Predicting Health Violations With NLP

Robert Cauvy

Business Problem

More than 48 million Americans per year become sick from food

Most cities' health inspections are conducted at random

Objective: Reduce time spent on spot checks at clean restaurants that have been following the rules closely — and improve health and hygiene at places where food safety issues are higher risk

Data Sources

13,061 Unique Records

9,158 Never Failed

3,903 Have Failed

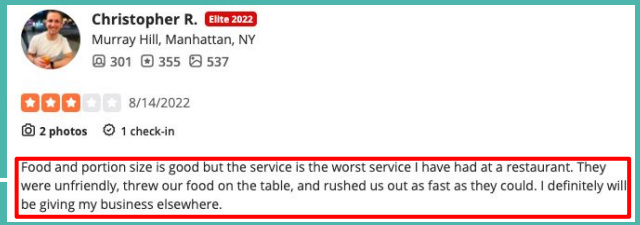
NYC Open Data Portal

- DOHMH New York City Restaurant Inspection Results

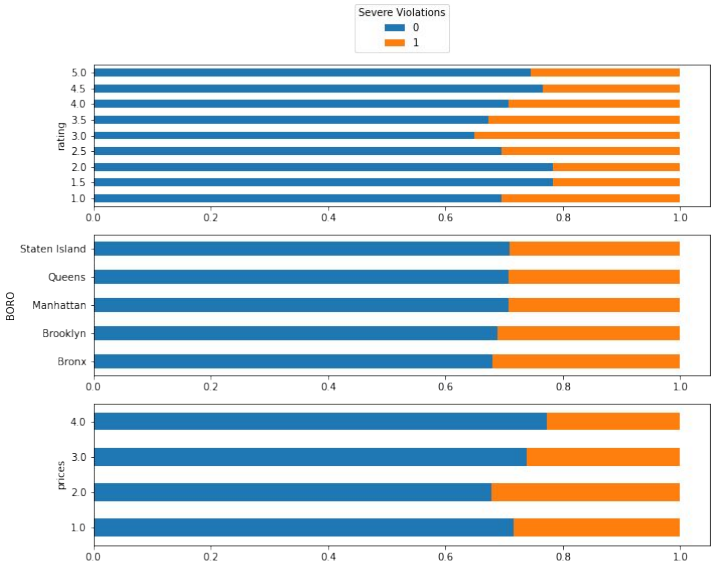
Yelp API

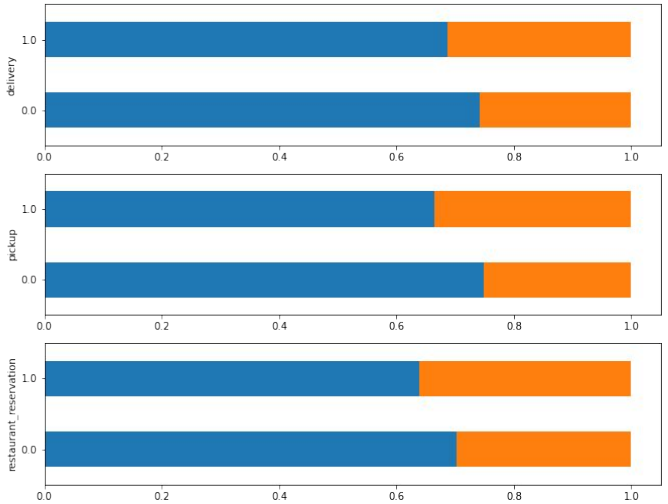
- Yelp Fusion Phone Search

Yelp User Generated Business Reviews



Categorical Exploratory Data Analysis





Exploratory Text Data Analysis

Models

Bag-of-n-grams	Fitting & Tuning	Results
• Naive Bayes	• Count Vectorized • Tf-idf Transformed • Stopword Removal • Unigrams and Bigrams • Tuned with GridsearchCV • Optimized For Recall	71% Accuracy / 22% Recall
• Logistic Regression		57% Accuracy / 60% Recall
• Decision Trees		53% Accuracy / 65% Recall
• SVC		58% Accuracy / 51% Recall

Best Model

DecisionTree Classifier

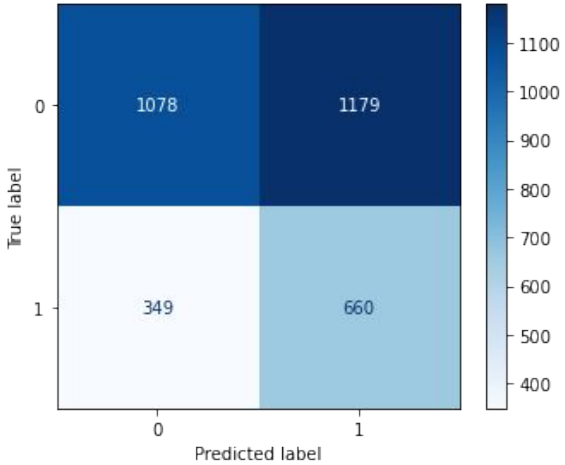
{'class_weight':'balanced',

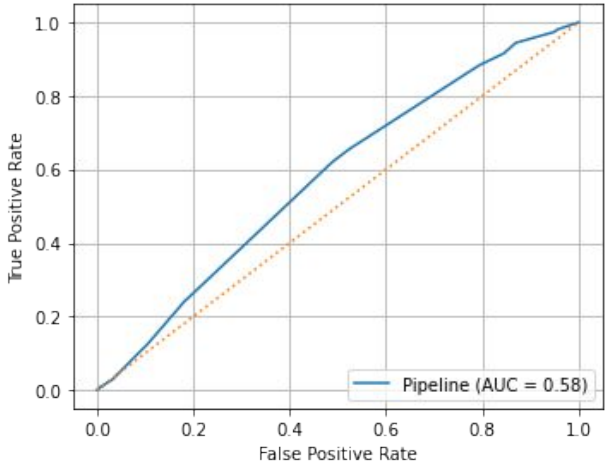
'criterion': 'gini',

'max_depth': 4}

- Highest Recall
- Highly Interpretable
- Robust to overfitting
- Computationally inexpensive

	precision	recall	f1-score	support
0	0.76	0.48	0.59	2257
1	0.36	0.65	0.46	1009
accuracy			0.53	3266
macro avg	0.56	0.57	0.52	3266
weighted avg	0.63	0.53	0.55	3266





Next Steps

- Revisit Target Variable
- Preprocessing methods
- Hyperparameter Tuning
- Deep NLP Methods
- Deploy model for public use
- Target Restaurant Stakeholders

Contact

Robert Cauvy
rcauvy@gmail.com

linkedin.com/in/robert-cauvy/

Complete analysis available at
github.com/rcauvy

Appendix - Deep NLP

Word Embeddings	Preparation	Results
• Empty Embedding	• Skip-gram and Continuous Bag-of-Words Embeddings •	55% Accuracy / 38% Recall
• Pre-Trained Word2Vec		32% Accuracy / 97% Recall
• Pre-Trained GLOVE		TBD