Executive Summary of Yelp Data Analysis

Tuesday Group 5 - Jiatong Li, Qilin Hong, Xinyu Zhang

In this executive summary, we summarized our process and results of analyzing Yelp reviews data. Goal 1 is to provide suggestions to business owners, Goal 2 is to predict ratings and compete in Kaggle. We first talk about our Kaggle prediction model briefly, and then focus on the whole process of providing suggestions. Our target is Brunch & Breakfast restaurants in Las Vegas. We'll show you our suggestions based on analysis at the end of this summary.

Review Analysis

For the kaggle prediction model, we mapped each Yelp review into a 128 length real valued vector but we didn't limit the total number of words. Also, we constrained each review to be 150 words due to the length of majority reviews and the trivial difference on the Kaggle prediction vperformance. Finally, we used long short-term memory (LSTM) network with dropout regularization to reduce overfitting and we got a perfect score, which is 0.68, on the Kaggle by using 1 million training samples.

Rate	1	2	3	4	5
Number	16,870	14,684	21,643	41,347	74,979
Percent (%)	9.95	8.66	12.77	24.39	44.23

In this step, we assigned positive label to 4/5-star reviews and negative label to 1/2/3-star reviews respectively because most business owners hope 4-star reviews over 5. As the table above shown, the number of positive labels is twice the amount of negative labels. Even though the adjective words always have strong effects on predicting the label, it doesn't make any difference for the suggestions in goal 1. Thus we removed non-noun words in review texts and we would like to extract informative features (noun) from review tests. Based on these information, we could train different kinds of classifier and then get the most informative words.

The first classifier is Naive Bayes classifier, which uses the Bayes theorem to predict the probability that a given feature set belongs to a particular label. We could learn about our data from the most informative features to the least informative features. The corresponding results and the probability of a feature pair belonging to each label are listed in nb_informfeatures.txt (https://github.com/KiRinHong/ratingsInYelp/blob/master/data/nb_informfeatures.txt). While the negative features account for large portion of the informative features.

Another classifier we used is maximum entropy classifier, also known as a logistic regression classifier. The maximum entropy classifier converts labeled feature sets to vectors using encoding. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature text. You could get the corresponding informative features from me_informfeatures.txt (https://github.com/KiRinHong/ratingsInYelp/blob/master/data/me_informfeatures.txt). and the numbers shown are the weights for each feature. For example, '-1.332 overprice==True and label is positive' tells us that the word 'overprice' is negatively weighted towards the positive label. The informative words obtained by this classifier are more balanced. After we merged them, we selected some interesting features that the business owners are not familiar with to do further analysis.

Suggestions based on words

Given the informative word list, we ranked the significance of the occurrence of a word based on One-Way ANOVA analysis, which is performed on the stars of reviews to the occurrence of each word. The p-value of each word indicates the significant differences of ratings between reviews with or without the word.

Rank	Merged Words	Type of Suggestion	Topic
1	smile, Rachel, Ashley, Jovany, Jenifer, Mike	service	staff
2	dirt, hair, cigarette	service	hygiene
3	minute, hour, takeout	service	efficiency
4	refund, credt, policy	service	payment
5	vegan, hazelnut, chipotle, broth	food	menu
6	dry, overcook	food	cooking
7	bland, flavorless, tasteless	food	cooking
8	burnt	food	cooking
9	overprice, cost	profit	pricing
10	voucher, advertising	profit	promotion

Note: * denotes that the p-values are less than 0.05.

Among the words given in the list, majority of them are regarded as significant. We merge the words with similar meanings and analyze them into eight topics belonging to three types: service, food, and profit.

- **Service**
  - From staff aspect, we found customers give high ratings when the service attitude of the staff are kind. Thus, we suggest staff in restaurants smile more and the business owner might consider awarding excellent workers like Jovany.
  - From hygiene aspect, we suggest the restaurants keep the environment clean and ban indoor smoking.
  - From efficiency aspect, we found when the customers complain about the time of waiting using “minute” and “hour,” or choose “takeout” indicating not enough seat and long waiting time, they tend to rate low. Thus, we suggest the business owners be more efficient in serving dishes and arrange tables to reduce customer waiting time.
  - From the payment aspect, we recommend businesses process payment and refund more efficiently. For example, American Express Gold Gard would give you 4 times posints when you are dining at U.S. restaurant, which you could use them to lots of different parterner airlines.
- **Food**
  - From menu aspect, reviews mentioning “vegan” and three ingredients like tend to have a better rating. Thus, we suggest the brunch restaurants add vegan-friendly menus, and pay more attention to dishes with ingresients like chipotle, a kind of Mexico chili usually in sauce.
  - From cooking aspects, we suggest businesses to make dishes savory rather than bland, control the cooking time carefully, and mind the temperature of dishes.
- **Profit**
  - From the pricing aspect, restaurants receive a lower rating when customers think it is expensive, and when they think the price or meals in reality are not corresponding to the propaganda. Thus we recommend restaurants to balance between the pros and cons gained from pricing or promotion.

Suggestions based on attributes

We calculated the average score of each brunch & breakfast restaurant in Las Vegas by averaging all of the scores from the same businesses. We got 448 business owners in total. We obtained dozens of attributes variables from the business json files, and the average score from the review json files.

