tripadvisor

Frameworks & Methods Final Project

04.25.2019

Anderson Nelson
Cass Ernst-Faletto
Mayur Bansal

Table of Content:

Introduction

TripAdvisor an American company that currently holds the world's largest travel site. It is a travel and restaurant website that hosts restaurant and hotel reviews as well as accommodation bookings and forums. We think analyzing TripAdvisor data is interesting as there is a high volume of traffic on the site, approximately 490 million unique monthly visitors. There is currently 730 million reviews and opinions on TripAdvisor with 7.5 million accommodations, restaurants, and attractions. TripAdvisor covers 136,000 destinations with 4.9 million restaurants on its website1.

The dataset we have chosen contains restaurant information for 31 European Cities. It is important to note that this dataset only contains restaurants that are listed in the TripAdvisor database, we do not have a list of restaurants in each city. Each restaurant has information about its ranking in the city, cuisine style, rating, price range, number of reviews and the reviews written by customers.

The goal of the project is to evaluate the data from the restaurant owner standpoint: Trip advisor provides a lot of insights to consumer on how to select places to travel and how to evaluate travel options. For this analysis we wanted to provide data to help restaurants make better decision pricing, location, menu options. With this insight, restaurant owner can pivot their marketing campaign and offerings. We wanted to accomplish this by answering the research questions.

1. Understand the factors that impact restaurant ratings and price
2. Evaluate the data to understand user preferences and rating rationale
3. Evaluate the eating trends of the different cities
4. What impact does healthy options have on customer's rating of restaurants
5. Create a prediction model that classifies the restaurant ratings and prices range

Data Cleaning

We discovered that 3.24% of all the data was missing or NA. When we explored the missing data further, we discovered that only 3 columns contributed to this figure: Ranking, Rating, and the number of reviews.

Ranking: There are 9,651 Null and NA values which represent 0.04% of the entire dataset. The rank value represents the rank of the restaurant among the total number of restaurants in the city, and there are a number of reasons as to why it could be missing:

- The restaurant is new, or the restaurant doesn't get much traffic or
- They regularly receive customers who decline to initiate a rating or a recommendation on Tripadvisor

We can conclude that even if the restaurant had 1 rating, it's popularity would likely be on the lower end, since the factor that propel rankings is number of reviews. As a solution, we categorized the missing values and as 1+ the last ranking.

Rating: There are 9,630 are missing, and 41 values that are categorized as -1. The reason why they are missing is similar to the ranking.

There are a number of ways to address the missing this:

- Replace the missing values with the mean or median
- Replace the missing values based on stratified random sampling of 1-5 ratings in the
- Random numbers between 1-5
- Replace with a static value

The restaurants that do not have a rating also didn't have a ranking. They are either new or no one has reviewed them yet. The best course of action is to replace with 0.

Price Range: is provided in $$-$$$ format and we needed to modify it, in a way that R can recognize for modeling purposes. We converted the $ to numerics, and we also converted to a range a range of (Low, Medium, High).

Cuisine Styles: The cuisine styles were all listed in one column, separated by commas. To use cuisine styles in prediction and to use it for further analysis we had to seperate each cuisine style into individual columns and give these columns a binary outcome. We did so first by getting rid of white space, parentheses, and quotes. Once we were able to get rid of these special characters we used mtabulate from the qdapTools package to split cuisine styles at each comma.

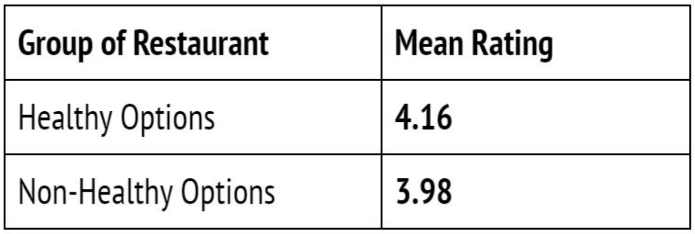
Data Exploration

Healthy Trends

The 2018 Canadian Healthy Eating Consumer Trend Report reported that 29% of individuals are more likely to visit a restaurant that offers some healthy options. Interestingly these individuals don’t always end up ordering a healthy dish [2]. How does this information translate into the rating of restaurants? Are individuals more likely to rate the restaurant higher because it had this healthy options tag?

