# Locating Fast Food for Success

A Data Driven Investigation

# Objective/Stakeholders

- 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

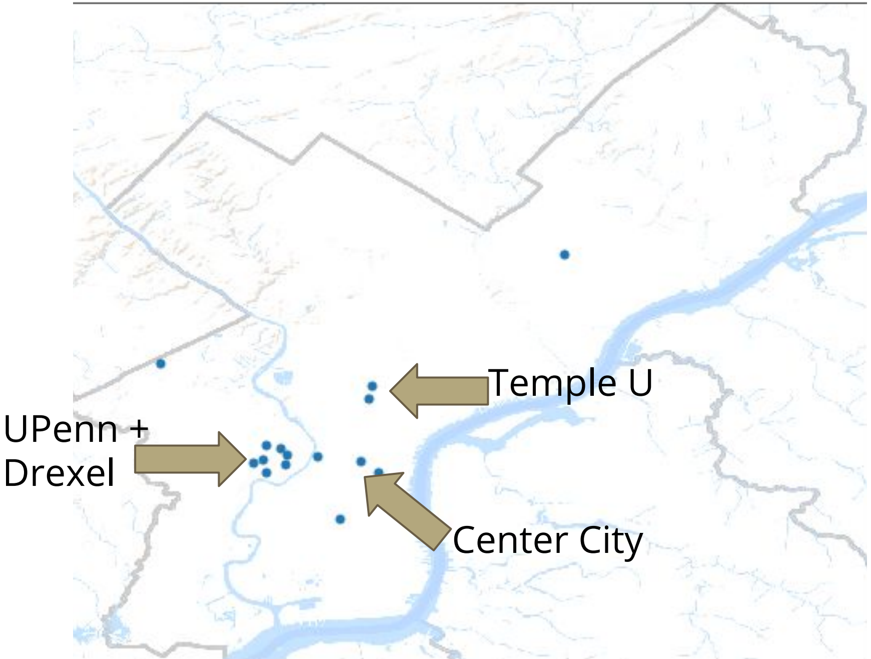
- Census Data is subsetted to just Philadelphia Data
  - Philadelphia CBGs pulled from Safegraph CBG geometry file with a streaming JSON reader
  - CBGs with data gaps are filtered out, along with CBGs labeled Philadelphia but not matching the codes for Philadelphia