By analyzing the levels and counts of each attribute, we found that:

- There are some attributes that have more than half amount of missing values, eg. “HappyHour” and “ByappointmentOnly”
- There are some attributes that are so imbalanced that the number of one specific level may be too small, e.g., “BusinessAcceptsCreditCards” and “RestaurantAttire”
- There are some attributes that is empty, which means they are not for out target -- brunch & breakfast restaurants, eg. “BusinessAcceptsBitcoin” and “AcceptesInsurance”
- There are some attributes that are too complicated to analyze, thus we ommited in this project, eg. “Ambience” and “Music”
- We kept some attributes that worth analyzing, those are: "Alcohol", "NoiseLevel", "RestaurantsPriceRange2", "RestaurantsReservations", "BikeParking", "RestaurantsTableService", "WiFi", "RestaurantsTakeOut", "OutdoorSeating", "RestaurantsGoodForGroups", "HasTV", "RestaurantsDelivery", "Caters" and "GoodForKids".

After this, we performed One-way ANOVA analysis to the average star and each of these attributes. The null hypothesis is there is no difference between each group, and if the p-value is less than 0.5, we reject the hypothesis. For example, the p-value of attribute “Alcohol” is 0.275, thus we retained the null hypothesis; and the p-value of attribute “RestaurantTableService” is 1.56e-09, thus we rejected the hypothesis.

As a result, the following attributes are signifigant: “NoiseLevel”, “RestaurantsReservations”, "BikeParking", "RestaurantsTableService", "Wifi", "OutdoorSeating", "HasTV", "RestaurantsDelivery" and "Caters". After this, we could generate the following insights:

- NoiseLevel: Quiet places tend to have high scores
- RestaurantsReservations: Restaurant Reservations does not have good influence on rating
- BikeParking: Having Bike Parking lots tends to have high scores
- RestaurantsTableService: Table Service will drag the cores down

- Wifi: Even though the initial anova shows Wifi is significant, but from the plot and analysis, the number of paid Wifi is too small. If we omitted ‘paid’ and ‘NA’, it was no longer significant. Thus, we cannot give a suggestion based on Wifi.
- OutdoorSeating: Outdoorseating will pull up the rating
- HasTV: Restaurants that has TV tend to have high scores
- RestaurantsDelivery: Restaurants that can delivery always have good rating
- Caters: Catering will have a good influence on customer rating

Finally we deleted Wifi from the above nine, we put the rest 8 attributes into the final analysis. We used RandomForest to calculate the important score of these attributes so that we can give a sequence of importance to business owners.

Attributes	NoiseLevel	RestaurantsReservations	BikeParking	RestaurantsTableService	Outdoor
IncNodePurity	11.01	7.93	9.32	14.90	

The most important one is RestaurantTableService, followed by Caters, NoiseLevel and HasTV and BikeParking.

Attributes	Alcohol	NoiseLevel	RestaurantsPriceRange2	RestaurantsReservations	BikeParking	R
p-values	2.75e-0.1	4.48e-03	3.66e-01	4.99e-02	9.00e-05	

Attributes	RestaurantsTakeOut	OutdoorSeating	RestaurantsGoodForGroups	HasTV	RestaurantsD
p-values	5.73e-02	2.44e-05	1.52e-01	3.18e-05	1.

Conclusions

In conclusion, to improve the review ratings, we give the following suggestions:

I. General suggestions - Service, Food, Pricing and Promotion:

1. Service - Improve service with kind smiles, clean environment, high efficiency, and convenient payment transactions.
2. Food - Extend menu to be vegan-friendly, use more well-welcomed ingredients like chipotles, keep dishes savory rather than bland, avoid overcooking, and mind food’s temperature.
3. Pricing and Promotion - Avoid overpricing and false advertising.

II. Specific suggestions - A list of Attributes in order of importance:

1. RestaurantsTableService - If you are not aiming at a high-end restaurant, table service is not always necessary.
2. Caters - Catering is an attribute you worth have. You can add this if you don't have.
3. NoiseLevel - Quiet is what customers want. A quiet dining environment will boost your rating
4. HasTV - Having a TV may increasing your rating. Please have this attribute if you can.
5. BikeParking - BikeParking is necessary. Please have this attribute if you can.
6. OutdoorSeating - OutdoorSeating is necessary. Please have this attribute if you can.
7. RestaurantsDelivery - Restaurants that can delivery always have good rating
8. RestaurantsReservations - Restaurant Reservations does not have good influence on rating.

Contributions

- **Jiatong Li**: Cleaned and merged raw data, performed One-Way ANOVA analysis for attributes, calculated importance scores based on RandomForest, gave suggestions based on attributes. She is responsible for the summary, presentation slides and github commitment for this part.
- **Qilin Hong**: Processed raw text data, constructed kaggle prediction model and ran LSTM, provided informative features list based on NaiveBayes and Entropy for further study. He is responsible for the summary, presentation slides and github commitment for this part.
- **Xinyu Zhang**: Merged and extracted data of our target, generated histograms in exploratory data analysis performed ANOVA analysis of word list, gave suggestions based on words. She is responsible for the summary, presentation slides and github commitment for this part.

References

- Natural Language Processing with Python. O'Reilly Media Inc. (http://www.nltk.org/book/)
- Reviews, Reputation, and Revenue: The Case of Yelp.com (https://www.hbs.edu/faculty/Pages/item.aspx?num=41233)