GreenEats: Navigating Top Vegan & Veggie Cuisine

STAT 628 Data Science Practicum, Module 3

Group 13: Kanishk Saxena, Shan Leng, Sreeja Kodati

1. Introduction

We have found that vegetarians always have a difficult time finding an appropriate restaurant or looking for popular dishes when eating outside. Taking Philadelphia as a sample city, we are curious to explore those popular and successful restaurants with vegan or vegetarian options. The goal of our exploration is two fold. On the one hand, we would like to summarize the key attributes and common features contributing to the popularity and success of these restaurants; and provide valuable information to help new business owners strategically position themselves in the competitive market of Philadelphia. On the other hand, we hope to extract the popular dishes from the reviews of these restaurants and recommend them to vegetarian customers.

2. Data Pre-Processing

2.1 Target Businesses

To obtain restaurants with vegetarian options, according to Yelp's API guidelines, we filter businesses by the "`categories`" column and only keep those with both the keywords "`Restaurants`" and "`Vegan/Vegetarian`", which leads to 1,629 businesses. To choose an ideal subset of restaurants for the following analysis, we group these businesses by "`state`" and "`city`" respectively.

Considering both sample size and geographical scope, we decide to select Philadelphia as our target city. Next, we exclude businesses with no more than ten reviews using the "`review_count`" column to ensure a reasonable sample size. Finally, we end up with 245 restaurants with vegetarian options and 47,991 reviews for these restaurants.

2.2 Data Processing for attribute summarization

With "`Vegan/Vegetarian`" restaurants in Philadelphia, an intuitive idea is to generate a list of business attributes these restaurants have as a part of their listing on Yelp. The corresponding data processing steps involve filtering the city's restaurants based on our definition of a successful business, which is a successful restaurant or business on Yelp that has more than ten reviews and a rating of 3.5 and above.

2.3 Data Merging/Processing

In order to learn the behavior of ratings, we worked on a linear model to predict the stars or ratings based on the trips made to a specific place (from Trips by Distance dataset) and on the mean income of people at a specific place (from the public US Census dataset) [`stars ~ mean_income + trips`]. We create 3 data frames for each of the datasets. For '`business.json`', we create a dataframe with columns of states and the ratings. For '`trips_by_distance`', we create a dataframe with columns of states and numbers of trips. For the US Census dataset, we create a dataframe with columns of the states with mean income.

We then merge all the three data frames based on the common states, which leads to a `merged_df` containing information on trips, median income, and ratings for the states that exist in all three dictionaries. The `merged_df` is used to create predictors (`X`) and the target variable (`y`) for regression. `X` contains the predictors, which are '`Trips`' and '`MeanIncome`'and y is the target variable — '`Ratings`'. The code uses statsmodels to fit an Ordinary Least Squares (OLS) regression model to the data. The `line model = sm.OLS(y, X).fit()` fits the linear regression model to predict the 'Ratings' based on 'Trips' and '`MeanIncome`'. This model finds the best linear relationship that minimizes the difference between the actual '`Ratings`' (y) and the predicted values based on '`Trips`' and '`MeanIncome`'(X). These predicted

values for the target variable are generated by applying the fitted model to the same data it was trained on, resulting in predicted ratings for each state based on their respective number of trips and mean income.

3. Discovery of Popular Vegetarian Dishes

We would like to use the vast review content from Yelp's businesses to detect popular vegetarian dishes served by those restaurants obtained in Section 2. By "popular vegetarian items", we mean menu items that are repeatedly mentioned a lot by customers and have an average star score of 4 or higher from reviews associated with them. We split our analysis into two steps: information extraction and summarization. First, we extract all mentions of potential menu items from businesses' reviews; and second, we summarize these mentions to obtain popular dishes to be suggested to new business owners and recommended to vegetarian customers.

