Picking the Optimal Location for Fast Food

It takes a substantial investment to open a new fast food restaurant and a franchisee to be only gets one chance to pick their location. A restaurant managed in exactly the same way could experience a large bump in patronage and profit just by placing in an area with more demand. Therefore, any person opening a new restaurant would do well to know those areas before they start to put down any groundwork. This investigation aims to provide a model for the potential entrepreneur to identify those areas.

We will only be able to analyze a small subset of the data that could possibly be gathered for any given area, and the analysis of this data will be further kept to a feasible scope for this study by constraining our investigation specifically to the city of Philadelphia. It could be of future interest to analyze demand in other cities in the future and see how the effects of our observed variables might differ for other cities.

Python notebooks and consolidated data corresponding to this report at https://github.com/jon-e-pizza/Springboard/tree/master/CapstoneFastFoodEstablishments

Gathering Model Variables

To measure the potential success of different areas for our fast food restaurant, we will have to rely on observations of current consumers. This means we will need to identify areas currently in Philadelphia where there is a high volume of fast food consumption, and then find the variables that best correlate to those areas.

The set of area variables that we choose to pick from will mainly be procured from the extensive measurements gathered in the US Census' American Community Survey. These data are offered in a convenient set of CSV files by Safegraph free of charge at https://www.safegraph.com/open-census-data. The units of area of the US Census are Census Block Groups (CBGs), and so we will be mapping our fast food restaurants to their corresponding CBGs to line up with our measurement variables.

In addition to census variables, we will also make use of Safegraph's recorded visitor count to each CBG, which they also offer free of charge at https://docs.safegraph.com/docs/neighborhood-patterns. Then from outside Safegraphs offerings we will also consult the City of Philadelphia's metadata on crime reports (http://metadata.phila.gov/#home/datasetdetails/5543868920583086178c4f8e/representationdetails/570e7621c03327dc14f4b68d/) to put together a crime count variable for each CBG. Finally we consult Walk Score (https://www.walkscore.com/) for numerical transit and walkability scores to add to our list of predictor variables.

For an output variable we purchase information about the locations of fast food restaurants in Philadelphia and the patronage experienced by each from Safegraph's data shop:

https://shop.safegraph.com. To develop our model we will need to tease out the features of the areas that contain the top performers in that list.

Census Data Choices

We use the latest US Census American Community Service from 2016 understanding that, while it will not perfectly reflect the current state of Philadelphia, it is likely to provide us the best possible picture of differences between different city areas.

For our investigation we need to extract observations specific to Philadelphia from the full set of census data. And then, as we would be unable to conduct an investigation of every variable in that survey, we will also need to make some arbitrary up front choices about what subset we do choose to investigate.

In our downloaded safegraph census data folder, there is a data folder containing various csv tables with one census category per table. The tables filenames correspond to their census table ids that can be seen at https://www.census.gov/programs-surveys/acs/guidance/which-data-tool/table-ids-explained.html (ex cbg_b00.csv). Each table is keyed by CBG, so we will need to subset each table to the set of CBGs that are located in Philadelphia.

To filter out this set of CBGs, the Safegraph census data provides the file geometry/cbggeom.json identifying the location of each CBG. To obtain just the Philadelphia CBGs from this file, the file size requires using a streaming json file reader and saving just the ids of those CBG objects in the file. Having produced and saved this list, Pandas makes it fairly easy to trim down our other data tables just to the entries matching these ids.

A breakdown of the content in each tables by column ids is available in the Safegraph metadata/cbg_field_descriptions.csv table. We see that there is highly detailed data available, so we will use these field descriptions to help us sum various fields to arrive at high level observations.

Among the information available, we amalgamate the following variables:

Census Filename	**File Data**	**Variable Obtained**
cbg_b00.csv	Overall Population	Total Population
cbg_b08.csv	Commuting	Total Working at home in CBG
cbg_b11.csv	Household Types	Count of Single occupant households
cbg_b11.csv	Household Types	Count of single parent households (at least one minor present)

cbg_b11.csv	Household Types	Count of two parent households (at least one minor present)
cbg_b14.csv	School Enrollment	Count of enrolled undergraduate and graduate students
cbg_b15.csv	Education Attainment	Count of residents with attained high school Diploma or GED
cbg_b15.csv	Education Attainment	Count of residents with Bachelors, masters, or doctorate
cbg_c17.csv	Poverty status	Total Households below poverty line
cbg_b19.csv	Income	Count making 40k and below
cbg_b19.csv	Income	Count making 40k-100k
cbg_b19.csv	Income	Count making 100k and above
cbg_b23.csv	Employment Status	Total employed count
cbg_b25.csv	Housing Characteristics	Total renter occupied households
cbg_b25.csv	Housing Characteristics	Total owner occupied households

Some of these fields are read directly from the census files, and some are summed from various fields. Our final output will combine all these variables into one python Pandas dataframe indexed by CBG.

