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. **Abstract**

Many people, myself included, allow star ratings to influence purchasing and dining decisions. Additionally, we know customer ratings influence how well a business does [3]. As the 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 for a given business type and 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 a positive star rating and use this information to make suggestions for business ventures.

Our findings led us to evaluate a subset of the data located in Nevada (NV). Using a decision tree and information gain we found that DDS (Dentist Offices) had the highest entropy rating, and our findings were found to have an accuracy rating of 0.8490305653048011. Additionally, we found restaurants and food as having the highest entropy rating for each category and found this with an accuracy rating of 0.7968682318288139. This led us to believe that opening a dentist's office in the food space could be unorthodox, but an interesting proposal. Either way, the data leads to the food space as one's greatest chance for high ratings, and dentist offices in the area are also providing services that tend to leave positive reviews.

2. **Introduction**

The question we are interested in is where a person should open a business given the Yelp dataset. This question requires us to take a look at the data to understand the target we are trying to achieve along with what we can say about the interpreted information. Since we are provided with two sets of star review information it is our analysis that we should use the business review stars as a gauge and since this is an average of all of the reviews that are provided for the business we can say that we are interested in businesses that maintain a high average star rating, and are not limited to a certain type of business but are open to evaluating any business models that receive ratings from yelp.

This question is important because as big data becomes readily available we are provided an opportunity to make smarter decisions, and the decision to open a business has a huge impact on the local community in both financing along with providing goods and services that people are interested in. This is a key component in growing communities and provides stability for the community. This is also important information for the entrepreneurs that are taking risks in local communities. As the amount of opening and operational costs increase, so does the liability. Knowing this, it becomes crucial to have any information you can gain in deciding to open a business. With the help of the yelp data set and statistical analysis, we hope to demonstrate the value of such techniques.

3. **Literature Survey**

S. Hegde, S. Satyappanavar and S. Setty [1] dive into the factors that contribute to a highly rated restaurant, 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 based on single-valued attributes (values that can be classified into true or false) and multi-valued attributes that they classify under ambiance. They find the high-frequency attributes by inspecting the max number of restaurants that have facilities concerning high-frequency attributes. They found that creditors 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 user's review in hopes that 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. They also 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 keywords. 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 businesses began to do well they also got more customers visiting, along with more check-ins. Generally though, I did not find anything on Kaggle that seemed to approach the problem the same way we will be.

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, user (personal ID) and business files since it is closest to our data set. 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.

The data that is provided from this challenge is six json files (Review, Business, User, Tip, Checkin, Photo). The business file is comprised of fourteen total attributes having seven categorical attributes, four nominal, and three ordinal attributes. In the business document, we find nine total attributes having four nominal, two ordinal, and three categorical. Finally, our User document carries 22 total attributes having sixteen categorical, four ordinal, and two nominal. Working with these attributes provided a few challenges when trying to interpret the information using our decision tree process since two of the attributes in the business table name, and categories were categories but did not have number values associated with each. Our solution to such a problem was to use get_dummies which is a python library that converts the categorical attributes into a matrix with each of the categorical attributes moved to columns and if a tuple row had a 1 from our target attributes then the column with the categorical name would place one in the matrix index.

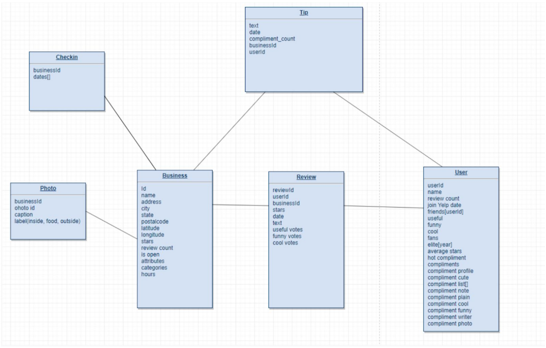

Figure 1: Attributes and data from the Yelp dataset

5. **Main Techniques Applied**

The storage and manipulation of our data was primarily accomplished locally, with pandas data frames. The reason why we chose this approach was because it seemed to be the most reliable way to work with the data without conversion errors. Additionally, if we were to attempt to upload the data to GCP MySql row by row it would have not been possible for us to accomplish in a reasonable amount of time.

