Food Happens in Vegas:
How can restaurants improve their Yelp profiles for success?

Vegas Foodies

Elizabeth Combs (eac721)

Anu-Ujin Gerelt-Od (ago265)

Wendy Hou (wh916)

Emmy Phung (mtp363)

Introduction to Data Science

December 9, 2019

Abstract

The purpose of this data mining project is to examine how restaurants can improve their Yelp profile to become more "successful" on Yelp in Las Vegas, Nevada. Differently from the traditional approaches to this dataset, our methodology defines "success" as a binary variable through an exploratory analysis of the restaurants' review counts and ratings on Yelp. Feature variables include categories and attributes that Yelp users can use to select which restaurant to visit. For this project, we ran Decision Tree, Random Forest, and Logistic Regression to explore key features associated with "success" and obtain recommendations for restaurants to improve their Yelp profile. Final results indicate that determinants of success vary by cuisine type.

I. Business Understanding

A. Business Question

Food goes beyond survival necessity and contains both the historical and cultural values of a region as well as the current trends of the region. Therefore, for a restaurant to be successful, it is crucial for the business to understand its adaptability and strategic fit with the region. To succeed in today's competitive scene, a restaurant is almost required to have an online profile. Studies have shown that, as user-generated content is becoming the norm, online reviews that are available on various business directory platforms not only provide users with great references about a product/service of interest but also help businesses grow and generate revenue (*Chang*). Yelp, as one of the most prominent applications in this market, uses the wisdom of the crowd to source its inventory and recommend relevant restaurants, events, or services to its users. According to *Yelp Newsroom*, there are 38 million unique mobile app users and 91.3 million unique web visitors every month. Having such a huge audience base, Yelp plays an important role in promoting business for a restaurant on its platform. In fact, Yelp usually attracts around 2.6 million diners for restaurants through its platform every month (*Fast Facts*). As a result, one can gain important insight into the success and popularity of restaurants in comparison to their competitors by exploring the Yelp dataset.

B. Data Mining Solution

In this project, we decided to utilize Yelp's dataset available on Kaggle to aggregate information on restaurants in Las Vegas and build a "success-formula" based on features we identified as important. Compared to other projects using the same dataset, ours creatively explores from a business-owner's perspective, recommending features for businesses to help them become more appealing and improve their performance. The project selects a target variable that includes two aspects that conventionally represent a restaurant's success: star rating and total number of reviews. By integrating these two criteria, the project uses the response variable to represent the success of a restaurant on Yelp. After evaluating the restaurants' overall success, the project identifies the top performers in Las Vegas and the most important features that correlate with the restaurant's success.

II. Data Understanding

A. Data Source and Sampling

Yelp provides a business dataset where each instance is a business's Yelp account. The data contains identification (ID, Name, Location, Hours of

Operation), ratings (Stars and Review Counts), and features (Categories and Attributes) of restaurants. Yelp also has data on reviews, users, and more that go beyond the scope of this data mining problem. In total, the Yelp dataset contains 188,593 instances of businesses in mostly Europe and North America with 14 features describing them (*Appendix 1*). To address our business problem, we selected a subset of our dataset using the 'Categories' feature to examine only restaurants with the keywords "Restaurant" and "Food". We have selected Las Vegas for this project since it is one of the cities with the most instances (4,064 restaurants).

B. Potential Bias

Our dataset has some inherent bias given that it only contains restaurants with a Yelp profile, excluding those that have less online presence. Additionally, Yelp has provided limited information on the dataset's latest update and thus, metrics may vary over time. For example, the study may be influenced by the exclusion of recently closed or newly opened restaurants. It is also important to remember that profiles are self-reported and are often used to drive traffic, so restaurants may add irrelevant features that are popular in searches. At the same time, there is a potential survey bias within the data since Yelp relies on user engagement. One example of this is within the star ratings where few restaurants have ratings below three-stars. (See *Exploratory Data Analysis* for the treatment of this data.)

