Locating Fast Food for Success

A Data Driven Investigation

Objective

- Observe Fast Food business in the city of Philadelphia to predict opportunities for potential franchisees

- Derive the attributes of the high traffic areas through statistical and machine analysis

- Minimize the risk and maximize the expected return on investment in opening a new Fast Food Location

Our Input Variables

- A subset of the census variables provided by Safegraph.com (http://safegraph.com/open-census-data). We base our models on the Census Block Group areas defined in this data set.

- Total visitors per CBG also provided by Safegraphs with their census data

- Crime reports as provided by the City of Philadelphia (https://bit.ly/37Bgy8D), which we sum instances per CBG

- Walk and transit score as obtained from Walk Score (https://www.walkscore.com/)

Output Variables

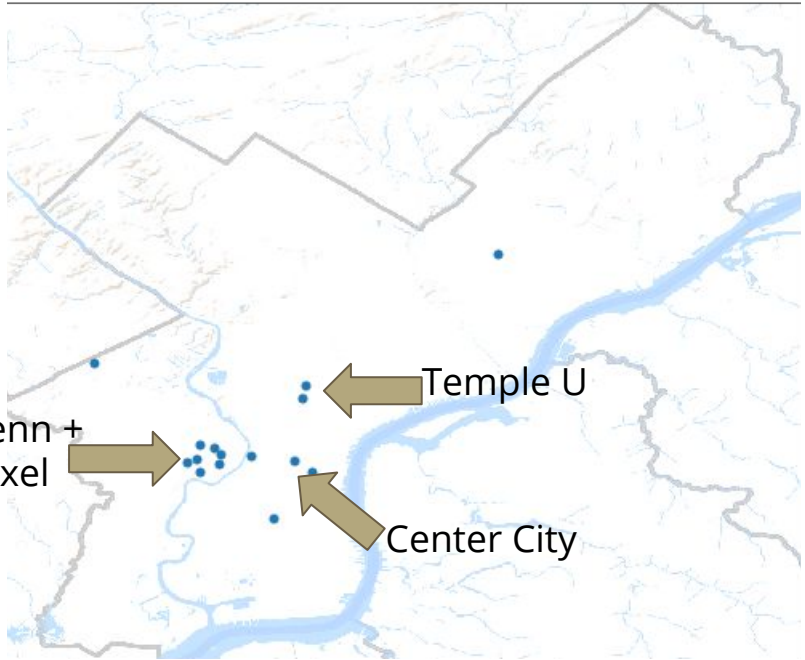
- A sum of fast food locations per CBG, procured by summary of limited service restaurants from Safegraph (https://shop.safegraph.com/)

- Total visitors to limited service restaurants in each CBG, again totaled from data obtained from Safegraph

Data Processing

- Census Data is subsetted to just Philadelphia Data
  - Philadelphia CBGs pulled from Safegraph CBG geometry file with a streaming JSON reader
  - CBGs with data gaps are filtered out, along with CBGs labeled Philadelphia but not matching the codes for Philadelphia
- Census Geocoding service used to get CBG for lat/long of each reported crime, as well as CBG of for address of each restaurant
- Google map api used to get addresses for the lat/long of each CBG, then used to do Walk Score and Transit Score lookup
- Data combined via Pandas and output to CSV

Successful CBGs



Variable Correlation with Total Consumers
Positive:

Undergrads and Grads	0.546434
raw_visit_count	0.546264
Crime Counts	0.376914
Transit_score	0.343890
Pop. that Works from Home	0.254979
Total Population	0.245195

Negative

Married Parent Households	-0.176664
Owner Occupied Households	-0.188451
Single Parent Households	-0.189105

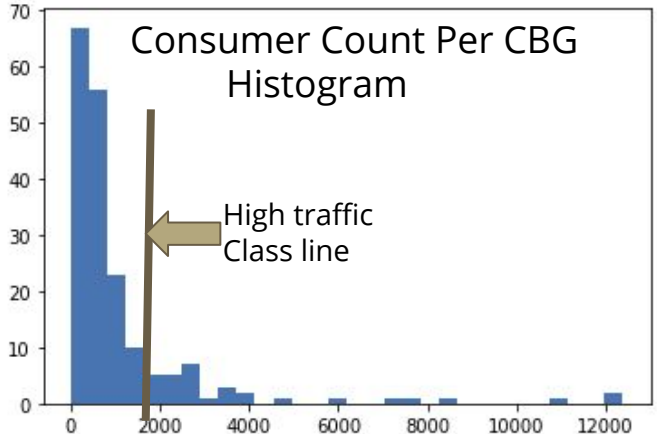
Top Variables for Market Share

The Percentage of all fast food consumers in the top 20% of CBGs for each variable

Supervised Learning

- Experimenting with Regression models to predict total consumers to a CBG, our best model is produced by sklearn's random forest and emphasizes:
  - 'Total Population', 'Undergrads and Grads', 'Degree attained', 'Households in Poverty', 'Below 40k Earners', 'Crime Counts', 'walk_score', and 'raw_visit_count'
  - With R^2 scores 0.89 on the training data and .61 on our test data

- We also use classification models to predict 'high traffic' areas.
  - Our best overall classifier is sklearn's Multilayer Perceptron Classifier:



	Training	Testing
Recall	1.0	1.0
Precision	1.0	0.67
AUC	1.0	0.86

Customer Segmentation

- We try a few unsupervised learning techniques, and are best able to isolate a cluster of successful CBGs by performing KMeans on a 6 component PCA transformation of a subset of variables identified by KNearestNeighbors:

col_0	0	1	2	3	4	5
Successful Location						
False	16	14	64	29	0	1
True	4	2	9	2	2	6

Cluster 4

High	Low
Single Occupant Households, Degree attained population, Households in Poverty, 40-100k Earners, Renter Occupied Households, Crime Counts, Walk Score, transit_score, raw_visit_count	Single Parent Households Married Parent Households

Cluster 5

High	Low
Population that Works from Home Undergrads and Grads	Single Occupant Households, Single Parent Households, Married Parent Households, Diploma/GED as Highest Educational Attainment, Households in Poverty, Income Earners of any Tier, Owner Occupied Households

Reference

A more extensive report and python notebooks are located in https://github.com/jon-e-pizza/Springboard/tree/master/CapstoneFastFoodEstablishments