Defining Positive Business Attributes for Ratings

From the Yelp Dataset

Chris Powell
Computer Science Department
University of Colorado at Boulder
Boulder, CO USA
chris.powell@colorado.edu

Donald Boles
Computer Science Department
University of Colorado at Boulder
Boulder, CO USA
dobo2888@colorado.edu

Jacob Bostick
Computer Science Department
University of Colorado at Boulder
Boulder, CO USA
jabo6853@colorado.edu

1. Problem Statement/Motivation

I, along with many other people allow star ratings to influence purchasing and dining decisions. Additionally we know customer ratings influence how well a business does [3]. As the old adage goes "the customer is always right", knowing this we can understand what the general population "likes" and "dislikes" with the help of sites like Yelp and others by taking advantage of statistical analysis on these large datasets. Our goal for this project is to evaluate influential attributes from the Yelp data set that contribute to positive reviews of a business. Using this information we hope to be able to make recommendations for key features that most highly affect star ratings. Additionally we plan to use this information to assess the best geographical region, the probability of star rating success givin business type, and will explore correlations for certain demographics leaving positive or negative reviews for a given business type. The majority of related work primarily utilizes sentiment analysis to justify ratings, along with finding key factors and ambiance that have an impact on the business. our goal is a little different in that we would like to explore attributes that seem to correlate with positive star rating and use this information to make suggestions for business ventures.

2. Literature Survey

S. Hegde, S. Satyappanavar and S. Setty [1] dive into the factors that contribute to a highly rated restaurant, in order to provide a model that one can use to make informed decisions when opening a restaurant. They do this by using three high priority tasks such as high frequency attributes, what days are crowded, and exploration into the location of the business. Looking at the review information they classify the attributes into categories based on single valued attributes (values that can be classified into true or false) and multi valued attributes which they classify under ambiance. They find the high frequency attributes by inspecting the max number of restaurants that have facilities with respect to high frequency attributes. They found that creadicars were the most influential attribute for hotels, along with Monday as the busiest day for restaurants.

M. Fan and M. Khademi, [2] attempted to use NLP to classify a users review in hopes they would be able to predict the given star rating for that user. To do this they grabbed the feature by finding the most frequented words categorized by their consistency in the corpus, additionally they applied four machine learning models to this corpus including Linear Regression, Support Vector Regression (with and without normalized features), and Decision Tree Regression. They found that Linear Regression performed the best for both the top adjectives and the general corpus of text.

Other similar work found on kaggle [4], explored the most rated business locations, and found some interesting observations about the key words. They also explored the top business categories by review count and used a histogram to display their findings. [5] They also explored the data and found that as business began to do well they also got more customers visiting, along with more chickens. Generally thought I did not find anything on Kaggle that seem to approach the problem the same way we will be.

3. **Proposed Work**

Ideally we would like to use regression across the list of attributes to predict business rating. Then we could review the data and propose businesses and test them on the model to see if it would predict a high rating. This would also allow us to hone in on specific attributes because we could have different models for different attributes, thus allowing us to create a business record comprised of the highest probability attributes and see if the interaction between them impacts the rating when predicted as a group instead of individually.

To get to the place where we can run our regression analysis we need to choose a few of the attributes that we would like to use and prepare them. Location is an attribute that is really intriguing to us and is split among address, city, state, zip code, latitude, and longitude. The location is available for a high percentage of the businesses on record which makes comparison better because of the large volume we have to work with. If we can't transform location into a variable for input into our model the plan will be to create our model based on other attributes of the business and then see the test data over a map where we can look for spots with high error.

K-means clustering will allow us to create clusters of the dataset. We can start with an arbitrary 192 clusters to get approximately 1,000 businesses per cluster. We can then iterate through the clusters and determine the averages for ratings to see which area has positive or negative effects on a business' rating. Then from there we could label the clusters to allow us to categorize rural or urban areas and whether that positively or negatively impacts the rating. Creating a visualization could help us refine the clusters.

The categories and attributes of a business we will convert into a new pandas object with a dozen of the most common attributes to start with. For instance, there are attributes like 'BusinessAcceptsCreditCards' and 'GoodForKids' that we could turn into binary for the regression, those examples already are but others are categorical. We would flatten the attributes with multiple categories into a series of boolean attributes.