3.1 Extraction

Recognizing menu items in unstructured text is very similar to the Named Entity Recognition (NER) process. In our analysis, we use "`flair`", a powerful natural language processing library in python, to extract potential menu items. We first apply a pre-trained ontonotes model, which can classify 18 different types of entities, to the reviews of vegetarian restaurants and extract entities tagged as "product" from the outputs, as shown in Figure 1.

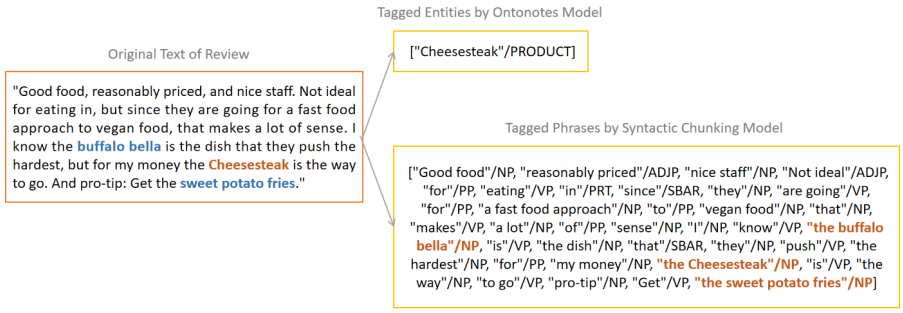

Figure 1. Example of NER with Ontonotes Model and Syntactic Chunking Model

It can be seen that the ontonotes model successfully detected "Cheesesteak" as a product, but it failed to identify "buffalo bella" or "sweet potato fries", which are also menu items. The reason could be that compared to the word "Cheesesteak", which is very likely to be a proper noun with an initial capital letter, the other two phrases seem to be too common to be identified as a "product". To address this issue, we induce another model for chunking noun phrases denoted as "NP" in Figure 1. To further obtain meaningful dishes, we filter the noun phrases with a vegetarian food list[1] containing 330 high-frequency common words, and retain only those phrases with overlapping parts with the vegetarian vocabulary.

We also conduct post-processing with respect to the detection results:

- Remove three types of words: Stopwords (e.g., "the", "a", etc.); General words (e.g., "dish", "brunch", etc.); Non-vegan items (e.g., "fried chicken", "Cheesesteak", etc.).
- Standardize the expressions (e.g., "mac & cheese" and "Mac n cheese" are unified as "Mac&Cheese").
- Only keep products and noun phrases that have appeared more than 10 and 20 times.

The above procedure leads to 82 products and 395 noun phrases for vegetarian dishes in the end.

3.2 Summarization

With Yelp's API guidelines, we use the "`stars`" of a review as the star rating for all the vegetarian dishes extracted from the corresponding text content. For each vegetarian restaurant, we calculate the average star ratings (ave_star) for every vegetarian dish served by the restaurant and retain the dishes with an

[1] We have manually created this list based on information provided by this website.

average of at least 4 to be displayed as recommendations in our Shiny-app. Next, We extract vegetarian dishes served by more than 10 restaurants and consider them as the most common ones. By further filtering these common dishes to have an average star rating of at least 4.5, we obtain 48 “star dishes” that are popular and common for both business owners and vegetarian customers.

