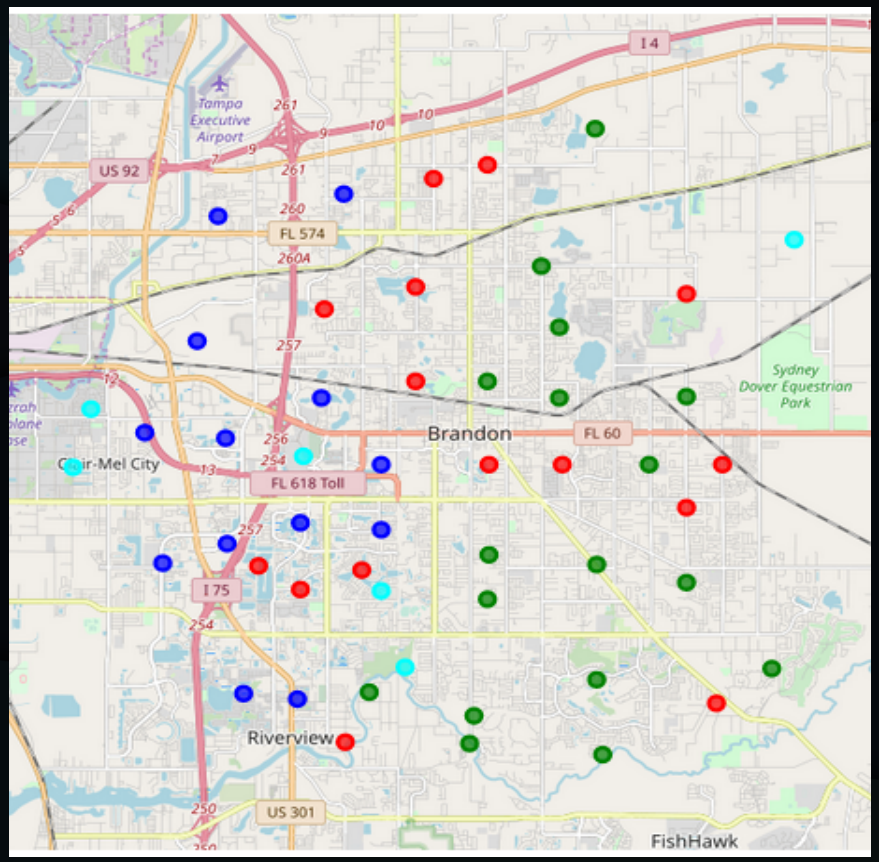
	Venue LikesRatio	blue Cluster	cyan Cluster	green Cluster	red Cluster
Venue LikesRatio	1.000000	0.092494	0.274983	-0.136126	-0.249542
blue Cluster	0.092494	1.000000	0.514373	-0.583997	-0.461265
cyan Cluster	0.274983	0.514373	1.000000	-0.535036	-0.635961
green Cluster	-0.136126	-0.583997	-0.535036	1.000000	0.347968
red Cluster	-0.249542	-0.461265	-0.635961	0.347968	1.000000



9. Select features: start simple, look for obvious patterns, keep options open
   - ✗ aggregated subpopulations
   - ✗ venue “competition”
   - ✔ nearby tracts per cluster
   - ✔ venue “cuisine”
   - ✗ worst-case: all dis-aggregated subpopulations
10. Try multiple machine learning algorithms with each feature set
11. Use best model to identify “ideal markets”

Results: Clustered Tracts

- Blue: I-75 corridor, commercial and apartments; many black and young-adult residents
- Green: further away from Tampa, many single-family homes; many white residents, few young-adults
- Red: along SR60, original Brandon with mix of single-family homes and business; no demographic trends
- Cyan: near Tampa except one to the east near farmland; many Hispanic residents, no age trend



Results: Best Model

- The best: support vector machine
  - 0.628 +/- 0.079 (roughly 5 / 8 correct)
- The good: logistic regression
  - 0.618 +/- 0.088 (still consistently better than chance, >50%)
- The bad: decision tree
  - 0.566 +/- 0.083 (<50% was within margin of error)
- The ugly: Multi-linear regression
  - Couldn’t get any combination of feature sets to reliably produce a positive variance score; using the full set of subpopulations resulted in a massively negative score

Results: Predictions

Cuisine	svmRatio	blue Cluster	cyan Cluster	green Cluster	red Cluster
Cat:American	Bad	1.18	0.41	1.11	1.11
	Good	2.63	0.81	0.63	0.63
Cat:Asian	Bad	1.48	0.49	1.01	1.01
	Good	3.00	1.00	0.00	0.00
Cat:European	Bad	1.18	0.41	1.09	1.13
	Good	2.63	0.81	0.69	0.56
Cat:Fast	Bad	1.19	0.39	1.11	1.15
	Good	2.44	0.83	0.67	0.56
Cat:Joints	Bad	0.77	0.15	1.23	1.31
	Good	1.91	0.70	0.87	0.83
Cat:Latin	Bad	0.82	0.29	1.26	1.26
	Good	2.11	0.68	0.76	0.76
Cat:Other	Bad	1.34	0.45	1.05	1.09
	Good	2.75	0.88	0.63	0.25

General Discussion

- Ambitious, ill-fated goals = wasted time; spend more time during planning phase researching potential variables to improve efficiency
- Individual features were not strongly-correlated with “likes ratio”; clustering was extremely helpful by reducing the complexity of the issue and allowing for more intuitive analysis
- Small-ish sample size due to API rate restrictions probably limited effectiveness of machine-learning algorithms meant to be used on thousands of records

Recommendations

- Recommendation for most cuisines: multiple “blue” tracts nearby, minimal “green” or “red” tracts; “cyan” tracts not much of a factor
- Asian restaurants are unique: must have “cyan” tracts, “blue” tracts are less important
- Both recommendations align with common-sense based on context of local economic geography, i.e. stick to the areas where most other restaurants already are, where the young-adults are rather than families and retirees; any possible racial trend is probably a symptom of the age trend rather than a true predictor, so not much need to worry about matching cuisines’ international sources to customers’ ancestral origins

Conclusion

- Telling the hard truth: this model does not provide much counter-intuitive or nuanced insight
- Positive framing! Plenty of opportunities for improvement and lessons learned about process
  - Start small and simple, expand scope and complexity after building reliable foundation (current model)
  - Introduce additional data to turn implicit contextual knowledge into explicit features:
    - economic data about residents
    - price ranges for restaurants
    - additional businesses not on Foursquare or non-restaurant businesses
    - Separating effects of race and age distributions (different cluster sets?