We would like to use the business name. Since these are a series of words we could flatten them into a set of binary variables. First we would iterate through the business names and create a new dictionary with the top 50 words, removing useless ones like 'the'. Then we would iterate through our dataset and create a support and confidence for those words and each of the number of stars. Therefore a typical word could be 'brewery', which we would then check for each record containing 'brewery' and determine the support for the word brewery and the confidence that if the record contains brewery it is 5% a 5 star review, 45% a 4 star review, 40% a 3 star review, and 5% a 1 star review.

The review table has additional information about the reviews a business has received and we can total those into additional attributes for a business. Useful votes, funny votes, and cool votes may not represent the sentiment of a review so it would be interesting to see how a business' rating relates to the votes those reviews received and that would be something we predict once we come up with our business that we think would get the highest rating.

The data set is large and we are still reviewing but anecdotally we reviewed a couple hundred counts of reviews and it appears that Yelp only provided businesses with 3 or more reviews, which will help us to avoid outliers with a single positive or negative review skewing our data. We can create a distribution of the 192,000 businesses and their review counts to remove businesses outside 2 standard deviations. Also, we would like to see the distribution of stars since it appears that Yelp only records the stars in half integer bins, so 3.5 and 4 but nothing in between.

What is the probability that a business is open based on it's number of stars? We should be able to create a support and confidence for each half digit interval and the probability the business is still open. This could support the correlation between high rating and business success, not causation obviously.

4. Data Set

Our data set comes from the Yelp Dataset Challenge, Round 13. Yelp is a website that hosts a platform for users to rate and comment on businesses and services. The data set "includes information about local businesses in 10 metropolitan areas across 2 countries [6]." Within the Yelp data set, we plan on using the reviewer, business, checkin, and user json files. Each team member will have access to the data set locally on his machine. The Yelp data set can be accessed at https://www.yelp.com/dataset/challenge.

5. Evaluation Methods

We can evaluate our models based on the residual sum of squares comparing the neural network and regression models we use. We will also be using support, confidence, and k-means clustering as methods that one could evaluate our project on.

6. Tools

We chose to program in Python for its ease of use and vast amount of applicable libraries. Some useful libraries include pandas and plot.ly. Pandas allows us to conduct data analysis. Plot.ly enables us to graphically display the analysis on the Yelp data set.

7. Milestones

The milestones for this project are the progress report, project final report, project code and descriptions, project presentation, and peer evaluation and interview questions.

The table below explains the time frames for each step towards the milestones.

Step	Time
Data Integration	Oct 19 - Nov 6
Data Cleaning	Oct 19 - Nov 6
Normalization	Oct 19 - Nov 6
Feature Selection	Oct 19 - Nov 6
Dimension Reduction	Oct 19 - Nov 6
Pattern Discovery	Nov 7 - Nov 22
Classification	Nov 7 - Nov 22
Clustering	Nov 7 - Nov 22
Outlier Analysis	Nov 7 - Nov 22
Pattern Evaluation	Nov 23 - Dec 6
Pattern Selection	Dec 23 - Dec 6
Pattern Interpretation	Dec 23 - Dec 6
Pattern Visualization	Dec 23 - Dec 6

ACKNOWLEDGMENTS

Yelp's dataset challenge was instrumental in having a dataset that we could analyze reviews, reviewers, and businesses.

REFERENCES

[1] S. Hegde, S. Satyappanavar and S. Setty, "Restaurant setup business analysis using yelp dataset," in 2017, . DOI: 10.1109/ICACCI.2017.8126196.
[2] M. Fan and M. Khademi, "Predicting a Business Star in Yelp from Its Reviews Text Alone," 2014.
[3] K. Floyd et al, "How Online Product Reviews Affect Retail Sales: A Meta-analysis," Journal of Retailing, vol. 90, (2), pp. 217-232, 2014.
[4] Extensive Exploratory Data Analysis of Yelp Business Dataset
[5] Factors affecting closure of a business on Yelp!
[6] Yelp. 2019. Yelp Dataset Challenge Retrieved from https://www.yelp.com/dataset/challenge