Predicting Fast Food Growth

Kenny Huang

SML310 Spring 2021 Final Project

One of the most pressing but overlooked problems facing our country is access to healthy food. In recent years, increases in obesity and diet-related diseases have faced scrutiny from the general public and public health officials, and as a result, much work has been done to understand the issue, its causes, and implications for the future health of Americans. In reports in 2009 and 2012, the USDA describes the prevalence of "food deserts" – areas in which it is difficult to access affordable and nutritious food. In these areas, often associated with crime and poverty, supermarkets or grocery stores that offer fresh fruits and vegetables are rare, and residents typically have to travel farther to acquire these healthy foods, rely heavily on pre-processed frozen foods, or eat more frequently at cheap and convenient fast food restaurants. Focusing on the latter, tne concern is that fast food restaurants understand this phenomenon and are opening locations in poverty-stricken areas to capitalize on the demand for its services.

In this report, I will seek to address this concern by looking at fast food restaurants in the United States and predicting the trajectory of their growth or decay based on a variety of social, economic, and behavioral factors. Although this topic has been researched before, this prediction approach is novel and can be built on in the future.

Data Acquisition

To answer this question, I sought to find as many features as I could that are relevant to fast food restaurants. Initially, I searched for datasets at the ZIP code level, but some of what I found was only available for counties or even states. I took what I could and did my best to incorporate them into the analysis. In the end, I downloaded the County Business Pattern [1] (CBP) datasets from the US Census Bureau, which included a detailed breakdown of businesses in the US by category. Specifically, according to the North American Industry Classification System (NAICS), businesses classified as supermarkets were tabulated as 44511, while limited-service restaurants, which I will use as a rough proxy for fast food restaurants, are noted as 722513. From the CBP datasets, I extracted these relevant rows to store the number of each type of establishment within each ZIP code.

From the Census Bureau, I was also able to access the median and mean household income values [2] within each ZCTA block. While this was by ZCTA block instead of ZIP code, it turns out that the ZCTA blocks were formed from census blocks to roughly estimate the ZIP codes and, for the most part, are conveniently numbered the same as their corresponding ZIP codes. Using the ZCTA to ZIP Crosswalk [3], I was able to properly convert the ZCTA codes to ZIP codes.

One aspect that I wanted to consider was obesity data, which I was able to find on the CDC website [4]. In particular, I wanted to explore the data on the proportions of adults who reported to eat less than one serving of fruits and and one serving of vegetables per day, which were alarmingly about 40% and 20%, respectively. This data, while interesting, was only available by state and on a biannual basis, preventing it from being as useful as I had hoped.

Finally, the remainder of the data was acquired via the Social Explorer [5], which is an easy-to-use interface to access data from a variety of official sources including the Census and American Community Surveys (ACS).

In particular, the ACS provided an incredible amount of information about a variety of social categories at a level of detail that I didn't know was possible. From here, I exported a number of tables focusing on demographics data.

For all of these sources, I extracted data only from 2016-2019, partly for quality control and partly for feasibility. Once I was able to get everything I needed onto my device, I utilized R and its data manipulation capabilities to generate data points of more than 50 features each that I could then use directly in my analysis.

ZIP Code-level Analysis

To begin the analysis, I pull in the ZIP code data that I had assembled. In total, there are almost 12000 data points detailing the changes in fast food restaurants and supermarkets, number of households, and median household income.

```
zip_data = readr::read_csv("fast_food_predict2.csv")
zip_data %>% dim()
```

```
[1] 11751     6
```

