POTENTIAL 'MICHELINS' AND WHERE TO FIND THEM

GROUP: Spyder
Yining Li, Chenjian Yang, Shenghui Hua

I. Preview

1. Introduction

Bread is the stuff for life! People love food so much that they would spend a decent amount of money for a fine dining. However, despite the fact that Michelin restaurants are the definition for the highest level of fine dining, not many people have Michelin experience. Price might be one of the issues, but some of the Michelin restaurants are really hard to reserve! This prompts our team to find those non-Michelin restaurants with Michelin quality. We believe that there are potential factors that can significantly determine the Michelin standard. Therefore, we use data from Yelp to construct our study by applying data analytic tools.

2. Data Source:

1) Yelp API: Restaurants info in New York, Chicago, San Francisco, and Washington D.C. where most of the U.S Michelin restaurants located (Due to the limit of Yelp API, we were only able to 1000 restaurant per city.)

2) Web Scraping: Wikipedia for Michelin list, and all restaurant reviews via Yelp (100 reviews upper bound for each restaurant, total 350,000+)

3. Preprocessing

1) Cleaning:

- Removed restaurant entries without important variables, such as price and review.
- Removed duplicate entries.
- Merge Michelin entries to dataset, if they were not there before.
- Create dummy features for logical variables. (e.g. 0 for non-Michelin and 1 for Michelin)

2) Final Dataset: 3800+Rows with the following fields:

- **Yelp data**: name, rating, review count, url, reviews, delivery, pickup, reservation
- **Location**: latitude, longitude
- **Price level**(dummy): $-pr1, $$-pr2, $$$-pr3, $$$$-pr4
- **City**(dummy): NY, CHI, WDC, SF
- **Categories**(dummy): French, American, Korean, Thai, Italian, Chinese, Spanish, Japanese, Mexican, Indian, British, Mediterranean, Wine Bars, Dim Sum, Seafood, Steakhouses

II. Exploratory Data Analysis

Before applying machine learning models on everything, we would like to find the rough Yelp profile for Michelin restaurants.

1. Restaurant Distributions

We first visualized restaurant distributions by categories and found out that elegant food such as French, Italian and Sushi has more weights on Michelin characteristics instead of fast food such as pizza and burgers, though the later ones are more popular for the whole dataset. Next we count the number of restaurants by star by city and realize New York has more Michelin one and two-star while San Francisco has more 3-star restaurants. It looks like both categories and cities should be important factors for Michelin standards.

2. Yelp rating vs. Michelin star

We plotted Yelp rating by Michelin star and the result indicated that those two standards agreed with each other. There was a clear increasing trend for the rating when adding star to restaurants. By using weighted average (base on number of restaurant with each star level), we calculated the average Yelp rating for a Michelin restaurant is 4.14/5.

3. Review Analysis

Reviews account for a large proportion of our dataset; we focused on analyzing the differences of the reviews between Michelin and non-Michelin restaurants. First, we conducted Vader and NRC analyses and found that compared with simple sentiments, these two types of restaurants showed significant differences (Michelin: 0.30 vs. non-Michelin: 0.28) in 'compound' sentiment, which is the reflection of multiple emotions. In other words, all restaurants receive both positive and negative reviews without obvious difference. Customer will not tend to give a higher evaluation just because it's a 'Michelin'. However, when combining all the sentiments, evaluation of Michelins is better than non-Michelins.

Second, by applying LSI model to compare the review similarities of Michelins and non-Michelins, it made sense by believing compared with non-Michelins, the topics of Michelin restaurants are more similar with other Michelins. The difference of Similarity Score* is shown as below.

Score values	NY	CHI	WDC	SF	Total
Michelin	2.16	4.22	3.43	3.71	**2.96**
Non-Michelin	-0.19	-0.14	-0.15	-0.20	**-0.17**

* For each Michelin restaurant, we compared its review with the review corpus of all restaurants, and found its Top-100 most similar restaurants list then calculated how many times each restaurant appeared in these lists, and used this scaled frequency as the Similarity Score. Theoretically, a higher score means more likely to be a Michelin.

In view of the significant distinction of Similarity Score between Michelins and non-Michelins, we decided to use it as a criterion to explore non-Michelin restaurants with Michelin quality, i.e., selecting non-Michelin restaurants

with the highest scores as the predictive potential Michelin restaurants. The prediction list is shown in Part IV.