The pandas dataframe allowed us the freedom to join the tables by using several merges. This was also possible because the data did not have any missing or Nan values. In addition, this led us to work with subsets of the data and transpose certain attributes to fit the current statistical analysis.

As stated earlier, we did not use the cube technique because we found some integrity issues with the data when converting the data to CSV format which was needed in order to use GCP's MySql for data cube processing. The other issue with converting our data set into a data cube was that there were no time stamp attributes which would make it difficult to organize the data by time.

The techniques that we did apply were information gain using a decision tree naive bayes classification and K means clustering originally. Information gain was chosen so that we would be able to assess each of the attributes in a reasonable amount of time while still having clarity on what's going on in the process. This method was chosen over the genie index because the data set was still too large which meant that it had a lot of noise. This could therefore have negatively impacted the results of a genie index.

The reason we chose naive bays was because it seemed like a pretty straightforward choice after we decided what metric we wanted to use for succeeding businesses. Using bays on top of this helped us pull however many results we needed to map on a visualization.

The last statistical analysis we used was K means clustering which was used to categorize the latitude and longitude elements based on the geographical location. After receiving the results from K means clustering, we added an additional attribute to label each of the clusters into a categorical type attribute.

We utilized clustering to make decisions on what information we wanted to further classify or inspect. This led us to use techniques such as K-means clustering to filter the data based on a geographical region. We did this in order to ensure a reasonable data size for further processing. We also wanted to use the data related to regions in order to find the most interesting locations for the questions we wanted answered. Each of the areas had some interesting insights but we felt that Nevada would suit our question the best because it had one of the highest average/mean star ratings and one of the lowest business open-to-close ratios.

Ideally we wanted to use information gain 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.

Our goal was then to get to the place where we can run our decision tree. To do this 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.

6. **What we’ve done/why we did it**

In taking the first few steps toward that end we have processed the data in a way that makes the most sense given the breadth and depth of the dataset. Since the fully merged dataset is over 6 million tuples, it became important to find a subset of the data to use so we are still able to process in local memory. With this goal in mind, we first decided to review the businesses in a cluster-based approach to see if there were any interesting geographical region that made the most sense for us to evaluate. Our chief concern was that business interactions are largely influenced by locality and community. We also wanted to be able to understand the context of the data, to spot global outliers.

Another consideration of ours was in deciding what star attribute to use, calculating individual reviews given the business, or using the average business review. Using a histogram we learned that there must be some bias in the individual review, as it was skewed to both the one start and five-star side. Looking at the business average star rating we found it had a normal distribution. Seeing this confirmed our suspicion that individual reviews were influenced by multiple biases. For instance, the type of person that might leave a review in most cases would be someone extremely happy with the service or dissatisfied, and those that had their expectations met might not leave a review. Additionally there could also be a situation where there are several different products by a producer, but if the individual review coming in are in response to a single product that the company might be trying out, or if the company is unaware of a defective product, there might be spikes of negative reviews that do not reflect the company, but rather the product that is being reviewed. On the other hand, we saw a close to a normal distribution for the business average stars which gave us more confidence in these attributes' effectiveness.

To get a preliminary look at the natural layout of locality information we grouped the business by the state to see what states contained the most business entries. From this, we found 11 primary states, with a few others which we concluded, were just noise. To confirm the groupings in these specific states we decided to use K-Means clustering and display a visualization. We felt comfortable using K-Means clustering because there did not seem to be enough outliers to skew the results. Another consideration of ours was the density of the data, so we first tried to use DBSCAN clustering since it's a density-based model; however, it became unusable since the processing time took to long. This led us back to using K-Means since we could predefine K making the process much quicker. To ensure we had a reasonable K value we used the elbow method to assess the optimal value for K. What we found was that the chi-squared distance plateaued quickly after the first clustering. This suggested that most of the data points were already naturally grouped. Using the previously obtained information we first decided to try K = 11; however, we found that one of the groups "AZ" split into two separate groups. This was due to the low sample size of a few clusters which caused the smallest two clusters to merge, ultimately causing the largest cluster to split. We then tried K = 10, which allowed us to view a more natural grouping.