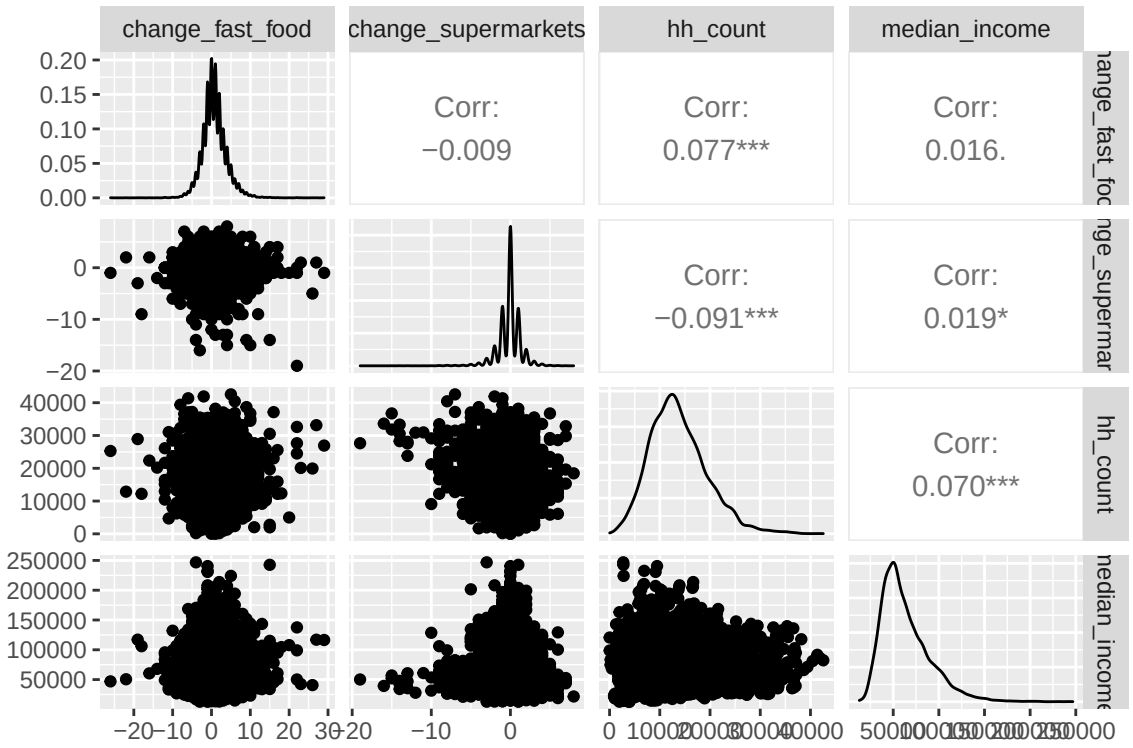
```
zip_data %>% head()
```

```
# A tibble: 6 x 6
  zip    year change_fast_food change_supermarkets hh_count median_income
  <chr> <dbl>            <dbl>                <dbl>    <dbl>         <dbl>
1 01001  2017                4                    0     7460         57694
2 01002  2017               -3                    0     9976         52379
3 01020  2017               -3                    1    12491         58780
4 01035  2017                1                    1     2316         58953
5 01040  2017                0                   -3    15403         37954
6 01060  2017                0                    1     6563         57485
```

Exploratory Data Analysis

Let's quickly explore these variables:

```
zip_data %>% select(-zip, -year) %>% ggpairs()
```



For now, the pairwise plots don't reveal any clear trends relating to the change in fast food restaurants. However, let's dig deeper.

```
zip_data %>% summary()
```