- 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

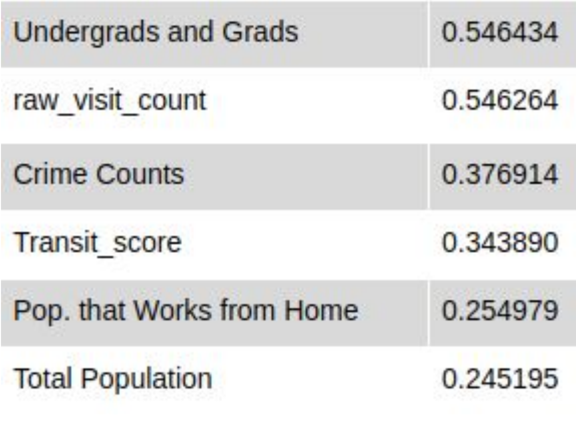


Temple U

UPenn + Drexel

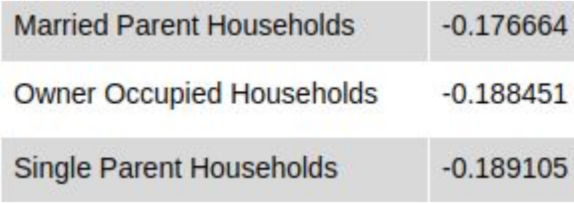
Center City

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 |

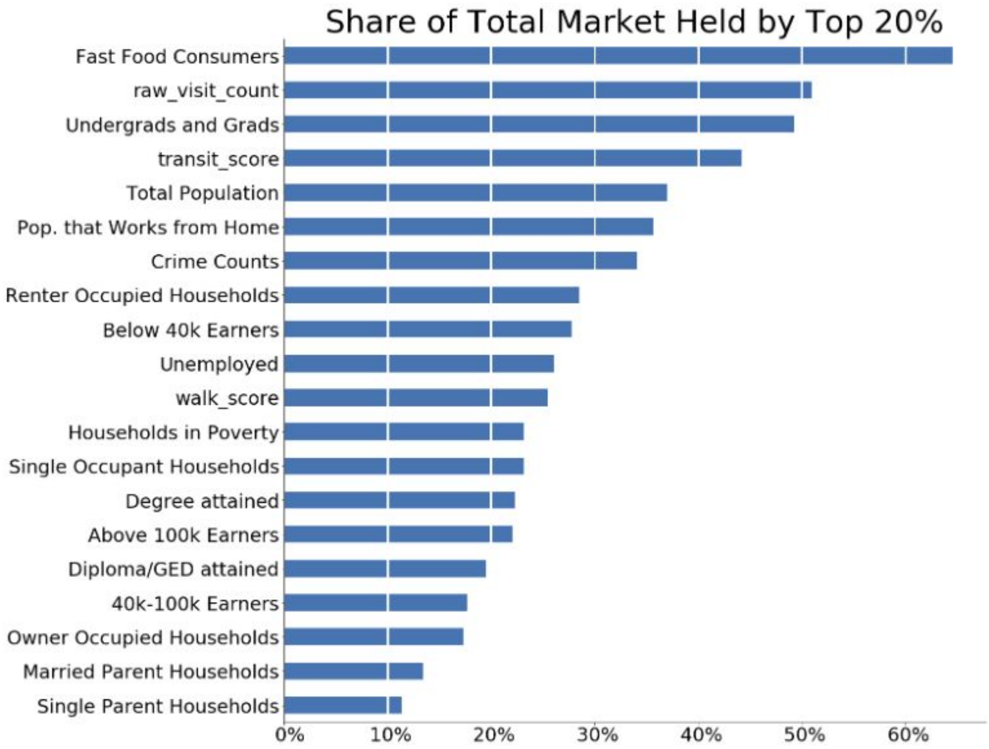
Negative



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

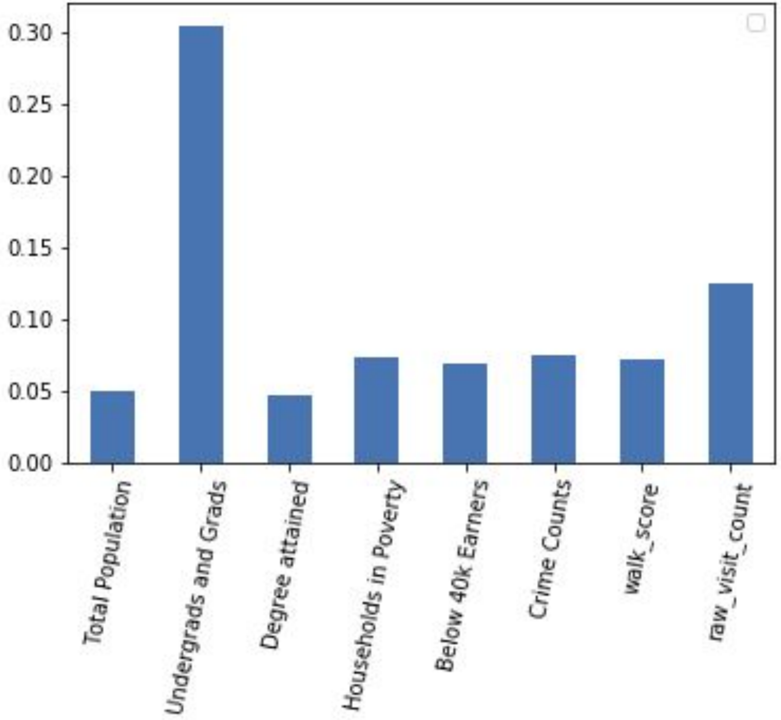
# Top Variables for Market Share

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



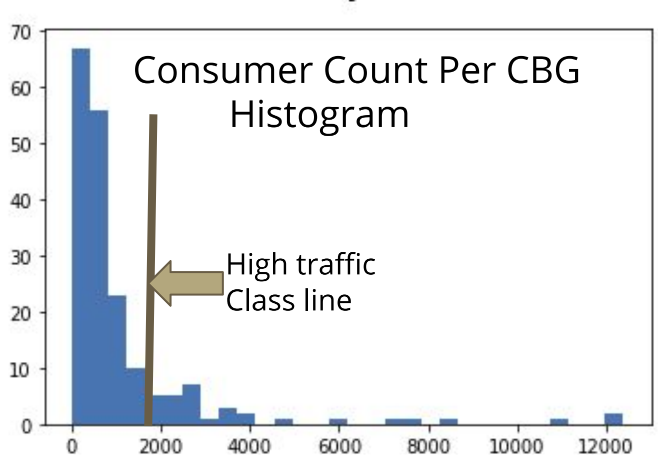
# Regression

- Experimenting with Regression models to predict total consumers to a CBG, our best model is produced by sklearn’s random forest
    - Top 8 variables account for 81.5% of our model’s decision: '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



# Classification

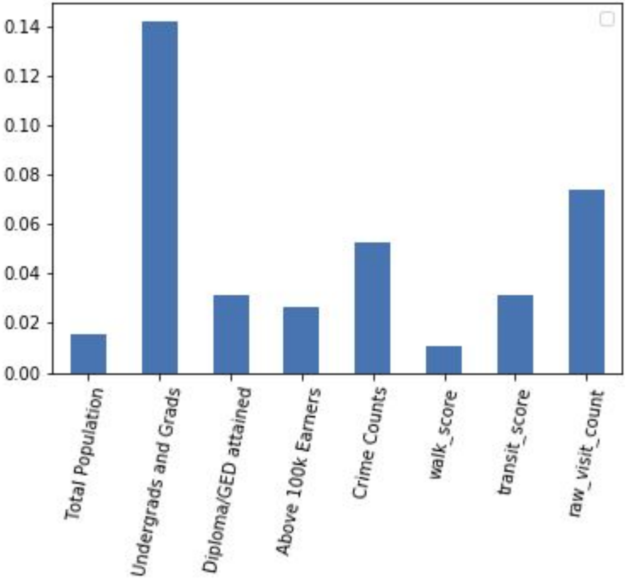
Using classification models to predict ‘high traffic’ areas.