C. Data Accuracy

The dataset has a high degree of missing values since the data are mostly self-reported by businesses. For instance, only a small number of restaurants report having 'Alcohol' in their Yelp profile. However, we suspect that more restaurants serve alcohol than what was expressed in this dataset. We chose to view the attributes that are labeled "Not Reported" as a "False" since users searching on Yelp online would not find these restaurants using these attributes. As a result, alcohol should not contribute to that restaurant's "success". The treatment of these missing values is discussed further in the *Data Preparation* section below. Finally, we have avoided data leakage by leaving out any information on review counts and stars, which went into the creation of our binary "success" variable. Since the other features were derived from the categorization and attributes of the restaurants, there should be no other proxy variables for our "success" metric.

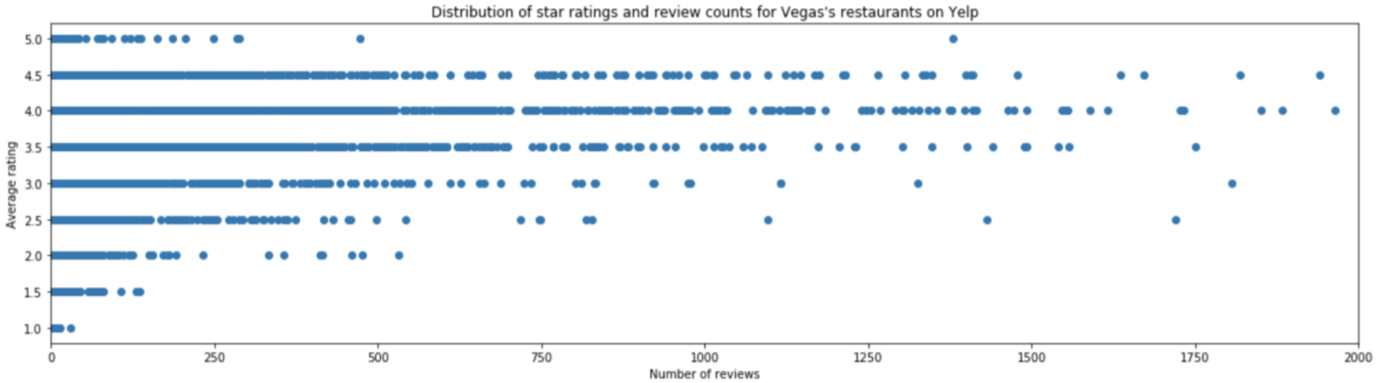

Figure 1. Current distribution of star ratings and review counts for restaurants in Las Vegas.

D. Exploratory Data Analysis

We started by looking at the distribution of star ratings of all the restaurants in our sample data, which was Yelp's default indicator of success. The majority of our restaurants were rated between 3.0 and 4.5 (*Figure 1*). Furthermore, a lot of highly rated restaurants had very few reviews (fewer than 10). The number of reviews also gave a trustworthy signal of whether that restaurant was popular. Opposite to star rating, we observed too much variance in the variable 'Review_Count'. Top-performing restaurants could have thousands of reviews while others had about 200+ on average. Furthermore, we acknowledged that popularity does not necessarily imply success since there could be both negative and positive reviews. Therefore, we chose to combine the two aspects to define success for a restaurant, which is different from other projects that use the same Kaggle dataset. Plotting these variables together, we were able to detect our successful instances (*Figure 1*), which were marked by a high number of reviews and ratings. Details for the specific thresholds and methods used to construct our target variable "success" will be discussed in *Data Preparation*.

Next, the team examined the attributes that could potentially contribute to these restaurants' success. After reviewing all the feature variables given by Yelp's dataset, we decided to extract data from the variables 'Categories' and 'Attributes' (*Appendix 1*). From the sample of Las Vegas restaurants, we observed 387 distinct categories, including food types ('Dessert', 'Barbeque', 'Pizza') and cuisine types ('Thai', 'American', 'Mexican'), and 81 attributes ('Parking', 'HasTV', 'Alcohol'). *Figure 2* shows the distribution of restaurants by cuisine type. We had a relatively even number of restaurants in each region, except for North America, but the ratings and reviews were distributed unevenly among groups. Therefore, we decided to compare restaurants within the same groups and specify a success formula for each cuisine.