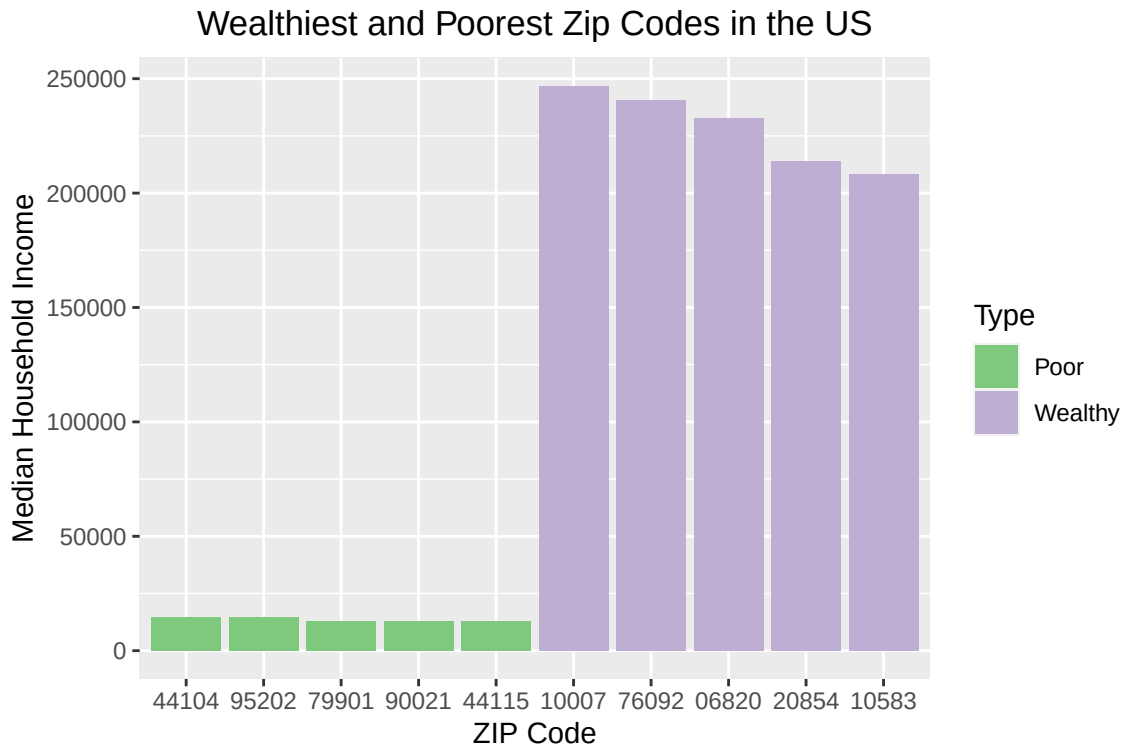
```
     zip                 year      change_fast_food   change_supermarkets
 Length:11751       Min.   :2017   Min.   :-26.0000   Min.   :-19.000
 Class :character   1st Qu.:2017   1st Qu.: -1.0000   1st Qu.: -1.000
 Mode  :character   Median :2018   Median :  1.0000   Median :  0.000
                    Mean   :2018   Mean   :  0.7627   Mean   : -0.135
                    3rd Qu.:2019   3rd Qu.:  2.0000   3rd Qu.:  1.000
                    Max.   :2019   Max.   : 29.0000   Max.   :  8.000
    hh_count     median_income
 Min.   :   21   Min.   : 12800
 1st Qu.: 9731   1st Qu.: 44512
 Median :13073   Median : 57353
 Mean   :13732   Mean   : 63552
 3rd Qu.:17053   3rd Qu.: 77250
 Max.   :42546   Max.   :246813
```

From the summary data, we note that the median of the `change_fast_food` column is 1. Since our overall objective is to predict the trajectory of the growth of fast food restaurants, we will create a binary label variable that will be the target of our prediction models. Because the median value is 1, this label can simply be whether the change was positive or not. This happens to be an impressively close 51/49 split, which will ensure that the predictions are balanced.

```
zip_final = zip_data %>%
  mutate(change_fast_food_sign = as.integer(change_fast_food > 0)) %>%
  select(-zip) %>%
  select(change_fast_food, change_fast_food_sign, year, everything())
zip_data %>% summarize(pos = mean(change_fast_food > 0),
                       neg = mean(change_fast_food <= 0)) %>% as.numeric()
```

```
[1] 0.5076164 0.4923836
```

Finally, looking at the 5 wealthiest and poorest ZIP codes, it's almost jarring how large the disparity is. Even though this isn't related to the main objective of the analysis, I think it's valuable to include as a nice refresher of reality.



One final step before running some preliminary models would be to normalize the numerical variables. This reduces interpretability but is important to prevent the regression model from unfairly preferring one feature over another.

```
zip_normalized = zip_final %>% scale() %>% as.data.frame()
zip_normalized[, 1:3] = zip_final[, 1:3]
zip_normalized %>% select(-change_fast_food) %>% head()
```

```
  change_fast_food_sign year change_supermarkets   hh_count median_income
1                     1 2017          0.09146324 -1.1021144    -0.2190836
2                     0 2017          0.09146324 -0.6599785    -0.4178567
3                     0 2017          0.76913319 -0.2180183    -0.1784688
4                     1 2017          0.76913319 -2.0060679    -0.1719989
5                     0 2017         -1.94154661  0.2937066    -0.9573301
6                     0 2017          0.76913319 -1.2597439    -0.2268999
```

