San Francisco Restaurant Health Scores

A supervised learning capstone

Outline

1. Research Question
2. Overview of Dataset
3. EDA and Feature Engineering
4. Models, results, and drawbacks
5. Who can use this model and for what purpose?

What are we trying to learn?

Can a restaurant's consumer reviews help predict its health score?

Why does this matter?

2011:

> CDC estimates 48 million people get sick, 128,000 are hospitalized, and 3,000 die from foodborne diseases each year in the United States.

Source: https://www.cdc.gov/foodborneburden/estimates-overview.html

Not much to work with...

About 50,000 records from San Francisco Restaurants.

Pulled from the Kaggle open data set found here.

Business Information

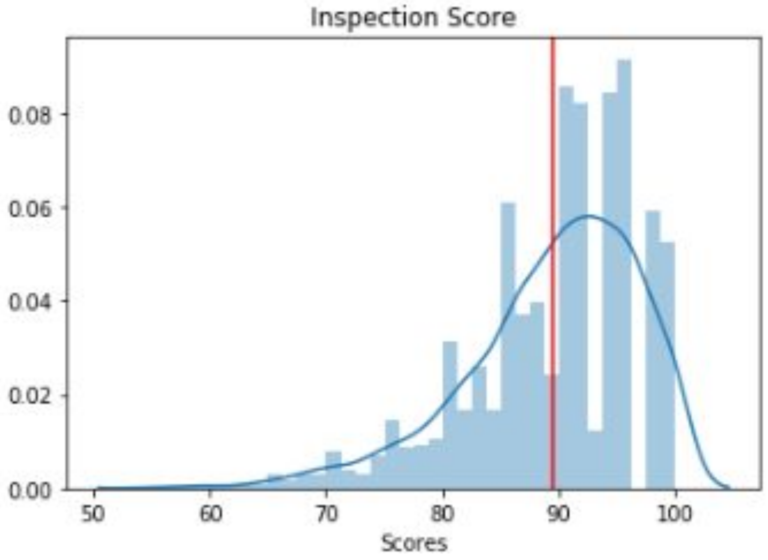
Name
Latitude
Longitude
Phone Number
Postal Code
Neighborhoods
Polic Districs
Fire Districts

Inspection Information

Inspection Score
Date
- Earliest: 8/2/2016
- Most Recent:: 8/1/2019

Violation Category
Violation Description

What does our inspection score look like?



Mean: 89.53

Standard Deviation: 7.43

Other Noticings:

- Long tail
- High scores almost discreet

Research Question

Overview of Dataset

EDA & Feature Engineering

Models, Results, & Drawbacks

Who can use this model and how?

Data Cleaning

1) Drop rows where there the scores and violations are incongruous
    a) Scores with 100 but had violation ids/descriptions
    b) Scores with less than 100 but no violations
2) Scores with 100 and null descriptions are changed from null to 'No Violation'
3) Duplicate Entries Removed
4) Filled null values of basic business information with the median
5) About 12,000 unique inspections

Feature Engineering via Data Mining

Kaggle open Data

Added: Number of Violations Count and Risk Category Counts

Yelp Fusion API

Average Yelp Review

3 Most Recent Reviews

Restaurant Price

TextBlob API

Average Sentiment Analysis

Research Question — Overview of Dataset — **EDA & Feature Engineering** — Models, Results, & Drawbacks — Who can use this model and how?

How do certain features compare to our target

feature	r_squared	p_value
num_violations	0.714909	0.000000e+00
high_risk_count	0.629531	0.000000e+00
medium_risk_count	0.483993	7.941033e-227
no_risk_count	0.351163	7.068552e-113
low_risk_count	0.277726	1.375526e-69
yelp_rating	0.101012	2.922536e-10
review_sentiment	0.084302	1.471651e-07
review_rating	0.070498	1.117630e-05
price	0.046187	4.026364e-03

Violation Information is at the top followed by Yelp Data

Feature Selection

1) All names (as dummies)

2) Calendar Information (as dummies)
   a) Date
   b) Day
   c) Year
   d) Month

3) Numerical Features
   a) SF Data - Neighborhoods, Fire Prevention Districts, Police Districts, etc
   b) Yelp Data - review sentiment, review rating, price, overall rating