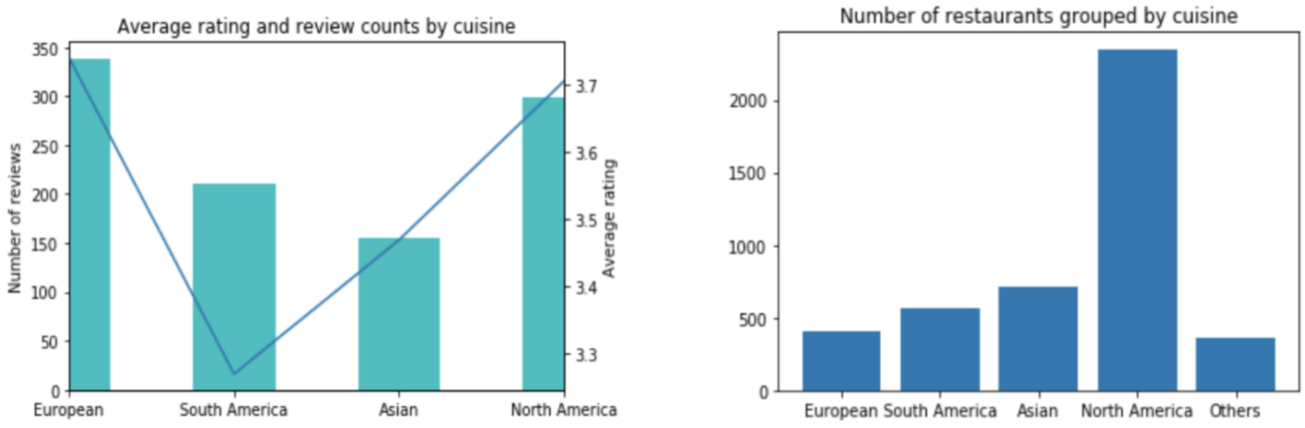

Figure 2. Descriptive statistics by cuisine type.

III. Data Preparation

A. Data Cleaning & Feature Engineering

As mentioned above, the team extracted data from 'Categories' and 'Attributes' which were not in a clean, tabular format within the data file. The team generated hundreds of binary variables (dummy indicators) to account for each category and attribute being present within a restaurant. This allowed us to identify missing attributes in restaurants and were reviewed for sparseness and applicability. Attributes with more than 50% null values were excluded as well as features that were observed in less than 1% of the restaurants because they provide limited marginal value for modeling. We also used a naive approach to exclude attributes such as 'HairSpecializesIn_Coloring', which we suspected to be irrelevant to restaurants. Eventually, we ended up with 41 attributes to be considered for this project.

B. Target Variable & Categorization

After multiple trials, we assigned a binary variable for "successful" restaurants with ratings greater than or equal to 4.5 stars and review counts greater than or equal to 100. These numbers were chosen because they generated a reasonable number of successful instances (331 instances, 8.76%) and fit the intuitive definition of success for a restaurant. However, our dataset is negatively skewed due to these thresholds and we will talk about the effect it has on our models in the *Modeling & Evaluation* section.

We found a middle ground between spotting differences and overfitting by cuisine: we aggregated cuisines by the top 100 most frequently occurring categories by continent: North America, South and Central America, Europe, Asia (*Appendix 2*). We classified the top 100 words which categorize 90% of the data into the first four categories, and other words were classified into an 'Others' category.

