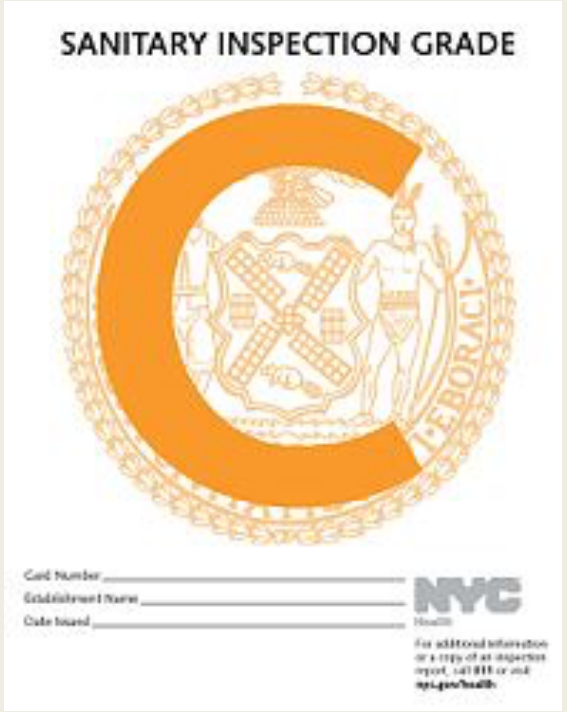
Jesse Kranzler
DAT-NYC-37
8-8-16

SHOULD I EAT THERE?

Finding Correlation between NYC DOH & YELP ratings

Or, for those of you who are visual learners…

DOES 

=

?

My primary data set

- NYC Open Data contains many data sets relating to all sectors of life in NYC. One of these is the results of DOH's restaurant inspections (data set available here)
- The data includes business names, addresses, violation descriptions and many more germane data points. For a full data dictionary, please refer to my iPython Notebook.

Why am I doing this?

- NYC Department of Health restaurant ratings are not exactly the best representation of a restaurant's quality.
- Restaurants with B or C ratings often lose a huge percentage of their clientele over a simple letter grade.
- Recently there has been controversy around certain DOH standards actually being prohibitive for chefs. For example, it is currently mandated that all chefs wear gloves while handling raw proteins. This has been a major problem for sushi chefs who consider the relationship between fish, rice, and the chef's hands to be the key factors to making quality sushi. Additionally, chefs are required to freeze fish for a minimum of 15 hours before serving it raw, which has a huge impact on the fish's flavor and texture. (Read more here)
- Because of some of these quirks in the DOH rating system, I believe that a restaurant's letter grade is not necessarily an indicator of quality.

How will I test this hypothesis?

- Yelp is one of the largest sources of customer reviews on the internet. In addition to detailed reviews & photos on each business, Yelp sums all of these up in an easily digestible 0-5 star rating.
- Since Yelp has an API that offers free access, I decided to use it to add the dimensions of Yelp Star Rating & Yelp Review Count to my existing dataset.

Using the API

- There were three problems right off the bat in connecting my dataset with Yelp's API
  - *1. I'd never used an API before*
  - *2. The DBAs in my dataset did not match business names in Yelp*
  - *3. My dataset included 451K+ entries while Yelp's daily API call limit was 25K*
- Overcoming these obstacles
  - *1. The holy trinity of Anthony, Zunayed, and ye olde Internet taught me the ways of API usage*
  - *2. While the business names didn't match, I noticed that business phone numbers were consistent across both data sources.*
  - *3. Since this project is just an exercise, I chose a random sample of 40K entries, which I crawled in 20K chunks over 2 days then concatenated my results.*
- To view my API calls, cleaning of the data, and uploading my final list, please view my notebook

Exploratory Analysis

- Now that my data was (relatively) clean, I wanted to make a few plots to see which predictors had an impact on my target variable (”GRADE”)
- Refer to my notebook for charts and explanations of each.
- During this process, I decided it would also be worthwhile to create a new column in my dataset (”weightedavg”). This way I was able to slice my dataset by GRADE and see how each rating compared to the average for that letter grade. I did this because a vast majority of my samples were rated A, so I figured creating weighted averages would help avoid skewness down the line.
- I also put together crosstabs to evaluate GRADE by Cuisine and Zip Code with a percentage breakdown of each.