“A place at the table: vegan fine dining is on the rise across Europe” confirms that Veganism and Vegetarianism is spreading across Europe. Our analysis led to a few interesting insights about how Veganism is becoming a big trend in European Cities[3]. We grouped restaurants into groups with healthy food options and restaurants without these healthy food options. Healthy food options include “Gluten Free Options”,”Vegetarian Friendly”,”Vegan options” and “Healthy” cuisine types.

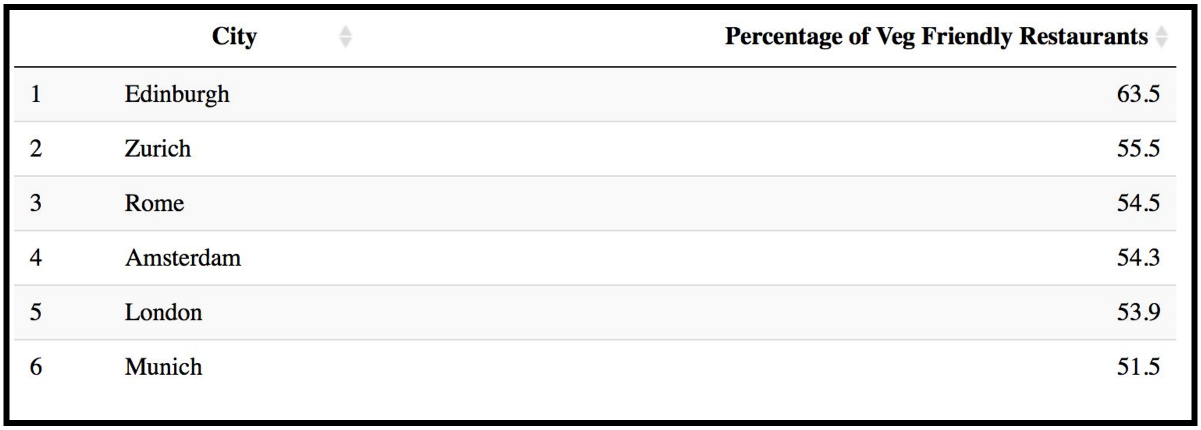
Around 20% of the restaurants listed provide all these healthy options to customers. When evaluating the mean price, we observe that restaurants with more healthy options tend to get a better rating than those which do not provide healthier meals to its customers. Below is the mean rating for restaurants that have the healthy options tags as compared to restaurants that do not have this healthy option tag.



Group of Restaurant	Mean Rating
Healthy Options	**4.16**
Non-Healthy Options	**3.98**

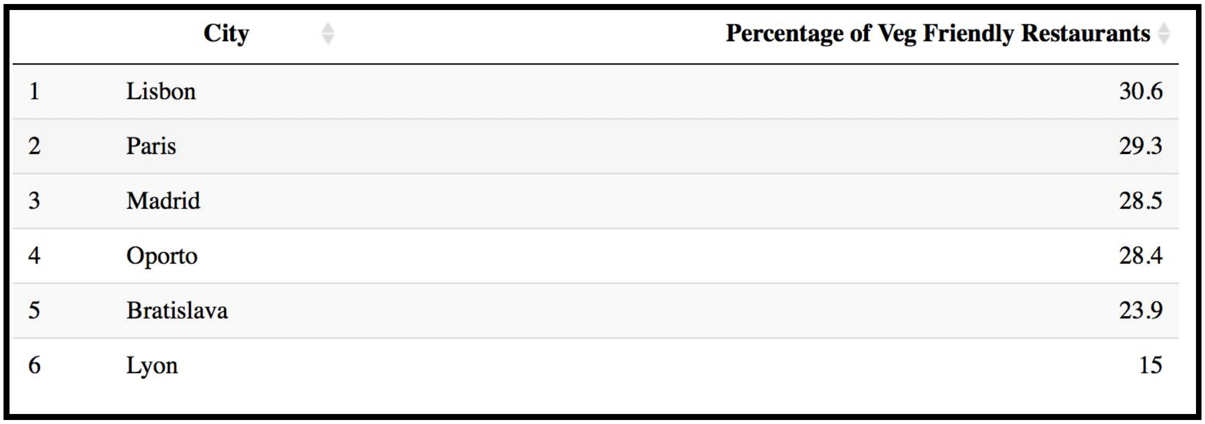
We conducted an analysis to determine if the findings was significant, and p-value of the difference is 0.3. While it’s not significant it’s an indication of how restaurant owner might want to think about marketing their restaurants.