In addition, the result of topic analysis via LDA tells us the similarity and differences between Michelin and non-Michelin review topics:

- Similarity: Positive words like '*Like*', '*Good*', '*Great*' appeared many times in both review corpuses. We conclude that positive reviews account for the majority of reviews on both Michelins and non-Michelins, which also explains why 'unipolar' sentiments like '*positive*' or '*joy*' didn't show significant distinction.
- Differences: For Michelins, words '*Menu*', '*Service*', '*Dishes*', '*Chef*' and '*Experience*' appeared more times, while some specific fast foods names appeared frequently in non-Michelin reviews, like '*Burger*', '*Chicken*' and '*Pizza*'.

Obviously, we can conclude that when customers patronize Michelin restaurants, they focus more on some artistic and quality points, such as dining experience, culinary style and dish design. While for non-Michelin restaurants, customers care more about how food tastes, i.e. the focus is the fundamental thing of a restaurant: the food itself. To some extent, the reviews topics difference of the two types of restaurants caused by the title of 'MICHELIN', which awards a restaurant an 'aureole' of artistry, enjoyment and cuisine professionalism.

III. Classification Models

We aim to classify Michelin vs. non-Michelin restaurants using several machine learning algorithms. Due to the nature of the problem and size of the dataset, we decided to apply random forest, logistic regression, linear regression and factor analysis.

1. Random Forest

To avoid overfitting, we used selectFromModel object from sklearn to automatically select the features with importance greater than the mean importance of all the features by examining only the training set. After fitting the data into the model, we got relatively low recall (33%). However, we did find that the most important features were all sentiment-related.

2. Logistic Regression and Adaptive Synthetic Sampling Approach (ADASYN)

Similarly, we used feature selection object to select the top 10 features. Judging by f-score, the performance was not better than random forest. The false positive rate was close to 0, suggesting a very imbalanced dataset (with roughly 200 Michelin restaurants and 3600 non-Michelins).

In order to fight this problem, we decided to utilize ADASYN to increase the minority Michelin class. The mechanics is as follows: first it finds the n-nearest neighbors in the minority class for each of the samples in the class, then it draws a line between the neighbors and generates random points on the lines, and lastly it adds random small values to the points to make them more scattered. We got our best overall results after applying ADASYN.

3. Ordinary Least Squares Regression (OLS)

Using the ROC curve, we decided the ideal threshold value is around 0.55. The linear model achieved both higher precision and recall than logistic regression without ADASYN. In addition, the regression coefficients gave us interpretive power. For example, for every 0.01 increase in Vader compound sentiment score, the restaurant is 2.5% more likely to be Michelin if we interpreted 0 and 1 as realized probabilities.

4. Factor Analysis with ADASYN

Factor Analysis is a method for data simplification and dimensionality reduction by learning logical structures and information relationships among the raw data. Considering the large amount of initial variables and the potential associations among them, we tried this method to extract some comprehensive factors from the original variables; these new factors can explain restaurants characteristics from certain different dimensions separately. Then we utilized scoring method to evaluate each restaurant's performance on the new factors and regarded the best performed ones as Michelin restaurants. Here we also applied ADASYN to avoid data imbalance.

Finally, we evaluated our models by five critical metrics on the test set, the result of which is illustrated below.

Metrics	Random Forest	Logistic Regression	Logistic w. ADASYN	Linear Regression	Factor Analysis
Precision	95%	77%	**90%**	79%	74%
Recall	33%	50%	**91%**	56%	74%
FPR	0%	1%	**11%**	1%	26%
F-score	64%	63%	**90%**	65%	74%
Accuracy	97%	97%	**90%**	97%	74%

IV. Conclusions

Based on the five models we constructed, we found that overall publicity and customer experience seem to be the differentiating factors for being a Michelin restaurant. Specifically, sentiment scores have higher relative feature importance in random forest model and higher significance in OLS regression than categories, ratings, and price.

In terms of classification algorithms, despite high precision and accuracy, classification models based on Yelp data and sentiment scores have relatively low recall due to imbalanced dataset. However, with ADASYN over-sampling, we were able to strike a better balance in precision (90%) and recall (91%) using logistic regression.

Prediction list of potential Michelins via Similarity score:

NY	**CHI**	**WDC**	**SF**
Traif	Bellemore	Gravitas	Homestead
Boulud Sud	Les Nomades	Corduroy	Chez Panisse
Manhatta	Arbor	1789 Restaurant	Pappo
Henry's End	the Albert	A Rake's Progress	The Wolf
The Eddy	TWO Restaurant	Marcel's by Robert Wiedmaier	Juanita & Maude