**Overview:**

The highly sought after Michelin Restaurant Reviews are coming to Washington D.C.! D.C. will be the fourth active city in the United States to have Michelin reviewed restaurants; the others are San Francisco, Chicago and New York City. Restaurants that make the cut can receive one of four ratings: ‘Bib Gourmand’, 1 Star, 2 Star, or 3 Star.

**Objective:**

News of these reviews has created quit a buzz with foodies and non-foodies alike causing many to start putting together their lists of who they think will receive these predacious honors.

I do not have subject matter expertise in fine dining, but by using Data Science I can build a model that can predict if a restaurant is likely to receive a Michelin Star Rating.

**Process Overview:**

By using data from restaurants that have already been Michelin rated I can created a model based on price, cuisine type and reviews to predict what restaurants in Washington D.C. will be likely candidates for Michelin stars.

**Data Acquisition:**

Utilizing the web scraping library ‘Beautiful Soup’, I was able to gather the information of all the Michelin rated restaurants within the United States from the Michelin website. This data was used as the train data to build and refine the model. I choose to only gather information from the restaurants in the U.S. to maintain cultural similarities. (See Appendix 1)

Rather than scraping Yelp, Google Reviews or Zagat to gather data for all the restaurants in the Greater Washington D.C. area, I instead sought out some subject matter expertise. As opposed to open source reviews, which tend to vary highly in terms of quality and consistency. The data by which I will be testing my model on to predict restaurants in D.C. was the ‘Washingtonian list of 100 best restaurants in Washington D.C. for 2016’.

(See Appendix 2)

The main rationally behind using this data was that the structure of these reviews was very similar to the structure of Michelin reviews, especially the text reviews. On the other hand, this does subject me to the bias of the creators of the list as well as the likely possibility that several restaurants that will get Michelin ratings do not appear on this list. This was made apparent when the Bib Gourmand list was released and only half of those were in my test data.

I was not worrying about the possibility of under predicting because of the size of my test data (more than 100 D.C. Restaurants get ratings), because there are only 140 Michelin Star restaurants in the US and the most any one city has is 71 (New York City).

**Cleaning:**

Much of my data was in good form when I scraped it due to consistent values on pages. All the data needed to be converted to numeric values in order to used in model construction. I intended on using three features for my dependent variables; ‘Price’, ‘Cuisine’, and ‘Review’ however, creating numeric values for these caused my number of features to jump to over 16,000.

Price – Price was not listed as a continuous value by either Michelin or The Washingtonian, but instead was either a range or a symbol. Fortunately their scale was the same and could both be converted into values with the range of 1-to-4.

(See Appendix 3)

Cuisine – Some of the restaurants listed by the Washingtonian had two or even three cuisine types, for simplicities sake I decided to only keep one type as Michelin Restaurants only have one type. Dummy features were then created to convert the categorical feature to several quantifiable features.

Review – The Review feature was converted to numeric features using a TFIDF Vectorizer.

After converting the ‘Price’ feature, but before Getting Dummies for the Cuisine feature, I merged my train and test data so that I could TFIDF and Get Dummies all at once so that both the train and test datasets would have the same number of features.

**Feature Selection**

Fit a Random Forests model and then called Feature Importance to get the feature importance percents. None of the features had any significant importance as far as predictability, so I could not just take the 20 or 30 best.

I took two feature importance lists for further testing, one of the top 250 features and another of all features with importance greater than 0 (this was still over 14000). Using selected features produced models that performed much worse than ones that used all the features. I ended up not using any feature selection.

**Model Selection:**

Initially, I planned on using Logistic Regression and Support Vector Machines to predict probabilities and set a threshold on the probabilities to assign a start value (i.e. a likeliness of 95% + would get 3 stars.). The predicted probabilities varied greatly across those with and without stars and the accuracies were terrible to say the least. K-Nearest-Neighbors was performing very well and I decided to further pursue the KNN model.

**Optimization**

Used Grid Search Cross Validation to get optimal parameters for my model. Running GridSearchCV on a KNN with over 16,000 features is highly computationally expensive and not possible within the given time on my personal computer. An Amazon Web Services Instance with 256 Gb of ram and 64 cores was used to run the GridSearchCV (still took over an hour).

**Results:**

Running my KNN with the optimized parameters did not produce ideal outcomes in that it was only predicted 0’s and 1’s with no 2’s or 3’s. Essentially my model does not believe there are any restaurants in D.C. worthy of those ratings. Here are my either predictions

(See Appendix 4)

**Proposed refinements.**

Use a single third party source to gather all my reviews and information that could be used as a dependent variable. This will ensure consistency across writing styles and vocabulary between train and test. This method will only use the scores from Michelin.

Instead of dropping additional cuisine types, the Get Dummies method could have been used in conjunction so a restaurant could have multiple cuisine features.

Try some different classification models.

**Appendix 1**

<https://www.viamichelin.com/web/Restaurants>

**Appendix 2**

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

**Appendix 3**

Before

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Restaurant | Type | Price | Review |
| 0 | ‘Michelin Example’ | Japanese | 'From 50 USD to 74 USD' | ‘This restaurant is the bee’s knees” |
| 1 | ‘Washingtonian Example | Japanese, Sushi | $$$$ | ‘I ate food here. It was good!’ |

After

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Restaurant | Type | Price | Review |
| 0 | ‘Michelin Example’ | Japanese | 3 | ‘This restaurant is the bee’s knees” |
| 1 | ‘Washingtonian Example | Japanese | 4 | ‘I ate food here. It was good!’ |

**Appendix 4**

My classifier only resulted in 1 star predictions. Essentially, these stars are likely to get stars. It’s just uncertain how many.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fiola Mare** | **Masseri** | **Fiola** | **Obelisk** |
| **Preserve** | **Del Campo** | **Woodberry Kitchen** | **Centrolina** |