Machine Learning

Having completed an initial review of Census Block Groups (CBGs) in Philadelphia and comparing those areas with high traffic of fast food consumers with those without, we look to machine learning modeling techniques to investigate the predictive capability of 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

For the purposes of predicting where to locate a new business we would most 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 an area would fall into this higher business volume class. We can then use classification algorithms to develop models to predict these areas, and then separately evaluate how well we think need is currently being met in those areas. For both regression and classification we expect that our predictions will perform better in many cases if we scale our widely varied measure 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 success 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. We regard the variables given the highest coefficients in these models (after scaling the variables) to indicate the most important features in selecting our high success areas. Due to the small size of our data set we end up doing a grid search with only 2 folds for both cost functions.

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 bettew with our data. We fit the decision tree ensemble based RandomForest and GradientBoosting 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. We note then that our emphasized RandomForest variables may be most valuable when it comes to estimating total visitorship to a CBG and note them down: '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'.

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, there is one more model we would like to have, which is, given this expected consumer count as a variable, as well as our previously emphasized variables, what areas appear to 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 evaluate 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 distance of points and this time actually produce a cluster that is 6 of 7 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'