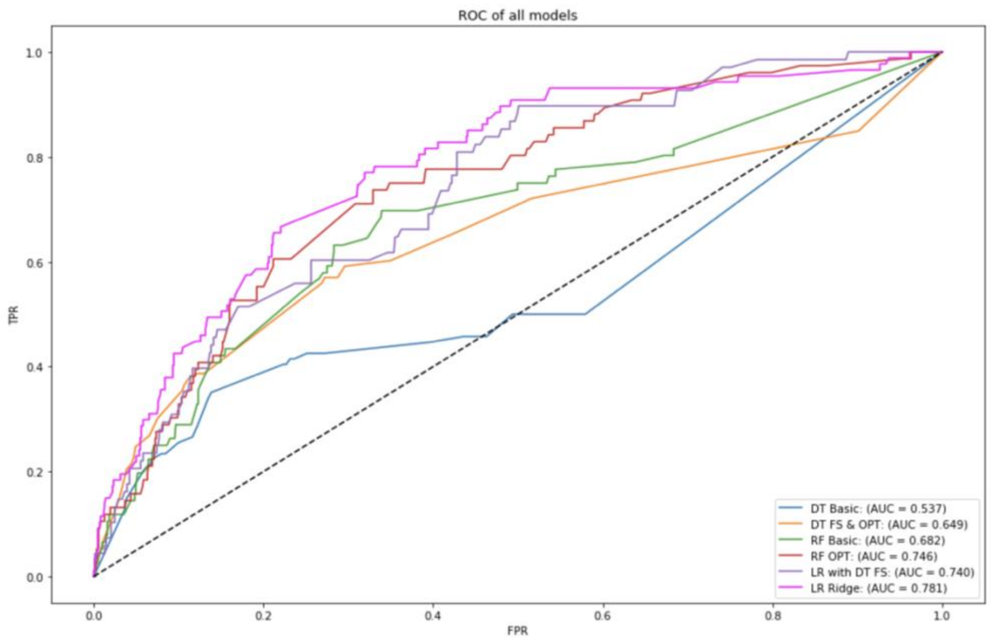

Figure 3. Performance of the data mining algorithms before and after optimization.

IV. Modeling & Evaluation

We used three different algorithms for this project: Decision Tree, Random Forest, and Logistic Regression, which were chosen as the most appropriate for modeling our binary response variable, “success”.

A. Evaluation Metric

As our main evaluation metric we used Area Under the Curve (AUC) for three reasons. First, it is a common, easily interpretable metric for binary classifications. Second, graphs can be plotted together for comparison across all of our models, identifying the best performers. Third, and most importantly, AUC is robust for skewed datasets (*Jeni, etc.*). As mentioned before, our dataset is unbalanced, with less than 9% of instances classified as “successful”. However, AUC tends to be more optimistic in its prediction if the dataset has a large amount of negatively classified instances, which is the case for this project (*Davis, etc.*). On the other hand, even though the individual evaluation of the model is affected, the comparison among different models using AUC should stay valid since the dataset is the same for all models; thus, their AUC should be equally optimistic.

B. Decision Tree

The decision tree algorithm was used as a baseline model for our project as it would help compute the information gain of the features to help classify our target variable. The advantages of Decision Tree are that the feature importance and their relations clearly visualize a decision-making process, and normalization of the data is not required. The disadvantages are that the prediction accuracy is low, and it tends to overfit the data, especially compared to Random Forest, as it utilizes a single tree.

We split our dataset into 75% for the training set and 25% for the testing set. We ran a Decision Tree classifier with all the default settings that we named "DT Basic", and the resulting AUC of this model was 0.537. This Decision Tree serves as a good comparison for the model improvements discussed below. (See *Figure 3* for a comparison of all models). We selected the top 36 attributes out of the 41 with normalized feature importance above zero. Some of the most important ones were 'BusinessParking_Valet' (importance score 0.169), 'Asian Fusion' (0.099), and 'Barbeque' (0.086). We proceeded to implement hyperparameter optimization with a grid search to control the complexity of the model, which improved the AUC to 0.649.

C. Random Forest

To further improve performance, we ran a Random Forest model, since the trees are "uncorrelated and protect each other from their errors" (*Yiu*). However, we have less control over the algorithm. We ran a model, "RF Basic" and got a higher AUC (0.682). We optimized this model with a randomized search cross-validation technique that would go through five different parameters: number of trees, features, depth, and the sizes of the splits and the leaves. The Random Forest algorithm with the best hyperparameters returned an AUC of 0.746, which proved it to be a far superior algorithm to the decision tree.