3.3 Key Findings About “Star Vegetarian Dishes”

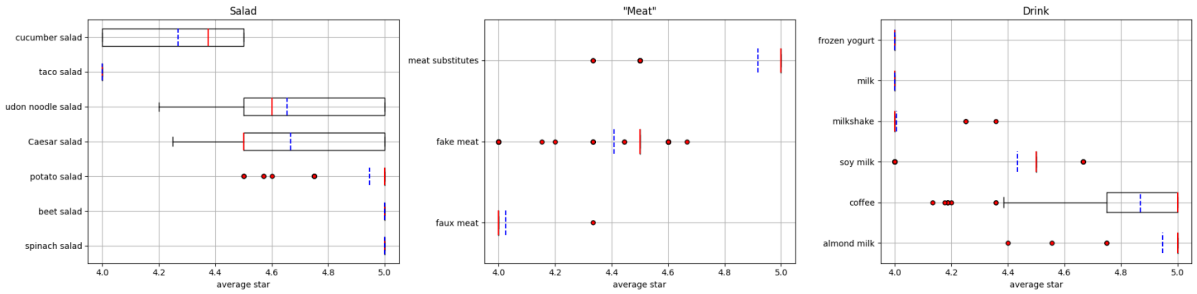

Figure 2. Boxplots of Average Star Ratings across Restaurants

- Salad. Compared with the very broad “salad” dish (ave_star = 4.40, same below), more specific ones such as spinach(5.00) and potato(4.95) salads are more popular, while cucumber(4.27) and taco(4.00) salads are less favored.
- “Meat”. Meat substitutes(4.92) seem to be easier to be accepted by vegetarians and can be considered by new business owners. Fake(4.41) and faux(4.03) meat are less popular.
- Drink. Low-lactose drinks like almond milk(4.95) and coffee(4.87) are more popular than milk(4.00) and milk derivatives.

New business owners may utilize the above information together with Figure 2 to develop menus. Besides, burrito(5.00), spring roll(4.97) and bagel(4.96) are safe items to be served since they all have a very high average star rating across restaurants. Also, business owners need to be careful with seasoner as spices(4.07), sauce(4.06) and ginger(4.00) are not very well-received.

4. Key Findings from Attributes (for successful businesses)

The main intention of the analysis is to find out if there were any common attributes among restaurants with ratings of 3.5 and above in the city of Philadelphia. Only attributes that were mentioned at least five times overall were considered.

Based on the common business attributes listed for a successful restaurant, some generalizations about the characteristics that contribute to its success can be made. For example, having “RestaurantTakeOut: True” as a common attribute potentially conveys that successful restaurants often prioritize convenience by offering services such as takeout and delivery.

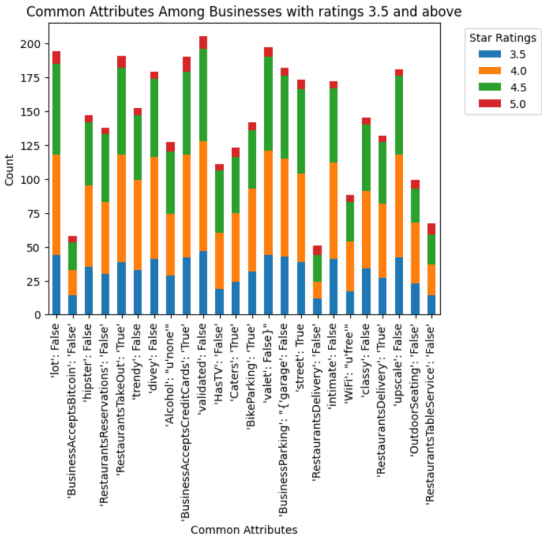

Figure 3. Barchart plot of common attributes among successful businesses.

Each of the common attributes from the barchart plot can be interpreted like the above to make generalizations that highlight the importance of understanding the preferences and lifestyles of the target customer base. Ultimately, the success of a restaurant is a multifaceted outcome that results from a combination of factors, including the quality of food, service, atmosphere, and effective business strategies.

5. Key Findings from the other 2 datasets

After our extensive analysis we find that the values we get provide insight into the performance of the regression model. An R-squared value of 0.2382 indicates that the model explains a moderate portion of the variance in the ratings based on '`Trips`' and '`MeanIncome`' and there may be other factors influencing the ratings that are not captured by these variables. MSE measures the average squared difference between the predicted values and the actual values. A lower MSE indicates that the model's predictions are closer to the actual values. In this case, an MSE of 0.3285 means that, on average, the squared difference between the predicted and actual ratings is 0.3285. Lower MSE values are preferable as they indicate a better fit of the model to the data. Even when we plotted the predicted ratings v.s. actual ratings graph we see the values to be around the first degree polynomial line, as shown in Figure 4.