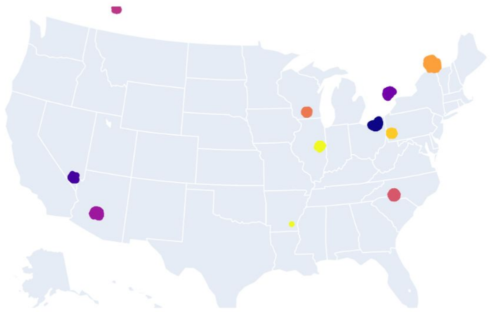

Figure 2: plotly graph displaying dataset clusters

Once we were able to identify clusters clearly, we started to iterate over the cluster and derive statistics for each of them. We wanted to identify a cluster that would be of interest for our project and then to drill down into it to allow use to work on a subset of data that was more manageable.

The metrics we decided derive were: open rate, average reviews per business, and average stars.

The open rate is the percent of businesses that are open as a percent of the total businesses in the cluster. The second derived metric was the number of reviews per business in that cluster by measuring the average. The third derived metric was average stars where we looked at the rating of the businesses in the cluster.

We also gathered each of those metrics for the entire dataset. The average nationally for reviews of a business was 34 and the average stars of those businesses was 3.6. The average for open businesses was 82.5% open.

Star's mean per cluster =
AZ 3.7071783749471012

NV 3.696218730219886
QC 3.634898312418866
WI 3.610895696006204
PA 3.577598502406846
NC 3.539541619443395
OH 3.505374880936182
IL 3.4647850854479545
AB 3.3856143856143857
ON 3.3563926872325784

Review_count's mean per cluster =
NV 61.79425929493354
AZ 35.327955282832555
NC 26.116861856189395
WI 25.130283055447848
PA 25.06141914779818
ON 22.77672720744442
IL 21.219057483169344
OH 21.129133215403456
QC 19.03147987884033
AB 12.08066933066933

Ratio of closed businesses per cluster
ON = 0.20606804105203314
IL = 0.20041429311237702
NV = 0.1858436304593114
WI = 0.18301667312911984
QC = 0.17265253137170056
AZ = 0.17241500916913532
AB = 0.16458541458541454
PA = 0.15920841504724548
NC = 0.15627754690844986
OH = 0.14648251462784057

We decided to choose the Nevada cluster because it was in the top three for all three of those categories.

To save space and time we decided to merge the tables in a pandas dataframe, but also decided to filter out for anything that was not in the NV k-cluster group. This allowed us to pair down to 2 million tuples, but the other advantage was that we were able to write out the result to a file so that we could import it in at a later time saving the previous steps taken for the merge.

The final thing we decided to do with the data before moving on to a classification tree was to add a column which we called star_target. This was a binary attribute that added a 1 if the business stars equaled five or zero otherwise. The reason we did this was so that it would be easier to compare against when we needed to select a target variable for our model. One of the main reasons that we chose information gain using a decision tree was that it seemed like it would be an approach which would allow clarity at each step of the process, and we chose to work with information gain rather than the gini index because our data set was large and included a lot of noise. Making this decision we evaluated all of the attributes of the merged file using pruning at a tree depth of 5. Using this method we found with a .9456603 accuracy rating. However, the biggest issues with this approach is that the business name and business category does not get process because it is categorical information that is not transposed into floating point numbers. This lead us to break down the data into two smaller groups one for categorical attributes and one for the name attributes.

To accomplish this we needed to tokenize each individual group and name into its own sub element since each name and category attribute had multiple entries. To accomplish this we tokenized each tuple string and placed it back into an array holding the column attribute align with the star_target. This allowed us to create sub data frames with category and names as separate dataframes with the necessary information for us to run our decision tree on. The last hurdle was to deal with was enumerating each of the categories, which we did by using get_dummies which turned each variable into a matrix having each name as its own column which crossindexes the target attribute.