4) Does not include violation information

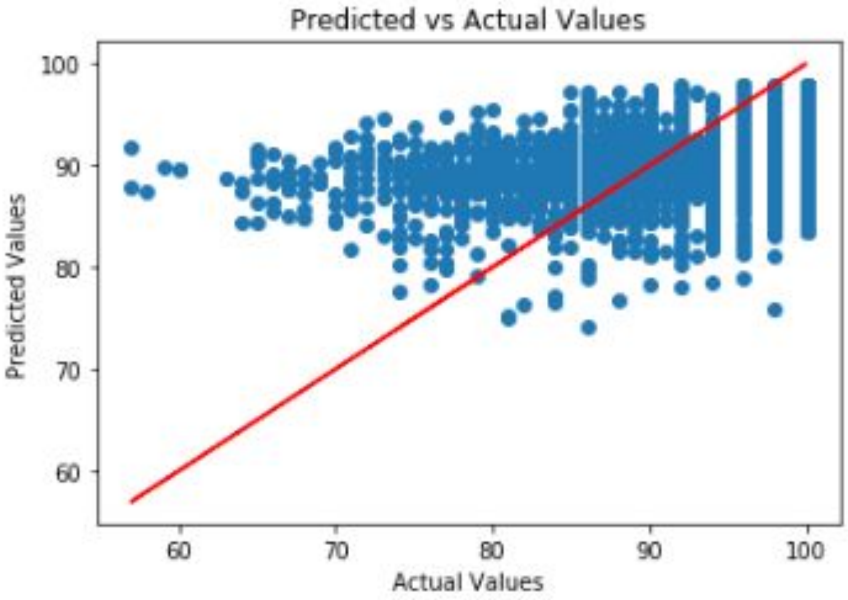
What does our dataframe look like?

Step	Number of Records	Number of Columns
Original	53,732	20
Remove duplicates - add & remove features	12,020	22
Merge Yelp Data	3,876	26
Used Features	3,876	1,202

How did our models perform?

Model	Test R^2	RMSE	MA % Error
Random Forest Regression	41.43%	5.47	4.71%
Bagging Regressor	41.37%	5.48	4.67%
SVR w/stand	36.70%	5.64	4.91%
Ridge	36.08%	5.72	5.11%
Gradient Boosting Regressor	34.02%	5.81	5.15%
KNN w/ Standard - Weights	15.73%	6.57	5.45%
SVR w/PCA	25.30%	6.18	5.45%
LRM w/PCA	21.97%	6.32	5.60%
KNN w/ Standard - Uniform	19.47%	6.42	5.28%
AdaBoost	6.43%	6.92	6.31%

Random Forest



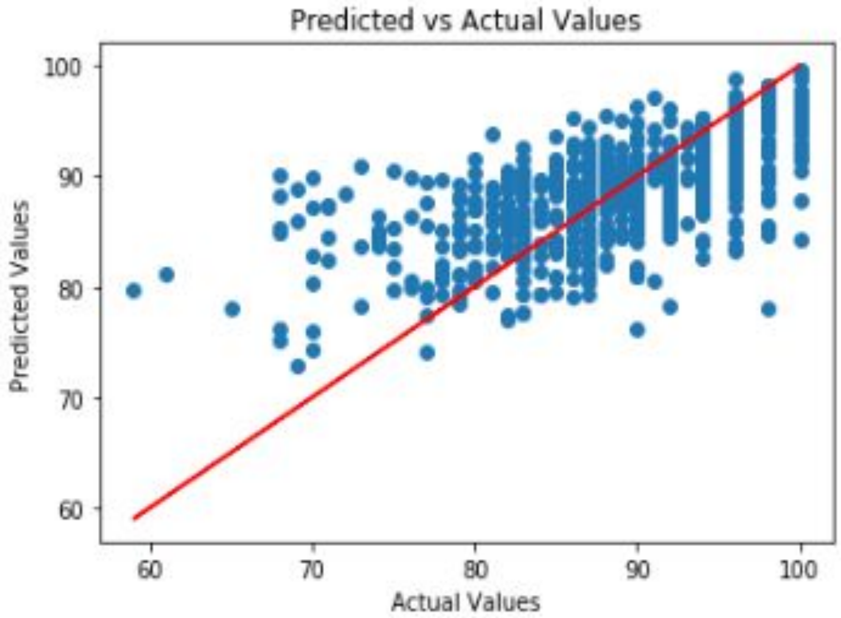
Noticings:

1) Overvaluing small values
2) Almost all predictions between 80 - 100

How important are the yelp features?

columns	importance
review_sentiment	0.055099
yelp_rating	0.042255
review_rating	0.037749
Neighborhoods	0.028515
Analysis Neighborhoods	0.027998

Bagging Regressor



Noticings:

1) Over predicting low scores
2) Almost all predictions between 80 - 100

How did our models perform without the Yelp data

Model	Test R^2	RMSE
Ridge	28.91%	6.57
Bagging Regressor	16.57%	7.11
Random Forest Regression	14.52%	7.2
KNN w/ Standard - Weights	-0.55%	7.88

Comparison

Model	Test R^2	RMSE % Change
Ridge	7.17%	14.86013986
Bagging Regressor	24.80%	22.37521515
Random Forest Regression	26.91%	9.589041096
KNN w/ Standard - Weights	16.28%	27.50809061
Averages:	**18.79%**	**18.58312168**

Where can this model improve?