Preliminary Model and Results

Now that we have the data ready, let's try plugging this into a basic logistic regression model. We will compute the percentage of positive truth values as well as our model accuracy, which we hope will be larger.

```
model1 = glm(change_fast_food_sign ~ change_supermarkets + hh_count + median_income,
             data = zip_normalized, family = "binomial")

# accuracy of model if predict all 1's
zip_normalized$change_fast_food_sign %>% mean()
```

```
[1] 0.5076164
```

```r
# accuracy of model
(as.integer(model1$fitted.values > 0.5) == zip_normalized$change_fast_food_sign) %>% mean()
```

```
[1] 0.5292316
```

As low as 53% accuracy is, this is a really promising result. Because the best you can do without information is only 51%, the fact that the model was able to get 53% accuracy on that many data points indicates that it most likely picked up some valuable insights that allow it to consistently perform better than random. However, even so, 53% accuracy is far too low to actually use, so we need to do something to reduce the noise in the data.

County-level Analysis

As we could see from the ZIP code-level analysis, the subdivisions are too fine-grain and contain too much noise for us to derive any meaningful information from it. One possible solution is to make predictions at the county level instead. Fortunately for me, the US ZIP Code database [6] provides the necessary information to map ZIP codes to their respective counties. By joining the tables and aggregating the data, I produced the equivalent tables for the county-level data. From here, I added additional, predominantly social-related, data that was only available at the county level. Notably, the social-related data was also limited to counties with at least 60000 residents, so this greatly reduced our sample size but also improved its quality.

I later further aggregated the data to the state level, but I found that the sample size of just 150 data points (50 states over three years) was too small to draw any meaningful conclusions. However, I had acquired the food behavioral data on daily fruit and vegetable consumption among adults that I wanted to include in the analysis. To do so, I made the assumption that neighboring counties would behave similarly and thus appended each state's value to all counties within that state. Looking back, this was a very strong assumption that ignored many of the significant demographic and economic differences that exist within each state, and as a result these features that I thought would be useful ended up having a negligible impact on the remainder of this exploration.

At this point, the data that we would use for county-level analysis contained almost 2500 points of 53 features.

```
full_data = readr::read_csv('fast_food_county_full2.csv')
full_data %>% dim()
```

```
[1] 2428   53
```

```
full_data %>% colnames()
```

```
 [1] "county"                "year"
 [3] "state"                 "count"
 [5] "change_fast_food"      "change_supermarkets"
 [7] "hh_count"              "median_income"
 [9] "percentfruit"          "delta_percentfruit"
[11] "percentveggie"         "delta_percentveggie"
[13] "state.y"               "total_pop"
[15] "male"                  "female"
[17] "pop<5"                 "pop5_9"
[19] "pop10_14"              "pop15_17"
[21] "pop18_24"              "pop25_34"
[23] "pop35_44"              "pop45_54"
[25] "pop55_64"              "pop65_74"
[27] "pop75_84"              "pop>85"
[29] "white"                 "black"
[31] "american_ind"          "asian"
[33] "native"                "other"
[35] "mixed"                 "pop>25"
[37] "edu<hs"                "edu_hs"
[39] "edu_college"           "edu_bach"
[41] "edu_master"            "edu_prof"
[43] "edu_doct"              "pop>16"
[45] "labor_force"           "labor_employed"
[47] "labor_unemployed"      "not_labor_force"
```

```
[49] "average_household_income" "median_household_income"
[51] "gini_index"               "median_home_value"
[53] "median_gross_rent"
```

One interesting note is that I was able to acquire the median household income at both the ZIP code and county levels from the Census Bureau and the Social Explorer. However, there were some accuracy issues; not only were they from different sources to begin with, but I also chose to aggregate to the county median income values by merely averaging the ZIP code-level values. Fortunately, a quick check confirms that the values are similar enough (6% relative error), and this confirmed that I had correctly aggregated the ZIP code data.

```
full_data_ratios %>%
    mutate(diff = abs(median_income - median_household_income) / median_household_income) %>%
    pull(diff) %>%
    mean()
```

```
[1] 0.06134148
```