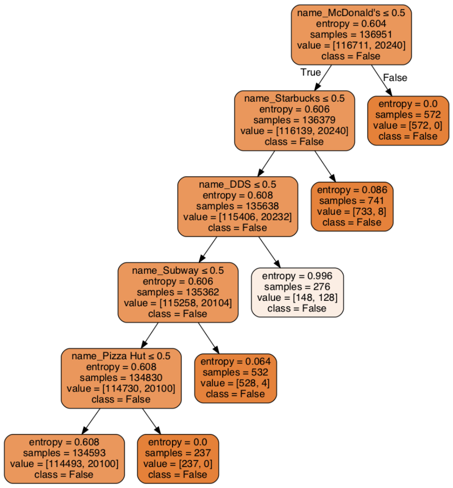

Figure 3: decision tree based on name

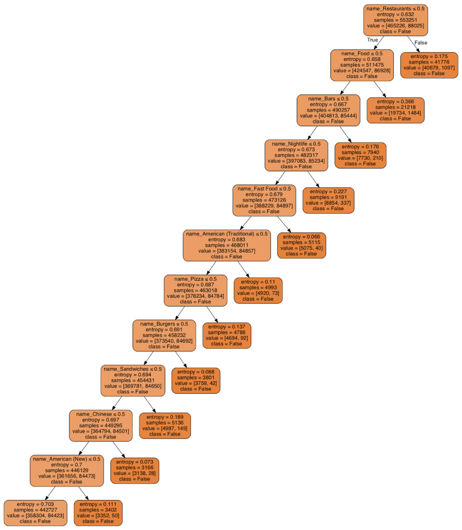

Figure 4: decision tree based on category

Next we used a naive bayes approach to determine the probability of each of the name attributes. This gave us insight into the actual probability of each of the business given the whole database. This also allowed us to be able to rank each of the business names so that we could visualize the top 100 business to see if there were any geographical interesting things happening

7. **Key Results**

Using K means clustering on the full data set we found Nevada to have one of the highest mean star ratings and business star ratings. It also had one of the lowest open-to-close ratios. From the decision tree we learned that dentist offices have the highest entropy which suggests that it is the highest attribute given from the decision tree. Also, looking at the business name attribute, we found that food also had the highest entropy for that attribute. Additionally, there did not seem to be a trend of these businesses local to Nevada, but it is interesting that people in this area really like their food and dentists.

8. **Applications**

The applications of our project fall into two categories: specific takeaways and ways that you could review our work for additional insight.

First, it’s easy to interpret that including ‘DDS’ in your business name is a significant indicator of being a highly rated business. What does ‘DDS’ mean? Doctor of Dental Surgery. So these are dentist offices? Yes, but maybe you can glean more from that. The word ‘Dental’ didn’t appear in the top terms, and with the split for ‘DDS’ being almost 50% 5 star rated, if it shared the same set of businesses it would have been one of the top terms as well. We think this might imply that some dentist offices include the doctor residing at that practice in their title while others are chains, where the doctors aren’t mentioned. The human connection to those doctors may be the reason that they are treated differently, their brands might be more about a person than a persona.

Second, additional insights from the decision tree and this approach. If you provided the decision tree with a proposed business name and set of attributes it would produce a probability of earning a 5 star rating. This would be helpful if you were considering opening a new business.

The decision tree could be pruned less to see what other attributes are desired amongst the population. For example, looking at figure 3, a less pruned decision tree would have more nodes on the right, so that the false nodes would have child nodes. The computation would have to be rerun to exhibit this behavior.

This approach could be used for existing businesses to get a baseline for what they should expect to be rated. If you were considering purchasing a business or owned a set of franchises, you could use the decision tree and the attributes of your business to determine how far off your rating is compared to what is typical.

While averaging rating and businesses nearby offers comparisons, this decision tree approach offers a different type of neighbor. Not a spatial neighbor, but a neighbor based on your features.

9. **Visualization**

Figure two shows eleven different dots. Those dots depict different clusters in the Yelp dataset. The states with dots in them are: Nevada, Arizona, Arkansas, Illinois, Wisconsin, upper Ohio, between South Carolina and North Carolina, Pennsylvania, in between New York and Vermont, and a cluster in Canada. We chose to focus on Nevada, which is the purple dot on the left side of the figure.

Figure three shows a decision tree based on name. Figure four shows a decision tree based on category. As described in the applications sections, using both of these decision trees together allows us to obtain a better understanding of how a business would do with a certain name and certain categories. Looking at figure four, fast food and bars are popular categories. Based on figure three, if that business selling fast food happens to be a McDonald's or Starbucks, then there is a higher chance that business would be rated higher than not. However, it is important to note that correlation does not equal causation. One explanation for this could be that there are more McDonald's and Starbucks than there are bars or other kinds of restaurants. Therefore, based on there being more of one kind of restaurant, then it is more likely that highly rated restaurants have that name.

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