Cities with most Vegetarian Friendly Restaurants



	City	Percentage of Veg Friendly Restaurants
1	Edinburgh	63.5
2	Zurich	55.5
3	Rome	54.5
4	Amsterdam	54.3
5	London	53.9
6	Munich	51.5

Cities with least Vegetarian Friendly Restaurants



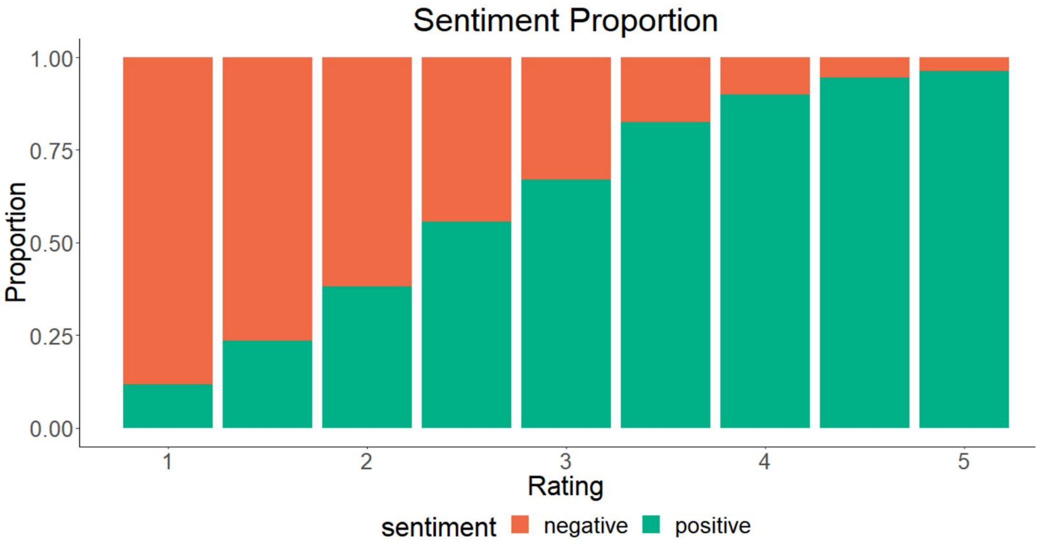
	City	Percentage of Veg Friendly Restaurants
1	Lisbon	30.6
2	Paris	29.3
3	Madrid	28.5
4	Oporto	28.4
5	Bratislava	23.9
6	Lyon	15

Interpretation:

London, Rome, Amsterdam feature in the list of cities with maximum proportion of Vegetarian Friendly restaurants whereas few cities in France (Lyon, Paris) and cities like Madrid have the least representation of Vegetarian cuisine options. furthermore, we found that this under-representation of Vegetarian friendly options in Madrid has a significant impact on the restaurant rating as 81% of restaurants offering vegetarian options to customers have received good ratings (4 or above) whereas only 58% of restaurants not offering vegetarian options receive good ratings. So as an restaurant owner who wants to open or market his restaurant, I would strongly look into this particular trend of Veganism in Europe. Similar observations are seen in the French cities.

Text Analysis

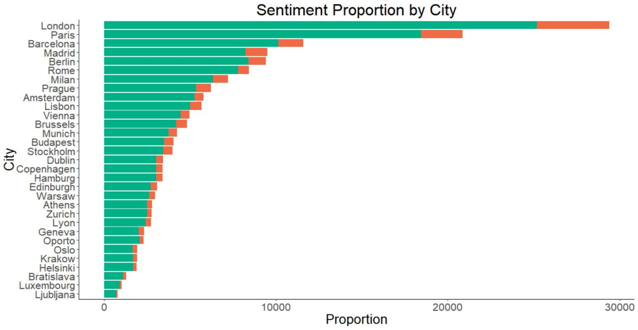
Before going further with our analysis, we wanted to gain insights into the reviews and how these are correlated with ratings. Specifically, are longer reviews correlated with better ratings? The correlation between review ratings and the number of characters is 0.031. This highlights a very small positive correlation between number of characters and ratings. Then we dove into number of words in each review and review rating. This correlation was also positive but smaller than the one discussed above, specifically, it is 0.014. Lastly, we

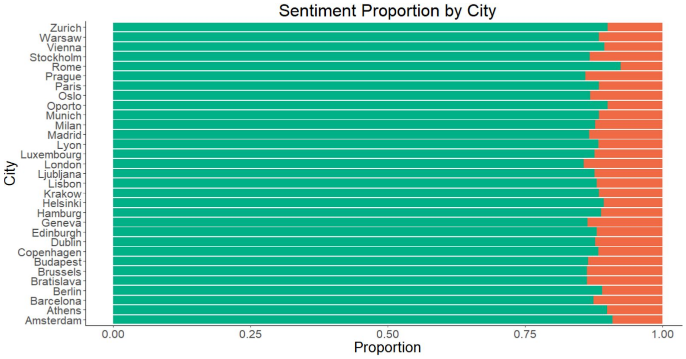


looked at the correlation between number of sentences and rating. This was the highest correlation found, of 0.037. All of these correlations between review length and rating show that the longer the review the higher rated it is. Nevertheless, the correlation is very small.