After this gathering step we find a small subset of our CBGs with entries of 0 in most of our variable columns and choose to remove these rows from our final investigation.

Additional Predictors

After collecting our census information for Philadelphia into one table by CBG, we have a few more data entries we would like to add to each row.

First we add the traffic each CBG experiences. We join the Safegraph data/cbg_patterns.csv total raw_visitor_count column to our census table on the CBG index.

Next we would like to procure a crime measure for each CBG. We turn to the city of Philadelphia for a total crime count metric to add to our table of predictors. The CSV of crime occurrences that we use has data from 2006 onwards. To have the best picture of the current state of Philadelphia, we will limit our crime counts to instances recorded in 2018. In order to total how many crimes occur in each CBG we will have to map each recorded Philadelphia crime instance to a CBG first. Philadelphia provides a lat/long location for every crime instance, so with this we can call the US Census geocoding service on each location and collect the CBG for each crime.

After the completion of this step, we add a total count of instances in each CBG to our table of variables. In a future analysis we may aim to further break down crime by type.

Finally we turn to Walk Score for their walkscore and transit score for each CBG. Walk Score requires an address and lat/long to perform a look up, so we act on the assumption that the central point of each CBG will have a walk score and transit score representative of the CBG as a whole. We get the central lat/long of each CBG to call Walk Score with from the Safegraph geometry/cbggeom.json file and we use Google's map API to get addresses for each CBG's central point. With these two pieces of information we collect the scores for each CBG and add them to our predictor table.

Output Variable

Our output variable that we will build our model around is the current success of fast food locations in Philadelphia. Using the Safegraph data shop we purchase data on each 'limited service restaurant' they have recorded in Philadelphia. As they do not record the CBG of each restaurant we first turn again to the US Census geocoding service to obtain the CBG of each location. Then to each CBG we add the total count of restaurants recorded in that CBG, and the total of visitors experienced by all of the restaurants in each CBGs. We are now ready to do investigation.