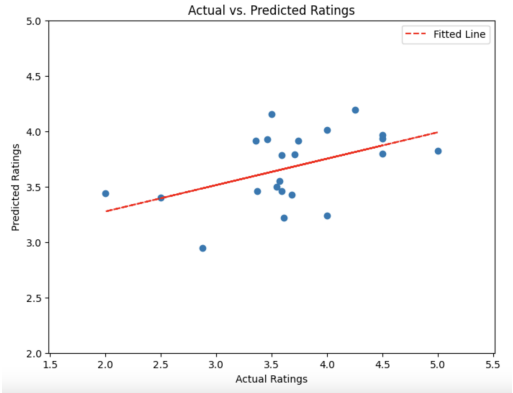

Figure 4. Ratings: Predicted v.s. Actual

6. Discussion and Conclusion

6.1 Discussion

For our NER part, we have managed to get a meaningful list of vegetarian dishes with the library "`flair`". But this process requires a lot of manual screening, which could be avoided with more advanced neural network NER models. However, due to a lack of labeled training data (i.e., reviews and tagged dishes), we were not able to apply neural networks for our case.

As far as future developments of the project are concerned, there are multiple ways of improving the quality of the designed shiny application. The application is currently intended to display the results using postal code to narrow the search. Having a filter by categorizing the city based on street information would be extremely helpful. A price range defined for each restaurant could be included as well. However, due to lack of detailed information about the restaurant's menu options, price range and other things, we could not implement the features in the application.

6.2 Conclusion

The main aim of our project is to help people explore the finest vegetarian and vegan restaurants in Philadelphia. Using customer reviews from Yelp, we have identified popular menu options for each restaurant in the area. The ultimate output of the project was a user-friendly application that serves as a tool for individuals seeking to explore good restaurants in the area. Users can filter restaurants based on postal code and click on markers through an interactive map interface to access detailed information and explore popular options. The application extends its utility to new business owners, offering valuable insights to make informed decisions and strategically position their businesses to resonate with the preferences of their target audience in Philadelphia.

Contributions and References:

Contributions	Kanishk Saxena	Shan Leng	Sreeja Kodati
Presentation 1	Responsible for slides 7-12 (Further Analysis and Results). Reviewed and provided feedback on all the slides	Responsible for slides 2-4. Reviewed and provided feedback on all the slides.	Responsible for slides 5-7. Reviewed and provided feedback on all the slides
Presentation 2	Responsible for slides 10-11 (talking about the other 2 datasets). Reviewed and gave feedback to all the slides	Responsible for slides 2-7. Reviewed and provided feedback on all the slides.	Responsible for slides 8,9 and 12. Reviewed and provided feedback on all the slides.
Summary	Responsible for summarizing data processing of all 3 datasets in order to use them in our analysis for our final linear model. Also adding technical details of the summary of our final model. Reviewing the other sections of the executive summary.	Responsible for Introduction, Data Preprocessing 2.1 and Discovery of Popular Vegetarian Dishes. Reviewed/edited the discussion and conclusion.	Responsible for summarizing the attributes and shiny app. Reviewed/edited the introduction, data cleaning, discussion and conclusion.
Code	Responsible for all the code to process all the 3 datasets, cleaning all the 3 datasets, extracting relevant information from all 3 datasets, seeing the common columns and values in all 3 datasets, trying to use all 3 datasets for our final analysis for the linear model.	Responsible for NER code and code to replicate Figures 1 and 2.	Responsible for summarizing the business attributes code. Generated a bunch of visualizations of the attributes grouped by different categories.
Shiny App	Reviewed and provided feedback on Shiny app	Reviewed and provided feedback on Shiny app	Worked on the Shiny app.