- Reviews are 3 most recent reviews - but health scores are from much earlier
  - Possible Solution: This is a living dataset - continue to gather data that matches until many features
- Not many records after pulling in yelp information
  - Possible Solution: See above or integrate a different dataset - google reviews?
- Does poorly predicting low scores
  - Possible Solution: Integrate some unsupervised clustering to discern target groups or work with potential stakeholders to determine

Who could benefit from this model?

1) Restaurant patrons
2) Restaurant management/owners
3) Health inspectors
4) CDC - potentially gather more foodborne illnesss data

What did we learn/gain?

Can a restaurant's consumer reviews help predict its health score?

There's potential, but more research and time would be needed to improve model's accuracy

Appendix

Duplicate Entries

San Francisco Open Dataset

Yelp Created Features

Textblob Created Features

Number of Violations - Visualization

Yelp Data - Visualization

Duplicate Entries?

business_name	business_postal_code	business_latitude	business_longitude	inspection_id	inspection_date	inspection_score	inspection_type
Parada 22	94117	37.769303	-122.451961	62051_20180514	2018-05-14T00:00:00.000	86.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20180514	2018-05-14T00:00:00.000	86.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20180514	2018-05-14T00:00:00.000	86.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20180514	2018-05-14T00:00:00.000	86.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20190509	2019-05-09T00:00:00.000	88.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20190509	2019-05-09T00:00:00.000	88.0	Routine - Unscheduled
Parada 22	94117	37.769303	-122.451961	62051_20190509	2019-05-09T00:00:00.000	88.0	Routine - Unscheduled

Created Features - SF open Dataset

Kaggle open Data

Basic Business and Inspection Information

- Total Number of Violations
- Counts of the different risk categories
  - No Violation
  - Low Risk
  - Medium Risk
  - High Risk
- This is a dynamic dataset

Created Features - GET Requests with Yelp's API

Yelp Fusion API

Average Yelp Review

3 Most Recent Reviews

Restaurant Price

Steps to take:

1) Pull in each businesses' unique yelp id

2) Get requests to pull in certain features:

   a) Restaurants average yelp review
   b) 3 most recent reviews
   c) The restaurant's price (converted from '$' to numeric 1-4
   d) The average rating of the 3 reviews

Created Features - Textblob and Sentiment

TextBlob API

Average Sentiment Analysis

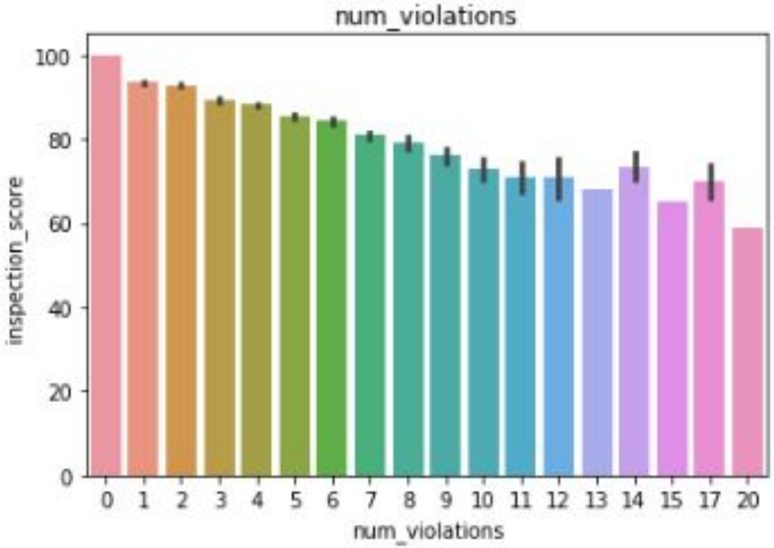
Information and how-to pulled from this site.