Modeling the Data

- I performed four different classification modeling techniques on this dataset: K-neighbors Classifier (KNN), Decision Trees, Random Forest, and Logistic Regression.
- I considered this a classification problem because my target variable contained three possible outputs (A, B, & C.)
- The first three modeling techniques were quite simple, but I had to take a slightly different approach when applying Logistic Regression.
  - *Since logistic regression requires a binary target variable, I dummy encoded my 'GRADE' column into boolean values ("isA", "isB", "isC")*
  - *I applied logistic regression on all three of these binary variables then took the weighted average of all three*

My results

Model	Accuracy Score
KNN	0.765825136106
Decision Tree	0.7448033555391064
Random Forest	0.7404391808536886 5
Logistic Regression*	0.9955584529977789

*While it might appear that Logistic Regression was the best model by far, this insanely high accuracy score is likely the result of the model overfitting. We could confirm this by running the model on our test set. We'd expect the model to perform very well on the training data, but very poorly on the test data.

But wait, there's more…

- Since the DOH dataset also provides a continuous variable for restaurant grade ('SCORE'), I decided to try applying Linear Regression to this dataset.
  - *A score of less than 14 points results in an "A" grade*
  - *A score between 14-27 points results in a "B" grade*
  - *A score of 28 points or greater results in a "C" grade*
- At the end of the day, my R^2 score was pretty dismal (0.0413), so let's pretend this never happened.

Having fun with the data

- I decided to test hypothesis on three of my preconceived notions about this data
  - *1. Certain cuisines are more likely to have C ratings than others*
  - *2. Certain neighborhoods are more likely to have C ratings than others*
  - *3. Larger chain restaurants should have less variability in DOH grade as they likely have unified codes on cleanliness.*
- The results were fairly surprising (see notebook for details and numbers)
  - *1. This proved true, but not as I expected. The top three cuisines were Chilean, Bangladeshi, and Eastern European*
  - *2. This also proved true in an unexpected way. All five boroughs were represented in my top ten neighborhoods.*
  - *3. This proved false. While some large chains had relatively small standard deviations in score, others had very large values.*

Moving forward: How could I improve?

- There are a few very simple solutions:
  - *Increase the size of the data set*
    - We only used about 20% of the dataset for this project. The more data we have, the more accurate our model can be.
  - *Replace NaN values instead of dropping them*
    - We could have filled NaN values in with column averages or done additional research to fill additional values in by hand (for example, we could replace the "Missing" boroughs by searching for addresses and zip codes)
  - *Feature selection*
    - We could choose different combinations of predictors to focus on

How can I improve: Additional data

- Additional Yelp data
  - *The $ rating from Yelp could be a very useful score. We'd expect $$$ restaurants to be fine dining "fancy" restaurants that should* have higher cleanliness standards
- Additional 3rd party data
  - *Sources like FourSquare would be a great complement to the Yelp data. Additionally, we could mark restaurants for their presence on various prestigious restaurant rankings (Michelin, The World's 50 Best, Opinionated About Dining, etc.)*
- Additional geographic data
  - *We could research external issues that prevent a restaurant from passing health checks (a subway running under the restaurant, proximity to a waste facility, etc.)*
- Owner / proprietor
  - *It's likely that restaurants in the same restaurant group / chain will have similar codes of cleanliness so we should expect grades to be fairly consistent (see examples below)*

Additional Data (Continued)

- Social Factors
  - *If employees are treated well they are more likely to care about their place of employment and put their best efforts towards keeping it clean.*
- Hype / Press
  - *Restaurants that are often in the public spot light might be looked on with additional scrutiny. We could compare the time of DOH rating to times of relevant reviews of the restaurant.*
- Etc…

Conclusions

- Like any real world dataset, there are infinite numbers of data points that could be added. I think we have shown that Yelp star ratings and DOH restaurant grades are not as highly correlated as one might have thought, but I'm sure there are other factors that would be better predictors.

Acknowledgements

- Thanks to my data sources: NYC OpenData & Yelp
- Thanks to my knowledge sources: Anthony & Zunayed
- And thanks to all restaurant workers who don't wash their hands after they shit for giving me inspiration for my project