Creating the labels at county level

We've already explored the data earlier, but we should still take a quick look at the new version to create the prediction labels. To start, counties saw as many as 528 new fast food restaurants open or 67 close, but most of them saw a change that clustered closely around a median of 2.

```
full_data_ratios$change_fast_food %>% summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-67.000  -1.000   2.000   6.194   8.000 528.000
```

We can now create our label accordingly. Almost 2/3 of counties saw an increase in fast food restaurants, so we can adjust and define our label as whether each county gained more than the median number of two fast food restaurants, which has a much closer 49/51 split.

```
full_data_ratios %>% summarize(`>0` = mean(change_fast_food > 0),
                               `<=0` = mean(change_fast_food <= 0),
                               `>2` = mean(change_fast_food > 2),
                               `<=2` = mean(change_fast_food <= 2))
```

```
# A tibble: 1 x 4
   `>0` `<=0`  `>2` `<=2`
  <dbl> <dbl> <dbl> <dbl>
1 0.637 0.363 0.489 0.511
```

Once we create our label, we can also begin to extract the most relevant features and normalize them for modeling purposes. We will also omit any rows with missing values, which I confirmed had no common relation and will thus not affect our analysis.

```
final_data = full_data_ratios %>%
    mutate(change_fast_food_sign = as.integer(change_fast_food > 2)) %>%
    select(1:5, change_fast_food_sign, everything())
```

```
normalized_data = final_data %>%
    select(5:9, 11, 13, 16:44, 50:55) %>%
    scale() %>% as.data.frame()
normalized_data$change_fast_food = final_data$change_fast_food
normalized_data$change_fast_food_sign = final_data$change_fast_food_sign
normalized_data = normalized_data %>% na.omit()
```

Feature selection

Now that we've created our normalized input features, we can directly compare them to see which values seem the most useful in explaining our target variable. One approach is to look at features individually, compute their average values across positive and negative cases, and then perform t-tests to see if the differences are statistically significant. However, because our data is normalized, we can easily compare features by just directly comparing the difference between the two average values. We are ranking these features this way because we are hoping that these differences are what determine whether a county experiences a gain or loss, so intuitively, bigger differences give more explanatory power.

	diff
change_fast_food	17.7267473300
change_fast_food_sign	1.0000000000
edu_hs	-0.6747877656
edu_college	-0.5395471943
change_supermarkets	-0.3568709498
hh_count	0.3464127278
pop>25	0.3439966050
median_gross_rent	0.2848975072
asian	0.2546661259
pop35_44	0.2244293754
white	-0.2206715047
pop65_74	-0.2187425974
pop75_84	-0.2114739687
pop<5	0.2066475397
pop55_64	-0.2023507684
average_household_income	0.2019883458
median_income	0.1903868737
pop45_54	0.1833796620
median_home_value	0.1831388653
black	0.1796666793
pop25_34	0.1713781618
other	0.1659069281
pop5_9	0.1615963723
median_household_income	0.1581533216
american_ind	-0.1549339208
delta_percentveggie	0.1458490179
edu_bach	-0.1445183273
edu_prof	-0.1328925261
male	-0.1301151417
female	0.1301151417
gini_index	0.1244518135
pop15_17	0.1151986519
pop>85	-0.1131342122
pop10_14	0.1034580133
edu_master	-0.0665545834
edu<hs	0.0659464755
pop18_24	-0.0609199712
unemployment	0.0388537223
delta_percentfruit	-0.0367118467
edu_doct	0.0161557744
mixed	-0.0097959193
native	0.0003723481

As we might expect, the raw change in fast food restaurants varies heavily between the two categories (also partly because it is not normalized). After this, we see other notable features: high school and college educated residents, change in supermarkets, age brackets, and county size indicators like the number of households or number of residents age 25 or above. We can briefly examine these and what effects we expect them to have on our models.

High School and College Education This feature tells us the proportion of the population above the age of 25 that only have a high school education or are currently attending college. Interestingly, the two differences are extremely negative, indicating that a lower proportion of less educated citizens actually correlates to a higher chance of gaining fast food restaurants. This runs counter to the initial concern that fast food restaurants open up in areas of poverty to take advantage of demand and would be worth exploring in future analysis.