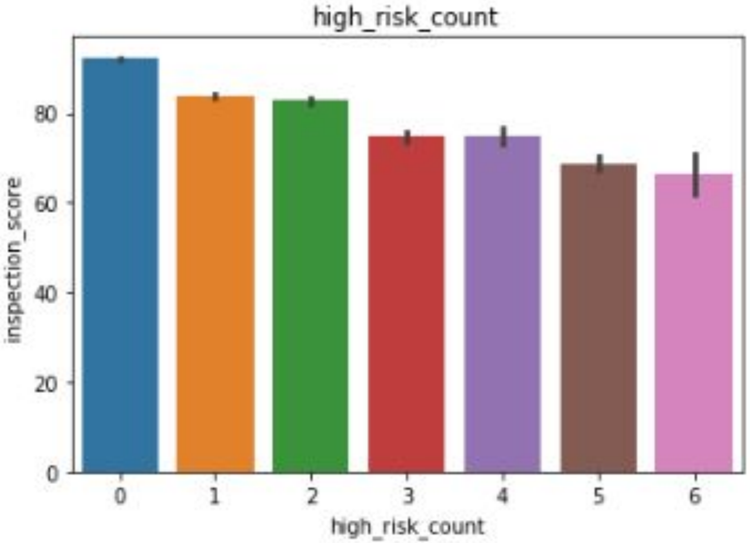
"TextBlob stands on the giant shoulders of NLTK and pattern, and plays nicely with both."

From the Yelp API:

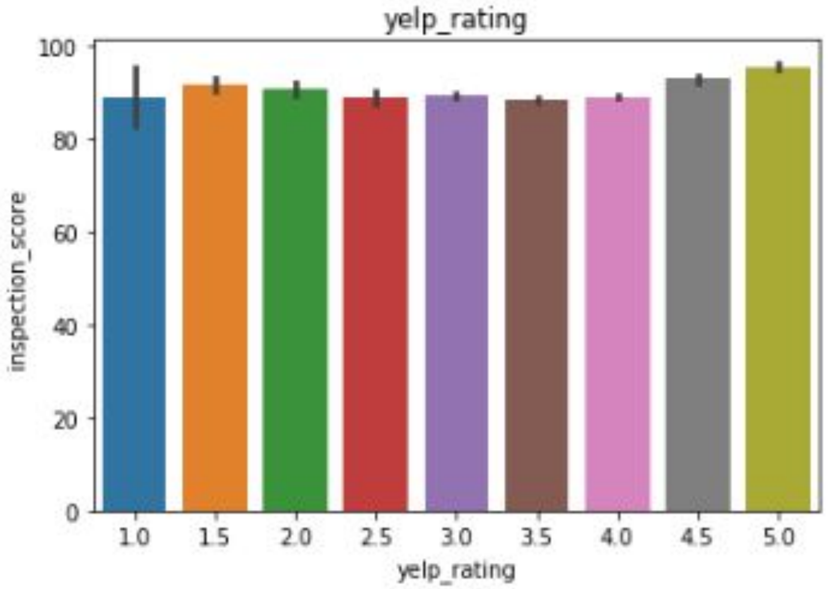
- Get the sentiment from the 3 reviews and average them
- Scale from -1 to 1

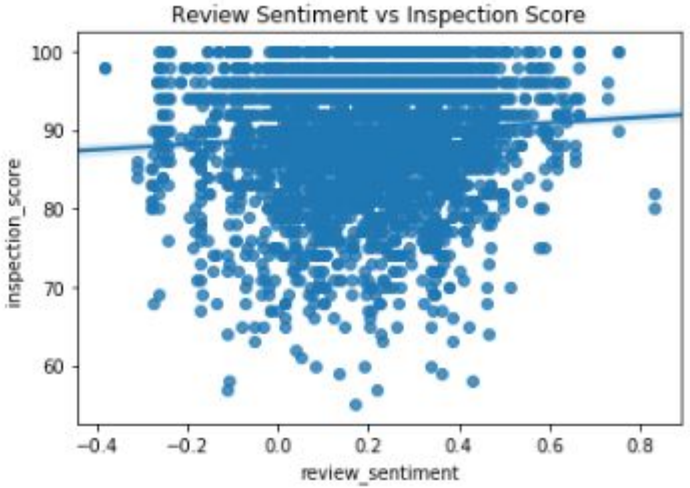
Number of Violations





Yelp Data


yelp_rating
inspection_score
100
80
60
40
20
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
yelp_rating


Review Sentiment vs Inspection Score
inspection_score
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
review_sentiment