Next, we analyzed review sentiments based on rating. This step in data exploration is to confirm our assumption that higher ratings would have less negative words and vice versa. This graph highlights that our assumption is correct. The lower the rating the more negative words it has. We were able to get this information with the bing lexicon.

We also wanted to look at sentiment by city to see if some cities have a greater number of positive or negative sentiments as compared to other cities.





The graph on the left depicts negative and positive sentiments from reviews by cities. Though there is information about review sentiments it is very hard to compare all the cities since they differ in the number of reviews they have. For example, London seems to have the most negative words used in reviews but this needs to be compared to the amount of reviews since it is also the city with the most reviews.

The graph on the right is more comparable and shows that approximately 80% of all reviews have words with pos

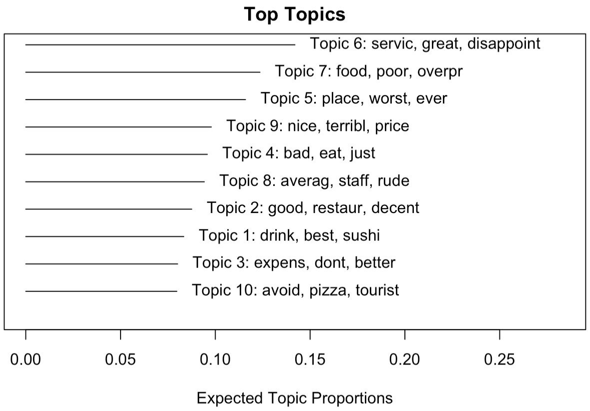
Topic Modeling

In order to extract the hidden semantics from a broad range of customer reviews, we decided to use Latent Dirichlet Algorithm (LDA) as a method to execute topic modeling. The idea was to not just look at the words in terms of the bag of words approach but also notice the probability distribution of words and get a range of meaningful topics for sentiment analysis.

To gain insightful topics, we separated our dataset into bad reviews and good reviews. This would be helpful in separating the topics which lead to poor and good ratings for a restaurant. Bad reviews correspond to restaurants having ratings less than 3 and good reviews are the restaurants rated above 4. The aim was to get certain set of topics which could be used to summarize the reasons which lead the customer to give a specific rating to the restaurant.

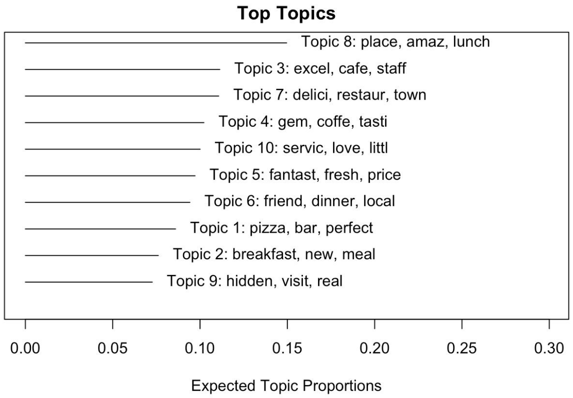
Following are the results from the LDA Analysis:

For Bad Reviews (Ratings<3), interpretations:



We find that negative words associated with staff, service, food and price are common in the identified topics. For example rude staff, disappointing service, overpriced food etc are some frequently occuring topics in bad reviews as expected. We also come across topics containing words like tourists and avoid which is interesting to find and can be helpful for restaurant owners to look for these topics which annoy customers.

For Good Reviews (Ratings >4), interpretations:



Similarly, we get expected words in our topics from positive reviews such as fresh, delicious, excellent staff etc. Additionally, we manage to get some really interesting topics containing factors like hidden, little, new, local which tell a lot about how customers look for these attributes while rating a restaurant. On further analysis, we find that “Hidden Gem”,”Local food”,”New spot in town”, “little and cosy cafe” are some sentences which often get repeated in positive reviews.