D. Logistic Regression

Finally, we decided to run a Logistic Regression, which would help us go beyond understanding which features tend to be highly associated with "success". We were interested in exploring the direction and scale of the impact that our features have on our target variable through interpretation of the coefficients. Using the important features suggested by our Decision Tree model, we achieved an AUC of 0.740, which underperformed the optimized random forest model. However, running Logistic Regression with all 41 features gave us a higher AUC (0.781), which confirmed that some features with low importance as predicted by Decision Tree still contained valuable information. The team also tried forward selection, using the recursive feature elimination (RFE) function from sklearn, which examines incremental improvements of the model when one feature is added. The RFE method did not change the set of features we had selected.

In attempts to improve our Logistic Regression model, we added regularization terms using a Grid

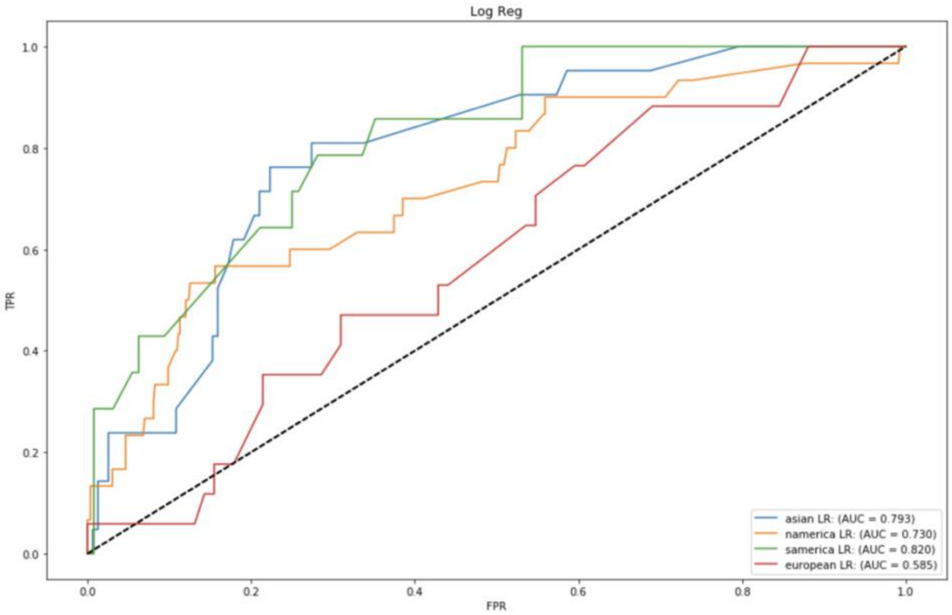

Figure 4. Results from running Logistic Regression after dividing by cuisine type.

Search CV technique to seek the best parameter C. Additionally, we ran bootstrapping to explore the effect of sample sizes on our model performance which indicates that the ideal sample size is within our available data range. In the end, the default parameters and ridge setting with full sample size remained as the optimal model.

E. Results and Further Iterations

Since our business problem is restaurant-type specific, we hope to provide applicable recommendations by subsetting our data to run individual Logistic Regressions by cuisine. This algorithm is a good candidate for this analysis as it is less affected by a small sample size. The AUC varies for the different cuisine types, ranging from 0.585 to 0.820 (*Figure 4*). This may be explained by the cuisine categorization methodology, which is better at categorizing Asian vs. European restaurants. Each model also has different features explaining the success of the restaurant. For example, the key attributes for an Asian restaurant include ‘Parking_Validated’, ‘Barbeque’, ‘Steakhouse’, and ‘Asian Fusion’. Barbeques and steakhouses turn out to be the most successful types of restaurants in the Asian cuisines. One possible explanation could be that Korean Barbeque and Japanese Steakhouses are often seen as higher-end restaurants that attract more reviews and attention on Yelp. More interestingly, the model identifies ‘Asian Fusion’ as a non-trivial feature. Restaurants that leave this feature out of their Yelp profile could lower their chance of showing up in a user’s search results for “Asian” food. For comparison, the list of important features for European restaurants include ‘Breakfast and Brunch’, ‘Coffee and Tea’, ‘Salad’. These results indicate that restaurants can add features to their Yelp profile based on typical searches by cuisine type to become more visible and successful on Yelp. Furthermore, while we are interested in reviewing specific attributes like ‘Music’, ‘Atmosphere’ or ‘Good for Lunch/Dinner’, the data is too scarce to extract any insight. Further analysis that include these attributes (when more data is available) would also bring value to our business objectives.