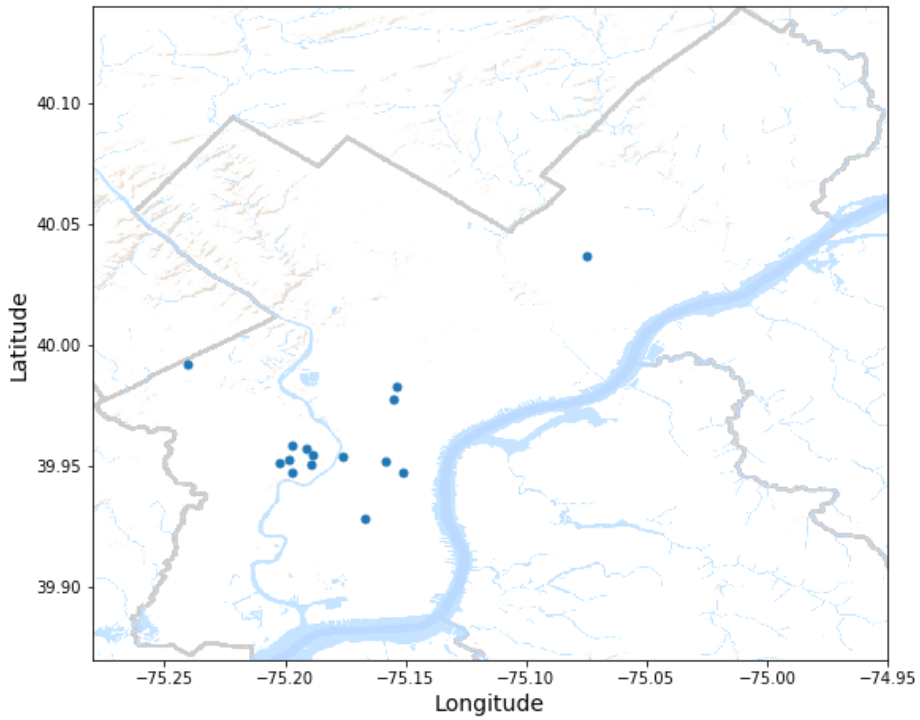
Data Analysis

In analyzing what areas are most successful for fast food establishments, we will not be able to derive any trend from areas that have no recorded fast food restaurants. We therefore start by filtering out the observations of any CBG that does not have any recorded establishment. This leaves 187 CBGs inside of Philadelphia. This is not a particularly large data set, and were we able to do this analysis again without the cost of procuring visitorship numbers being a limiting factor, we would endeavor to map CBGs of additional cities for a larger data set. Depending how many areas we could then analyze, it might also be interesting to do an analysis with CBGs grouped into relatively equal areas.

Among our remaining Philadelphia CBGs though, we would like to get an idea of which ones experience the highest consumership and what can be said about them. One of the first things we might want to know is where in Philadelphia are hot spots for fast food consumption. First we should identify isolate and examine high traffic CBGs.

We have to draw some arbitrary boundary to define what areas are the successful areas. For a first look, we can define any CBG with average visitorship to establishments at the 75th percentile mark or higher as a successful area. We find that this mark is 667 customers a month. As our ending interest is also where we might place a new restaurant, we narrow this set a little further by making sure if we remove 667 customers off the top of the visitor count in each CBG to represent a potential new entrant, that the average business still experiences at least that amount, meaning there is a good chance such an area can support another establishment.

Overlaying those CBGs on a map of Philadelphia, we get:



Which provides the immediate insight, to those familiar with Philadelphia, that university students are a primary consumer group to target for fast food as we see a dense cluster of successful CBGs in the ‘University City’ neighborhood. We also spot two successful CBGs North of downtown near Temple University that would further support this initial insight.

There are then 3 dots along the downtown corridor going east from University City indicating that raw visitorship is also likely to indicate a good spot to place fast food restaurants.

We look to our observed data to back up our initial assumptions by calculating the correlation of each variable to total count of fast food consumers in that area. Indeed we find that, among our gathered predictors, the highest correlated variables to total fast food consumer in a CBG are:

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

With University students leading the way in correlation to success of fast food establishments. Our other metrics seem to correspond directly to potential audience size, but with high volume of visitors seeming to be more relevant than a sitting high population as raw_visit_count and Transit_score both outweigh ‘Pop. that Works from Home’ and ‘Total Population’. We do also see that high crime counts seem to correlate with high fast food consumption, which might not be exactly what we expected as it seems a risk to the average business, but it may be that crime is correlated with other community factors that also correlate to high utilization of fast food restaurants. To tease out that exact connection is beyond the scope of our current study, so we will use this as a positive predictor as is, but might seek to tease out the underlying factors in a future study as replacement variables for this proxy.

We do also surface some negative correlations that might be worth keeping in mind to identify communities to avoid placing in. Namely, we see the negative correlations:

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

Showing family communities and communities with more established residents make less use of fast food establishments.

A single linear correlation value cannot necessarily capture how important each factor is though. While some community metrics may generally be associated with higher patronage, not all communities that are high in that metric are necessarily highly trafficked and this lack of a strict relation may not map well to strict correlation values.

To try to capture this in another way that is less sensitive to strict linear relations, we can group all communities that are high in a metric and see if as a group they capture more than their expected share of the market. We must again arbitrarily draw the line of what ‘high in a metric’ means. This time we choose the 80th percentile and above in each metric, and then map to what percentage of all fast food consumers in Philadelphia are accounted for by those top groupings of each factor. In a truly random world with a uniformly random split, these top 20% groups would account for approximately 20% of the total market, so we are interested in learning what kinds of communities experience significantly more (or less) than their business share. The chart below displays these measurements by total market share:

The first thing we see is that the top 20% of CBGs for total fast food consumer count account for almost two-thirds of the total market (more than 3 times their expected random share). A new business owner would therefore do well to be as informed as possible about the characteristics of one of these highly desirable top scoring communities. Some of our other metrics do surface as similarly over represented in their share of the market, though it does seem that they were captured by our correlation values. Again we find that the areas that should capture our eye the most when placing a new establishment are the highly visited areas (those with high visitor count and good transit) and those with high quantities of University students, all capturing more than double their expected market share. At the other end, we again see that family communities are unlikely to be high utilizers of fast food restaurants and should probably be passed over in the placement of a new restaurant.

