White Paper on Sources and Methods for the DC Michelin Challenge
Charles Rice
General Assembly Data Science Immersive

Overview

The Michelin *Red Guide* is considered the pre-eminent record of fine and high-quality dining around the world. With the announcement that DC would be receiving entries in the guide this year, we set out to use Data Science to predict the stars.

Sources

The primary source for the data used is reviews from a prominent aggregator of restaurant reviews whose name rhymes with 'whelp.' I did build a scraper, and a very good scraper, but The Site defeated my efforts, with a scraper that returned text that was messy, but erratically so. This meant that no reliable cleaning method could be used. Ultimately, I resorted to Import.io.

Reviews were selected for the existing Michelin restaurants, sorted in ascending order to minimize the effect of already having a Michelin star on the review. Reviews for DC were selected on the basis of 'best match' using The Site's own metric for choice, which seems to be a combination of number of reviews and stars. (538's piece on The Site and Michelin was the primary influence for considering The Site as the main data source.)

Methods

Data was scraped from The Site, both for Michelin Starred restaurants and for non-starred restaurants. For the starred restaurants, I pulled 60 reviews per restaurant. For candidate restaurants in DC, I pulled 20 reviews per restaurant. I realized only on Tuesday that I needed 'noise' restaurants – non-starred restaurants that might still be candidates. I only got about 12 of those, with 20 reviews per.

The data required extensive cleaning and processing, and a lot of hunting for the solutions. Ultimately, I got all reviews consolidated into a single row per restaurant.

These I vectorized using both TFIDF and CountVectorizer, just in case, with character ngrams of en- to tetragrams.

The vectors were run through Naïve Bayes, Random Forest, GradientBoosting and Linear SVM classifiers. Random Forest won, producing the highest F-1 score.

Further Considerations

The question remains whether sentiment analysis would add value. I am also curious as to whether we could get predictive value from the photographs.

Had I this project to do over again, I would gather a lot more data, including 'noise' restaurants. My current model returns far too many 1 star restaurants