Recommendations from the results of Topic Modeling

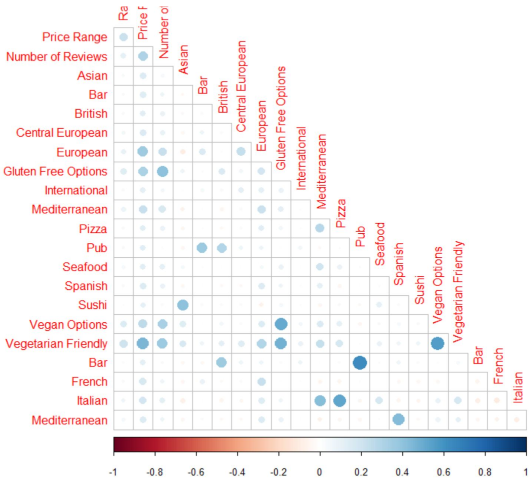
From our analysis of text reviews, we notice that predicting ratings from reviews is a confounding task as people have different ways to rate a restaurant and their interpretation of a good rating could be 3 or 4 depending on how they feel about it. To help restaurants identify their pros and cons, an automated rating predictor based on the reviews they provide would give a more accurate point of view to the restaurant owners. Instead of having customers write a review, we can give users an option to just select the topic which closely relates to their dining experience. The automated predictor would just give the rating based on the topics selected by the customer. This would make Ratings more effective and accurate to analyze.

Predictive Analytics:

a) Price Range:

We wanted to understand the factor that impacts price range. The results of this model would be used to help restaurants decide how to price their restaurant. Using the clean dataset, we wanted to create a prediction model that categorized the price range, and we used 246 variables to perform the prediction and interpret the result. We decided to use two models, random forest and regression. The rationale is that we wanted to focus on models that are interpretable.

Step1: Understand Correlation



We wanted to understand the relationship that exists among the variables in the data set, the best method is to conduct a correlation analysis of all the variables. Using all 246 variables produces little correlation among the variables. We filter the data set all the correlation with a value greater than .1. We were left with a dataset of 22 variables. Calculating the correlation of the to top variables produce results that are more insightful.

Observation: We saw correlation between Bar and pub, vegetarian and vegan, Italian and pizza, sushi and Asian, Mediterranean and Spanish. This indicates that there some overlap between the types of cuisines in the data

Step 2: Create a baseline model

We wanted to create a model that we can use to measure the performance of the models against. Using the 22 values outlined in step 1, we created a baseline model using a train set with TBD variables, we created a multinomial regression model and achieved a prediction accuracy of 71%. We wanted to created the best model under the circumstances in order to evaluate the results.

Step 3: Model building

We created two types of models: multinomial regression and random forests models using 3 sample sizes: 1,000, 5,000, and 10,000 and evaluated the results. We chose small sample size, given the mass amount of data that we created a larger model would have been computationally expensive.

Random Forest: We tuned each model, and the best model selected the best parameters. In our control we used cross validation with 10 samples 3 times, those operations were performed in a random search grid. We found that is better than grid search because it uses random combinations of the hyperparameters are used to find the best solution for the built model. It tries. random combinations of a range of values.

To optimise with random search, the function is evaluated at some number of random configurations in the parameter space[4].

```
control <- trainControl(method = 'cv', number = 3, search = 'random')
```

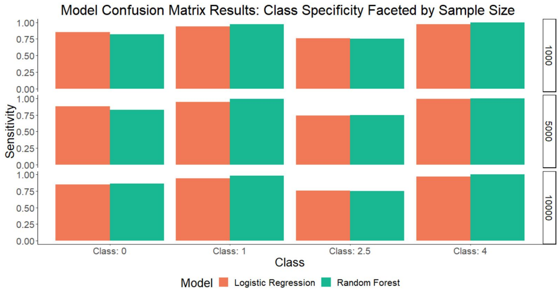
The random forest algorithm has 3 parameters to tune:
Mtry: Number of variables available for splitting at each tree node.
split rule : allows to consider extra trees as splitting rule
Min.node size: Depth of your trees

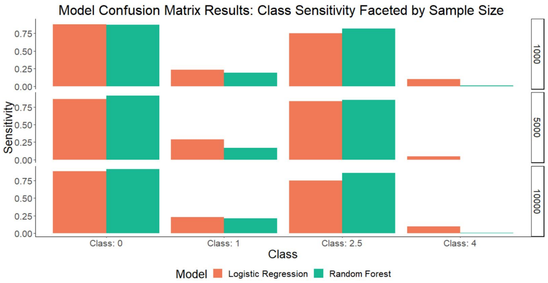
```
grid <- expand.grid(mtry = c(20,80,160), splitrule = 'gini',
                    min.node.size = 15)
```

We chose the parameters to balance for time and accuracy. It wasn't important to generate the most accurate model, but to generate a good enough model for the task , and the model was The model was evaluated on accuracy.