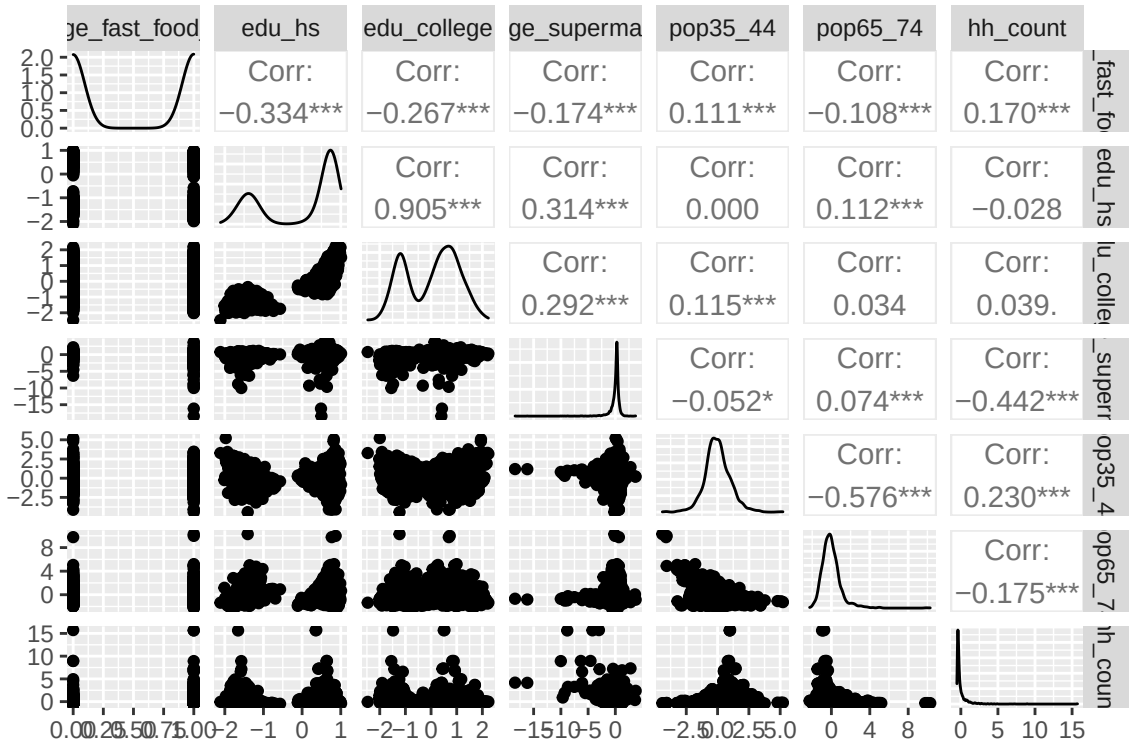
Change in Supermarkets Again, we see a strongly negative difference, indicating that supermarkets closing is linked to fast food restaurants opening and vice versa. This makes intuitive sense, because land/commercial property is a limited good, but this effect is actually amplified if you consider that the two establishments are also directly linked. Growing communities would attract both establishments and deteriorating areas would lose both, so the fact that the two are so antagonistically related is quite remarkable.

Age Brackets Here we note that the 35-44 age bracket has a positive difference, while the 65-74 and 75-84 bracket ones are negative. This aligns with what we might expect; young parents with budding careers and young children would be frequent customers, creating a larger demand for fast food restaurants in areas where this demographic is more common.

Size Indicators Other features like the number of households and the number of citizens above the age of 25 have positive differences. Indeed, established counties with larger populations have greater demand for fast food restaurants and will generally see growth.