V. Deployment

A. Deployment

In order to deploy our model, we have created a preliminary version of a recommender system to help industrialize feature recommendations for any restaurant on Yelp. Given the name of the restaurant that is categorized in one of our cuisine types, the system would output a list of similar restaurants along with their star ratings, number of reviews and attributes. By putting their profile next to the nearest competitors and comparing with their attributes, we could give restaurant owners an overview of their competitive advantages and disadvantages on Yelp. They can look at competitors’ attributes and test the recommended features in hopes of making their restaurants more visible on Yelp. For example, given a target restaurant, “I am Thai Cuisine” (average rating: 4.5, review count: 258), the recommender system would output its top five competitors along with their average rating and review count: “Chun Fai Chinese Eatery” (3.5, 203), “Thai Cuisine” (4.0, 178), “Hachi” (4.5, 238), “Sushi Hiro” (4.5, 257), and “Sushi Takashi” (4.5, 279), along with the attributes. To build this recommender system, we used an item-based collaborative filtering approach, specifically a cosine-similarity function, to detect similar restaurants given a target restaurant’s name and cuisine type. The selected features are again 41 key features selected from our data preparation steps.

Once the deployment has been tested for Las Vegas, we propose that it be deployed on other cities in North America. One important decision will be how to correct for the bias of the model being trained on the Las Vegas dataset. A potential first step would be to redefine “success” (our target variable) proportionally according to the difference between the markets. Another consideration for deployment in other cities could be that as the features in our model are fitted to the Las Vegas data, there may be some crucial differences when testing on data from different regions. Therefore, it is important to keep note of any features that have significant values of information gain and monitor changes in the model resulting from these features.

Further improvements include scraping Yelp frequently to create a more dynamic dataset of restaurants that are improving to see which features they added/removed, and if they are sponsored or have ongoing promotions. This would allow for the creation of a recommendation system that provides incremental return for restaurants.

B. Potential Risks

The analysis was completed only on restaurants with Yelp profiles in Las Vegas. There may be large differences between what makes restaurants successful on Yelp in different cities around the globe. Furthermore, the analysis may extend beyond cities to more rural areas, which may also be less comparable. For example, the attribute 'BikeParking' pertains more to metropolitan areas and may not be relevant to smaller cities. Additionally, given that the team removed features with less than 50% positive instances and replaced missing values with zero, important distinguishing features with few occurrences could have been too granular for our model to pick up. Therefore, the recommended features to add to a Yelp profile could be considered as only those that are popular.

C. Ethical Considerations

Given that restaurants self-report their attributes, there may be a risk of restaurants reporting ones they do not have in order to boost their Yelp profile in hopes of driving more traffic. Yelp may also have an incentive to add more attributes to increase overall engagement within the site to boost advertising revenue potential. If the business model was to be implemented, Yelp could consider checking the accuracy of the attributes that are reported by stores through an analysis of user reviews as restaurants add features to their profiles.

Another possible ethical risk is that some restaurants may offer incentives for customers to leave a review on Yelp. This could be in the form of free appetizers or desserts, as demonstrated by a card that one restaurant was giving out to diners, offering them a Limoncello as a thank you for visiting them on Yelp (*Lhardy Kitchen + Bar*). Although Yelp *Terms of Service* explicitly state that business representatives should not "ask for reviews and [not] offer to pay for them either", we cannot be certain that every restaurant follows this, leading to a biased and false rating on the website.