Step 4: Results Analysis and conclusion:

From the model the factors that impact price range are: Cities, Ranking, Ratings, Number of reviews. There are few select cuisines that are significant: American,Asian,Bar,Brew Pub, British, Central European, European, Fast Food, French,Gluten Free Options,Halal,Healthy,Korean and Mediterranean. We also evaluated the results of the confusion matrix to understand the how well the performed in predicting the multiple classes. The model performed poorly in correctly classifying the highest rated restaurants.





Step 6: Reduce the number of Parameters using PCA

Principal Component Analysis:
Since we have 256 variables in the model, we wanted to reduce the number of variables by feature face. The benefits are you have fewer relationships between variables to consider and you are less likely to overfit your model. There are many ways to achieve dimensionality reduction. However, by performing PCA we would lose the ability to interpret the variables. For the purposes of the analysis we wanted to understand what percentage of the variation could be explained by one variable. The model indicated the best dimension would only explain less than two percent of the variable of price range. To have flexibility in pricing it's necessary to have multiple cuisine type.

b) Ratings:

Another predictive model that we wanted to focus was to be able to predict Ratings based on the available variables. As ratings are closely associated with the cuisine types, it was important to understand the effect of specific cuisine types on the overall rating. In order to do that, we analyzed the cuisine types based on their occurences. Rather than going with all the available cuisine types, we just used the 50 most occuring cuisines in our Multinomial Logistic Regression model. Then we tried to use the 10 most occuring cuisines and noticed that accuracy does not improve after adding more cuisines to the predictive model after a point of time. To gain more clarity, we decided to use feature selection methods like Lasso regression which gave a sense of most important cuisines but the results were not helpful in improving the accuracy.

Using Multinomial Logistic regression model, we achieved an accuracy of 48% while predicting ratings based on the top 10 cuisine types (based on their occurences), Price Range, Ranking, length of reviews, Number of reviews, City and presence of Healthy options in the restaurant. Going further, we want to focus more on extracting sentiments out of reviews so that in future a really good predictive model is built which truly captures the sentiments of the people giving the rating to a particular restaurant.

Recommendation System

We wanted to recommend restaurants to customers based on their past reviews. However, out of the 125,527 reviews, only 201 reviewers did more than one review. With this dataset, we were not able to do user-based recommendation. We then decided to build a recommender system based on the popular method. The problem with this recommender system is that it recommends McDonald and Burger King for everyone as these were two restaurants with more than 1 reviews in the dataset.

Conclusions:

Restaurant owners must consider a variety of factors when launching or managing a restaurant.

1. As it relates to pricing the factors that a restaurant should consider are the type of cuisine, specifically recommending to explore recipes relating to: American,Asian,Bar,Brew Pub, British, Central European, European, Fast Food, French,Gluten Free Options,Halal,Healthy,Korean and Mediterranean
2. Europeans are trending towards healthier restaurant options, and there's an opportunity to build healthy conscious restaurants in Lisbon, Paris, Madrid, Oporto, Bratislava and Lyon.
3. Digging deeper into Topic Modeling using Latent Dirichlet Algorithm, we found that topics like “Hidden Gem”, “Little and cosy space” actually make a restaurant more likely to receive a thumbs up from customers whereas topics such as “Tourist trap” and “Rude staff” can impact the reputation of restaurant drastically.
4. After using predictive analytics to predict ratings, we realized that analyzing the current system of collecting reviews hampers the analysis as different people have different ways to express and review the restaurants. Therefore, having an automated review predictor would make more valuable contributions in building a better predictive model as the reviews would capture the true sense of how people feel of that particular restaurant.

References

1. http://ir.tripadvisor.com/static-files/6d4c71fd-3310-48c4-b4c5-d5ec04e69d5d
2. .https://www.technomic.com/newsroom/consumers-growing-interest-health-trends-opens-door-many-new-opportunities
3. https://www.europeanceo.com/lifestyle/a-place-at-the-table-vegan-fine-dining-is-on-the-rise-across-europe/
4. https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318