Machine Learning

Having conducted an initial examination of our Census Block Groups (CBGs) in Philadelphia with a mind for identifying areas with high fast food consumer traffic, we look to machine learning techniques to investigate whether we can construct predictive models on our variables beyond what simple correlation has shown us.

All described calculations can be seen at https://github.com/jon-e-pizza/Springboard/blob/master/CapstoneFastFoodEstablishments/learning/machine-learning.ipynb

Supervised Learning

When looking to locate a new business we would like to locate where we know we will achieve enough consumers to pay off business expenses. If we predict at CBG level variables we may be able to predict an expected amount of visitors per CBG through Regression models. We also know from our initial investigation that there can be a wide amount of variance in consumer counts, with high trafficked CBGs experiencing much more business than the average location. It may therefore be easier to fit a model simply to predicting whether a CBG would fall into the 'high business volume' class or not, so we will use classification algorithms to develop models to predict these areas. For both regression and classification we expect that our predictions will perform better in many cases if we scale our widely varied measures to the same influence via sklearn's Standard Scaler. We also split our data into a training and test set stratifying on our class of 'Successful Location' to ensure we have an even proportion of desired locations in each of our training and testing data.

Regression

In our initial data wrangling we already totaled fast food consumer count per cbg. We can pull this total as our output variable and then all other variables (excluding count of locations) can be put under consideration as input variables to predicting fast food restaurant traffic in an area.

We can look to simple linear regression to fit a line to our variables by least squares, but we know that not all variables have the same relevance to our desired prediction. We therefore attach cost functions to our regression for the purposes of feature selection. We try both Ridge and Lasso cost functions to perform this derivation, and use a GridSearch on our data looking for the hyperparameter alphas for the cost functions that produces the best R^2 scores. Due to the small size of our data set we end up doing a grid search with only 2 folds for both cost functions. We regard the variables given the highest coefficients by these models as the most important features in selecting our high success areas.

From our Ridge analysis we can confirm what simple correlation had told us previously, which is that the amount of University Students and the raw count of visitors to an area are the best predictors of fast food success in an area. Indeed an R^2 on a linear regression fit on just these two variables is already 0.51 for training data and 0.56 for the test data. While Ridge does

produce some other strong coefficients besides more than just these two variables, we find that none produce a more generalizable model as measured by R^2. When we use our Lasso cost function however, we are actually able to isolate a better generalizing subset of variables for linear regression, achieving R^2 scores of 0.54 for training data and 0.57 for the test data, with the isolated variables of:
'Total Population', 'Undergrads and Grads', 'Diploma/GED attained', 'Above 100k Earners', 'Crime Counts', 'walk_score', 'transit_score', and 'raw_visit_count'

We also use sklearn's RFE to see what set of features this emphasizes, but simply find 'Undergrads and Grads', and 'raw_visit_count' once more.

We also try non linear regression learning tools to see if a less biased tool can generalize better with our data. We fit the decision tree ensemble based RandomForest and GradientBoosting algorithms and in both cases end up with better generalizing models than we were able to achieve with a strictly linear model. First our RandomForest has R^2 scores 0.89 on the training data and .61 on the testing data, and then our gradient boosting model falls a little behind at 0.82, and 0.59. Given the top scoring of our RandomForest model, we note the variables it emphasizes in its prediction:
'Total Population', 'Undergrads and Grads', 'Degree attained', 'Households in Poverty', 'Below 40k Earners', 'Crime Counts', 'walk_score', and 'raw_visit_count'. We see the repeats 'Total Population', 'Undergrads and Grads', 'Crime Counts', 'walk_score', and 'raw_visit_count'.

And while being able to predict high consumership is an important model to have, in order to answer our original inquiry of where we should open a restaurant, a more important model may be: given this expected consumer count as a variable, as well as our previously emphasized variables, what areas appear to currently have fewer restaurants than we would expect?

We expand on our RandomForest Model as it generalized the best of our predictive models, and narrow to the variables that had been emphasized in that analysis. Our dataset is small however, with a small range of output, and unsurprisingly the model generalizes poorly for this use case, having an R^2 score of only 0.05 on the test data, even while achieving 0.94 on the training data. Given the caveat of this poor generalization, we still exercise this model to find that 2 CBGs in our (overfitted) training set, and 5 CBGs in our testing set seem to have lower establishment counts then would be predicted for their areas and could warrant a closer look.

