Using Online Reviews To Predict Restaurant Closures

By Luke McKinley

The Yelp Problem

The restaurant industry operates on razor-thin margins and with the advent of online review sites like Yelp, many business owners feel beholden to an army of armchair restaurant critics. In my professional background, I have spent hours managing online reviews and the people who write them. At their best, they can offer valuable feedback to business owners, and at their worst they are subversive attempts to receive free product at the threat of a bad review. It was my goal to see if online reviews can truly be a barometer of success for a restaurant, and in my analysis I have found that yelp reviews do not serve to indicate whether a restaurant will go out of business.

Collecting Data

The data was compiled by yelp and can be obtained at https://www.yelp.com/dataset

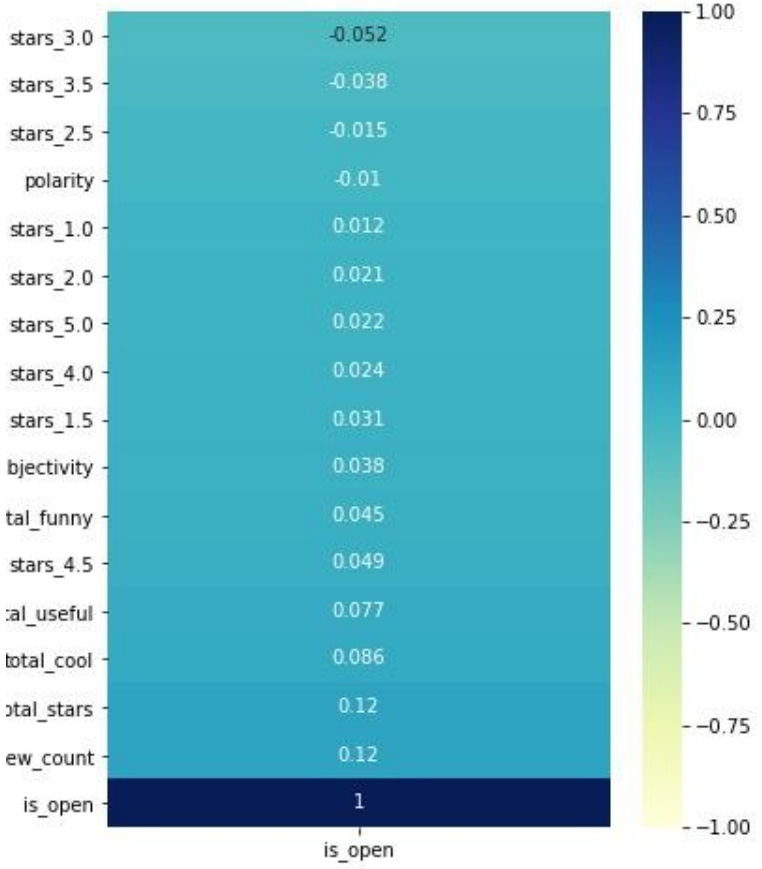
The original files contain over 8 million reviews of 209,000 businesses. For the scope of this project, I winnowed the dataset to include only restaurants. Vaex was used to convert the data from JSON format and TextBlob was used to perform sentiment analysis on 6 million reviews. The final dataset contains information on 50,000 restaurants in 73 different cities.

Conducting Inference

From the results, I can infer that all else being equal, an increase of average star value increases the odds of a restaurant being open by 13%. The correlation between being open and total_stars and total_reviews seems to be non-causal as restaurants that have been open longer will have more reviews.

Interesting Findings

- Nonlinear correlation of stars
- Negative correlation of 3 stars
- Coefficient of polarity is negative



Findings and Conclusion

My findings suggest that the content of yelp reviews offer no insight into predicting a restaurant closure. The baseline for open businesses in the dataset is 0.68. Performing logistic regression and principal component analysis yields a best score of 0.68 on testing data. Cross-vectorization and logistic regression only yield a slightly better score of 0.70.