References

1. “Fast Facts.” *Yelp Newsroom*, 30 Sept. 2019, https://www.yelp-press.com/company/fast-facts/default.aspx

2. “Yelp Dataset”. *Yelp Dataset Kaggle*, https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic_dataset_business.json

3. Chang, Hsin Hsin, et al. “The Impact of On-Line Consumer Reviews on Value Perception: The Dual-Process Theory and Uncertainty Reduction.” *Journal of Organizational & End User Computing*, vol. 27, no. 2, Apr. 2015, p. 32.

4. Davis, J., and M. Goadrich. “The Relationship Between Precision-Recall and ROC Curves.” *Machine Learning - International Workshop Then Conference*, 2006, p. 233. *EBSCOhost*

5. Guidelines, *Yelp.com*, https://www.yelp.com/guidelines

6. Jeni, L. A., et al. “Facing Imbalanced Data - Recommendations for the Use of Performance Metrics.” *Proceedings - 2013 Humane Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 245–251. *EBSCOhost*, doi:10.1109/ACII.2013.47.

7. Lhardy Kitchen + Bar, *Yelp.com*, 15 Sept. 2014, https://www.yelp.com/biz_photos/lhardy-kitchen-bar-coral-gables-2?select=FFtGgVBY5BdhFki2gwJDvA

8. Yiu, Tony. “Understanding Random Forest.” *Towards Data Science*, 12 Jun 2019.

Contributions

Each team member contributed to the research, design, and execution of the data mining problem. Team members met together weekly to consult and review together. The contributions below reflect leadership in particular topics:

- Elizabeth Combs (eac721): Data Understanding, Data Preparation, Logistic Regression, Results & Further Iterations, Deployment
- Anu-Ujin Gerelt-Od (ago265): Evaluations, Decision Tree, Random Forest, Logistic Regression, Deployment
- Wendy Hou (wh916): Business Understanding, Evaluations, Decision Tree, Random Forest, Deployment
- Emmy Phung (mtp363): Exploratory Data Analysis, Data Preparation, Logistic Regression, Results & Further Iterations, Recommender System

Appendix 1. Business Yelp Data Documentation

For more information on the dataset please visit: https://www.yelp.com/dataset/documentation/main

1) “**business_id**”: string, 22 character unique string business id ("tnhfDv5Il8EaGSXZGiuQGg")
2) "**name**”: string, the business's name ("Garaje")
3) "**address**”: string, the full address of the business("475 3rd St")
4) "**city**”: string, the city ("San Francisco")
5) "**state**”: string, 2 character state code, if applicable ("CA")
6) "**postal code**”: string, the postal code ("94107")
7) "**latitude**”: float, latitude ("37.7817529521”)
8) "**longitude**”: float, latitude ("-122.39612197”)
9) "**stars**”: float, star rating, rounded to half-stars ("4.5”)
10) "**review_count**": int, number of reviews ("1198”)
11) "**is_open**": int, binary (0 or 1) for closed or open ("1”)
12) “**attributes**”: json object format, business attributes to values ({"RestaurantsTakeOut: True, BusinessParking street: True…} ”)
13) "**categories**": an array of strings of business categories (["Mexican","Burgers", ..])
14) "**hours**": an object of key day to value hours, hours are using a 24hr clock ({""Monday": "10:00-21:00", "Tuesday": "10:00-21:00", ..})

Appendix 2. Table of Cuisine Categorization

The following terms categorize more than 90% of all restaurants into the first four categories. The 56 terms below were chosen out of the 100 most frequently occurring categories that were determined to fit into a single category.

Cuisine Type	Category Terms
Asian	chinese, japanese, asian, indian, sushi, dim sum, cantonese, ramen, noodles, thai, vietnamese, filipino, korean
European	mediterranean, italian, greek, french, creperies, tapas bars, middle eastern, halal
North America	pizza, sandwiches, delis, wraps, chicken wings, chicken shop, hot dogs, cheesesteaks, burgers, american, steak, southern, comfort food, cajun, creole, soul food, fast food
South America	mexican, latin american, tex-mex, tacos, salvadoran
Others	all other terms