Classification

We then move on to predicting whether locations can be classified as 'Successful' or not, a factor that should act as our first pass in narrowing down locations that we should consider for placement of a new restaurant. We use our previously obtained important features from LASSO regression to train two linear models for classification: LogisticRegression and SVM, aiming for models with high precision, meaning we can be sure an area predicted as successful is correctly predicted as such, and we check for a high area under the receiver operating curve (AUC) to match. In this case, with a bit of tuning, we do arrive at a SVM with precision scores of

1 on both the training and test data, and AUC scores of 0.93, and 0.83 indicating performance well above a random classifier.

We then again try non linear models for a point of comparison and discover a Multilayer perceptron (MLP) trained on our previous emphasized variables from LASSO, keeps the precision of the SVM model, but produces a higher recall (and AUC) making it a preferable predictive model as it can identify more valid locations where we could place a new establishment.

We also try our ensemble based classifiers again, as well as KNeighbors classifiers and a Gaussian Naive Bayes approach, but none seems to have the overall performance of the MLP model.

Unsupervised Learning

Finally we approach our data with unsupervised learning methods, seeing if we can discover relevant consumer groupings that had not been apparent to our supervised models.

We measure the Sum of Squared Errors for various counts of clusters in the KMeans algorithm to derive the most intrinsic cluster count of our data, but reviewing our clustering afterwards it's unclear we can use it to derive a successful segment of areas that might be more successful, as we don't find any cluster to be majority successful locations.

We try instead to cluster by KMeans after running a PCA on the variables that performed best for classification by KNearestNeighbors as KMeans also relies on relative distance of points and this time actually produce a cluster that is 6 out of 7 total locations marked as 'Successful' which indicates an actual segment we would like to pay attention to. We also isolate a 2 of 2 successful cluster which is a little less notable due to size, but we'd still like to see what this segment looks like as well.

Using the 6 of 7 cluster, and comparing to points that fell into other clusters, we define one segmentation of areas amenable to fast food locations as having high counts of:
'Total Population', and 'Undergrads and Grads'
And low counts of:
'Single Occupant Households', 'Single Parent Households', 'Married Parent Households', 'Diploma/GED as Highest Educational Attainment', 'Households in Poverty', 'Income Earners of any Tier', and 'Owner Occupied Households'
Which all seems to line up with locating in areas with universities.

Then our less reliable 2 of 2 cluster seems to identify 'downtown' areas as compared to other clusters as it has high:
'Single Occupant Households', 'Degree attained population', 'Households in Poverty', '40-100k Earners', 'Renter Occupied Households', 'Crime Counts', 'Walk Score', 'transit_score', and 'raw_visit_count'

And low:
'Single Parent Households', and 'Married Parent Households'

We also see if we can find a useful clustering with Agglomerative or DBScan algorithms but again don't produce majority successful clusters to target.

Next Steps

While our final successful location classifiers seem to perform fairly well at identifying areas for fast food consideration, and our regressors identified some areas we might want to look at further for placement of our restaurant, we still could use a better model for predicting how many restaurants an area can sustain. The best thing we could do to advance our models in this respect is to look to other cities for more data points. With enough data points we might be able to create regressors for restaurant count on just locations that fall into the 'successful' category which could be used to predict those locations among our identified successful areas that seem to have room for more restaurants.

Additionally, with more data points, we might be able to create entirely separate models for the two different segmentations we seem to have identified:

- Higher Education Areas
- Downtown Areas

Further, it does appear that more input variables might help us better predict the areas to locate in. For example, we'd like in a future study to uncover what variables underly the correlation between high crime areas and fast food success, and perhaps break down by type of crime as well.

Additionally we would also like to obtain numbers on the profit margins of restaurants to be able to identify areas that might be profitable even if they are not 'high success' areas. With such a model we could discover places with no fast food restaurants currently that might be highly similar to an area with just one currently profitable restaurant.

And if we can complete these models proficiently, we might be able to start generalizing further into the analysis of placing restaurants outside of cities where areas absent any fast food restaurants are bound to be more common, though we would expect that we'd still need to approach this as a separate market segment and do a bit more tailored tuning on our variables.