Consumer Count Per CBG Histogram

High traffic Class line

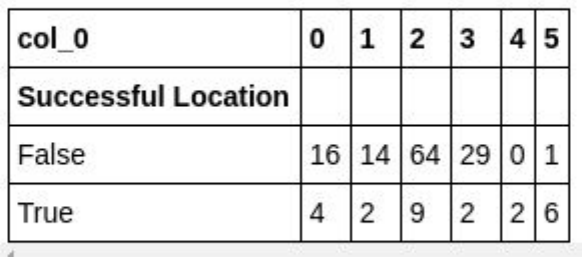
Our best overall classifier is sklearn’s Multilayer Perceptron Classifier with Lasso emphasized variables:

| | 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<br>Married Parent Households |

Cluster 5

| High | Low |
|---|---|
| Population that Works from Home<br>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 |

# Takeaways/Next Steps

- Target areas with universities and downtown areas with high single populations for new restaurant locations

- Gather data from more cities so a targeted regressor can be trained on just restaurant counts in successful areas. A regressor trained on our full data set for restaurant count was too poor for consideration.

- With more data, train a multi classifier for each of three classes and develop separate models for success in each
  - University
  - downtown
  - other

# Reference

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

# Appendix: Linear Regression

Data Sets have gone through SKLearn Standard Scaler, variables picked by function, then ran through LR

Best Ridge:

| alpha | Training $R^2$ | Testing $R^2$ |
|---|---|---|
| 100 | 0.53 | 0.55 |

| Significant Variables | Coefficients |
|---|---|
| raw_visit_count | 520.81 |
| Undergrads and Grads | 348.16 |
| Transit Score | 173.11 |
| Crime Counts | 166.14 |
| Total Population | 123.19 |
| Single Parent Households | -111.95 |

Best Lasso:

| alpha | Training $R^2$ | Testing $R^2$ |
|---|---|---|
| 10 | 0.54 | 0.57 |

| Significant Variables | Coefficients |
|---|---|
| raw_visit_count | 1163.82 |
| Undergrads and Grads | 735.43 |
| Transit Score | 315.52 |
| Diploma/GED attained | 96.41 |
| Walk Score | -125.30 |
| Above 100k Earners | -131.18 |
| Total Population | -175.79 |
| Crime Counts | -205.48 |

# Appendix Gradient Boosting Regressor

| n_estimators | max_depth | learning_rate | Training $R^2$ | Testing $R^2$ |
|---|---|---|---|---|
| 100 | 5 | 0.01 | 0.82 | 0.59 |

| Significant Variables | Feature Importance |
|---|---|
| Undergrads and Grads | 0.47 |
| walk_score | 0.17 |
| Degree Attained | 0.10 |
| Crime Counts | 0.07 |
| raw_visit_count | 0.04 |
| Diploma/GED attained | 0.04 |

# Appendix Classifiers

| | Logistic Regression | SVM | Random Forest | Gradient Boosting | KNeighbors |
|---|---|---|---|---|---|
| Best Variable Subset | 'Total Population', 'Undergrads and Grads', 'Diploma/GED attained', 'Above 100k Earners', 'Crime Counts', 'walk_score', 'transit_score', 'raw_visit_count' | 'Total Population', 'Undergrads and Grads', 'Diploma/GED attained', 'Above 100k Earners', 'Crime Counts', 'walk_score', 'transit_score', 'raw_visit_count' | Total Population', 'Undergrads and Grads', 'Diploma/GED attained', 'Above 100k Earners', 'Crime Counts', 'walk_score', 'transit_score', 'raw_visit_count' | Undergrads and Grads<br>raw_visit_count<br>Crime Counts<br>walk_score<br>Single Parent Households<br>Married Parent Households | 'Married Parent Households', 'Single Parent Households', 'Undergrads and Grads', 'Crime Counts', 'walk_score', 'raw_visit_count', 'Households in Poverty', 'Below 40k Earners' |
| Training Recall | 0.32 | 0.32 | 1.0 | 1.0 | 0.6 |
| Training Precision | 0.8 | 1.0 | 1.0 | 1.0 | 0.94 |
| Training AUC | 0.85 | 0.93 | 1.0 | 1.0 | 0.96 |
| Test Recall | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Test Precision | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Test AUC | 0.71 | 0.87 | 0.92 | 0.91 | 0.70 |