Pairwise Relationships In terms of pairwise relationships, we see obvious correlations between high school and college educations as well as the different age brackets, but nothing else particularly stands out.

```
normalized_data %>%
    select(change_fast_food_sign, edu_hs, edu_college, change_supermarkets,
           pop35_44, pop65_74, hh_count) %>%
    ggpairs()
```



Modelling Setup

Now that we've generated our data and examined what features we should include in the model, we can create the training and testing sets for the modelling section. Because we will be using relatively basic models with few hyperparameters, there is no need for a validation set for hyperparameter tuning, so we can just create the classic 80/20 split.

```
set.seed(42)
n = nrow(normalized_data)
idx = sample(1:n, 0.8 * n)
train = normalized_data[idx,]
train %>% dim()
```

```
[1] 1825   42
```

```
test = normalized_data[-idx,]
test %>% dim()
```

```
[1] 457  42
```

In addition, for the sake of simplicity later I have defined functions that compute the accuracy and confusion matrices of a given model:

- train_acc_log
- test_acc_log
- train_acc_svm

- test_acc_svm
- conf_matrix_log
- conf_matrix_svm

For the following sections, note that the dataset has about a 50-50 split in positive and negative data points. Thus, in our prediction models, getting an accuracy of above 50% would be indicative that our features indeed have predictive power. Once the confusion matrix is computed, the p-value is defined as the likelihood of seeing an accuracy at least as good as the model accuracy by randomly guessing. If the p-value is sufficiently small, then we can conclude that the model is consistently better than random.

Logistic Regression

First, let's try to create a logistic regression model. To test the waters, we can try just including the supermarkets feature.

```
log_model1 = glm(change_fast_food_sign ~ change_supermarkets, data = train, family = "binomial")
train_acc_log(log_model1, train$change_fast_food_sign)
```

```
[1] 0.6010959
```

```
test_acc_log(log_model1, test, test$change_fast_food_sign)
```

```
[1] 0.5929978
```

```
conf_matrix_log(log_model1, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 174 138
         1  48  97
```

Surprisingly, this basic model actually performs very well! We get training and test accuracies of 60% and 59%, respectively, the later of which corresponds to a p value of less than 0.0005. This is an extremely encouraging sign, and because the accuracies are close, we can try to make our model more complex without fear of overfitting.

```
log_model2 = glm(change_fast_food_sign ~ change_supermarkets + edu_hs,
                 data = train, family = "binomial")
train_acc_log(log_model2, train$change_fast_food_sign)
```

```
[1] 0.6613699
```

```
test_acc_log(log_model2, test, test$change_fast_food_sign)
```

```
[1] 0.6301969
```

```
conf_matrix_log(log_model2, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 175 122
         1  47 113
```

By adding the strongest feature of `edu_hs`, we improve our test accuracy to 63%, for a p-value on the order of 10^{-7}.

```
log_model3 = glm(change_fast_food_sign ~ change_supermarkets + edu_hs + hh_count + pop35_44,
                 data = train, family = "binomial")
train_acc_log(log_model3, train$change_fast_food_sign)
```

```
[1] 0.6783562
```

```
test_acc_log(log_model3, test, test$change_fast_food_sign)
```

```
[1] 0.6586433
```

```
conf_matrix_log(log_model3, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 166 100
         1  56 135
```

Now, by including features from all four categories I explored earlier, we're able to improve our model to almost 66%. In other words, we're right about twice as often as we're wrong, which is incredible considering that we are just using 4 simple features.

Support Vector Machines

Given our success with the logistic regression model, we can expect similar if not better results with an SVM model. We'll create models with the exact same features and see how they perform.

```
svm_model1 = svm(factor(change_fast_food_sign) ~ change_supermarkets, train, scale = FALSE)

test_acc_svm(svm_model1, test, test$change_fast_food_sign)
```

```
[1] 0.5798687
```

```
conf_matrix_svm(svm_model1, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 179 149
         1  43  86
```

```
svm_model2 = svm(factor(change_fast_food_sign) ~ change_supermarkets + edu_hs,
                 train, scale = FALSE)

test_acc_svm(svm_model2, test, test$change_fast_food_sign)
```

```
[1] 0.619256
```

```
conf_matrix_svm(svm_model2, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 171 123
         1  51 112
```

```
svm_model3 = svm(factor(change_fast_food_sign) ~ change_supermarkets + edu_hs + hh_count + pop35_44,
                 train, scale = FALSE)

test_acc_svm(svm_model3, test, test$change_fast_food_sign)
```

```
[1] 0.6455142
```

```
conf_matrix_svm(svm_model3, test, test$change_fast_food_sign)
```

```
          Reference
Prediction   0   1
         0 167 107
         1  55 128
```

Conclusions

Clearly, the SVM models are still objectively good but not as impressive as the logistic regression models. I suspect this is due to the fact that the data is still quite noisy and not able to capitalize on SVM's strengths. Regardless, we are able to achieve 66% accuracy on a binary classification problem, which is quite promising.

