Cheng Ji
Data Science Immersive
October 3, 2016

Washington DC Michelin Star Restaurants Prediction

Problem Statement:

Predict which restaurants in Washington DC will get Michelin star in 2016.

Data Collection Source and Method:

The restaurants in the training dataset are scraped from the following websites:

Chicago: http://www.chicagonow.com/chicago-food-snob/2013/12/chicagos-top-85-restaurants/

NYC: https://www.timeout.com/newyork/en_US/paginate?page_number={}&pageId=35907&folder=&zone_id=1202678

San Fransisco: http://projects.sfchronicle.com/2016/top-100-restaurants/

The restaurants list of DC is scraped from the following websites:

https://www.washingtonian.com/2016/02/08/100-very-best-restaurants/

https://www.washingtonpost.com/news/going-out-guide/wp/2016/10/10/these-are-the-d-c-restaurants-that-insiders-predict-will-get-michelin-stars/

If there are Michelin star restaurants that are not in the list, I added them manually into the list.

See ‘restaurants_scraper.ipynb’ for details.

The next step, I used the scraped list to make requests to Yelp API to get information as my features, including price, rating, cuisine category, and review counts. For some restaurants information that cannot be requested from Yelp API, I searched manually and added to my dataset.

See ‘Yelp_API_requests.ipynb’ for details.

Question: Why not just scrape restaurants from Yelp?

Answer: Thus I will have too much data, and most of which will be non Michelin starred restaurants. A highly unbalanced dataset may make it harder to train my models. So I need some sort of pre-filtered list.

Question: Why Yelp?

Answer: There is an empirical work on this problem on fivethirtyeight.com shows that the Yelp data is highly correlated to the Michelin starred restaurants of NYC in 2015:

http://fivethirtyeight.com/features/yelp-and-michelin-have-the-same-taste-in-new-york-restaurants/

I assume that in the other cities in the US, the relationship between Yelp and Michelin star would hold.

Question: Why not use reviews / text data?

Answer: There is no Michelin official reviews on any DC restaurant yet. If I use Michelin official reviews for my training set and reviews from other professional for DC restaurants, I will not have consistent features. The Yelp reviews are less reliable since they are not professional food reviews.

Models:

Four models were built: Random Forest, Gradient Boosting, SVM and Neural Network. Grid search was performed to optimize F1-score. Models are evaluated based on the score on the testing set. The score used to evaluate models was calculated in the way that this competition is set up. The best accuracy score was from neural network, while the best score of competition matrix was from random forest.

The final prediction was determined by a majority vote: only restaurants that are predicted by at least 3 models are included. The number of star was equal to the highest number predicted by any model.

See ‘models.ipynb’ for details.

Consideration and possible next steps:

If more time is given, I would try to scrape the accurate rating from Yelp (from Yelp API, the ratings are rounded to 0.5 level) and try to adjust review counts based on the open time of restaurants.