Looking at the complex logistic regression, we can get a sense for which features were the most important. Although the supermarket feature alone was able to create a proficient model, when combined with the other features in the complex model it became much less important and had the only non-significant coefficient. Thus, this is indicative of hidden interactive effects with the other variables, but even so we can still consider it as a valuable feature in our model.

```
log_model3 %>% summary()
```

```
Call:
glm(formula = change_fast_food_sign ~ change_supermarkets + edu_hs +
    hh_count + pop35_44, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1213  -0.9328  -0.7514   0.9644   1.7732

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.02064    0.05147   0.401   0.6884
change_supermarkets -0.01484    0.06917  -0.215   0.8301
edu_hs              -0.74399    0.05668 -13.126  < 2e-16 ***
hh_count             0.54442    0.09274   5.871 4.34e-09 ***
pop35_44             0.13385    0.05300   2.526   0.0116 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2530  on 1824  degrees of freedom
Residual deviance: 2231  on 1820  degrees of freedom
AIC: 2241

Number of Fisher Scoring iterations: 5
```

One thing that I'd like to point out is that looking at the confusion matrices, the models were all skewed toward predicting 'no', and almost all the improvement from adding features was manifested in the model generating positive predictions more frequently and correctly.

Summary

In this report I was able to create a logistic regression model able to predict the trajectory of fast food restaurant growth within counties with 66% accuracy. The model uses publicly available data and generalizes well on unseen test data. Notable conclusions include that contrary to previous expectations, an increase in education leads to an increase in fast food growth. Of course, the data is only for counties above a certain size from 2016-2019, so this surprising result may be from a combination of limited sample size and possible recent trends. Future work should be done to determine which of the two is more responsible.

Future Work

This project was a very preliminary look at a complex topic. In the future, more analysis can be done with more features and more years' worth of data (both before 2016 and after 2019). Possible additional features can also include the change in current feature values between years. In addition, there were notable omissions in the data used in this report; future iterations of this analysis would benefit from county-level detail on the behavioral CDC data and the social data for all counties regardless of size.

Reflections

Overall, this was a very interesting but challenging project to work on. The main difficulty lay in the different forms that each dataset took and needing to complete the necessary pre-processing to marry these datasets together into something useable for the analysis. In the past, most of my data analysis projects involved just one or two datasets, but this project involved data from many different websites, each with their own data dictionaries and unique documentation. To complete the data manipulation, I meticulously wrote 4 separate R scripts that created intermediate spreadsheets that I could verify were working properly. This project really tested my ability to keep track of the information that I had and what I didn't have, and I found that taking notes in a separate document and spreadsheet was very helpful in this regard. Throughout the process, I also encountered other difficulties that I learned from.

As I mentioned earlier, I was originally planning on completing the analysis at the ZIP code level, but ZIP code regions contained an average of less than 10 fast food restaurants. Even despite the significant noise, the results indicated that the features had some explanatory power and that it would be worthwhile to continue examining this relationship in a different way.

In aggregating the ZIP code data into county-level data, I encountered some minor issues. The biggest mistake was that I forgot that counties across states can share names; for example, there are 9 distinct "Jefferson Counties", 8 "Washington Counties", and worst of all, 5 "Orange Counties". The Orange County values were actually what tipped me off to my mistake, because it caught my eye that somehow 5 different counties experienced an exact 67-store decrease in the same year. Once I added a state column and then re-joined the tables with the additional 'state' key, the relative error in median household income also dropped from 40% to 6%. I was very fortunate to have caught that error early before the bulk of the analysis.

Acknowledgements

I would like to take this opportunity to thank Professor Hanke for his continuous hard work and guidance this semester, and Dr. Ofira Schwartz-Soicher, who was extremely helpful in brainstorming and finding the data that I needed to complete this project. Thank you for reading this far, and I'm always happy to discuss it further at kh19 [at] princeton [dot] edu.