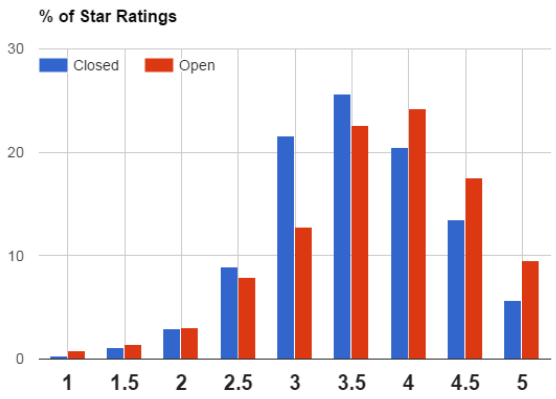
Exploring the warning signs of a doomed business

Michael Chin
A98115479
CSE 198 Assignment 2
UCSD

1. Dataset

Yelp.com is a business review website that lets users submit reviews on products and services that they have purchased. For this assignment, I analyze the dataset provided by the Yelp Dataset Challenge. It is a relatively large dataset, composed of over 1.6M reviews, 500K tips, 366K users, and 61K businesses. The structure of the reviews dataset includes the type, business, user, average star count, and text features. The businesses dataset on the other hand includes the type, name, average star count, review count, category, and open features. For my particular study, I will be focusing on the relationship between these two datasets, reviews and businesses, in the city of Pittsburgh, Pennsylvania.

On my particular subset of data, there are 3,041 businesses with 60,791 corresponding reviews. Of these remaining businesses only 2,670 are still open for business. The average star rating for any business within this subset is roughly 3.68 stars out of five. In my opinion, that is a fairly high rating, and gives me the feeling that the general user will rarely ever rate a business below 3 stars. One interesting observation is that the most common star rating for a closed business is between 3.0 to 3.5 stars, while the most common star rating for an operating business is between 3.5 to 4.0 stars. This seems to agree with the assumption that successful businesses tend to hover above the average, while less successful business tend to stay beneath the average.



Additionally, the average number of reviews that a business receives also follows the same structure as the average star count. The average number of reviews across all businesses is a moderate 23 reviews. As expected, the average number of reviews for open businesses is above the average, at 24 reviews. For businesses reported as closed,

however, the average number of reviews is at a meager 15 reviews. This makes sense as a reviewer needs to purchase a business's service of product before they can review it. So if no one is reviewing a business, then likewise, no one is purchasing anything from that business.

Lastly, the business dataset also contains a field named attributes. This field contains all of the miscellaneous information about a business, such as whether or it not it accepts credit cards, serves alcohol, or provides take-out. One particular sub-field, the "good for" field, caught my eye. The "good for" field contains sub-attributes such as breakfast, lunch, dinner, and late-night with values of either true or false. What is surprising, is that for restaurants that are labeled as closed, the "good for" column generally contains all falses. That means that reviewers believe that the restaurant is not good for breakfast, lunch, or dinner. It is alarming that a restaurant can be disliked by customers for any meal of the day.

From my analysis of the dataset, and further reading on articles beyond the dataset, there seem to be two different reasons as to why a business may close its doors for good. The business is either a failure and must shut down due to impending bankruptcy, or the business is too obscure and is not appreciated by the majority of its surrounding community. The first reason should be self-explanatory. One example of the second reason can be observed with the restaurant "Quiet Storm Vegetarian & Vegan Cafe". This restaurant holds a rather prestigious portfolio, having over 140+ reviews and an average star rating of 4.0 stars. The establishment was forced to close, however, as it was viewed by its landlord as a business that did not "...cater to a broader spectrum of people".

2. Predictive Task

As observed in the dataset, there are a number of features that seem to correlate with a business still operating. There also were two identifiable reasons for a business closing. With this information, I want to create a model capable of predicting the status of a business' operation. To measure my model's performance, I will split the dataset into a training set and a test set. The test set will consist of 30% of the businesses labeled as closed, and 25% of the businesses labeled as open, while the training set will consist of the remaining businesses not included in the test set.

The model itself will operate on two different distinct modes. First, the model will evaluate a business, using linear regression, based on the information provided by the business dataset. The algorithm used for linear regression consists of the standard form: $y = X\theta$. The features included in the linear regression will be the number of reviews, the average star rating, and the "good for" sub-column in the attributes field. Since only a subset of the businesses are actually restaurants, the model will only evaluate the "good for" sub-column on restaurants. For businesses that are not restaurants, the model will only take advantage of the star rating and the number of reviews. To represent the "good for" sub-column, the model looks at whether or not a restaurant received all falses. This is due to the fact that not all restaurants hold an evaluation for breakfast, lunch, and dinner. This preliminary step will serve as a baseline to dictate whether or not a business is successful or unsuccessful.

After labeling a business as either successful or unsuccessful, the model will then ignore all businesses that are labeled as successful. This is due to the fact that it is almost impossible to determine whether or not they will close. I will refer back to this later in my results section. For businesses that are labeled as unsuccessful, the model will further examine the review data associated with the particular business.

Using a list of phrases (unigrams and bigrams) generated by text-mining, the model will then use linear regression again to determine whether or not a business is as unsatisfactory as its star ratings describe it as. The text-mining process looks at the most popular words across all reviews, and then removes all stopwords. Based on the calculations provided by the linear regression, the model will then make a prediction based on the outcome of these two experiments.

3. Literature

The Yelp Dataset Challenge has been studied a countless number of times by competitors looking to win Yelp's grand prizes. One such study, *Personalizing Yelp Star ratings: a Semantic top Modeling Approach*, looks to correlate the behavior behind a particular user's review and their star rating. The most common method used to study the Yelp Dataset seems to be the use of a Latent Dirichlet Allocation (LDA) algorithm. This method is employed by researchers looking to observe the topics described by review text.

In regards to my predictive task, the University of Maryland has already built a similar model. Their piece of software analyzes the text of reviews for a business, and looks for keywords, such as "friend" and "good", that may indicate financial success. With their model, the professors were able to predict the impending shutdown of a business with a 70% accuracy.

4. Results

My model performed well on my test set, scoring a 77% prediction rate. I believe that this was due to the nature of my test set, since the number of operational businesses in the test set outnumbered the number of closed businesses 10-1. This is further reinforced by the fact that, of the incorrect predictions, only 32% of the predictions were false negatives, while the other 68% of them were false positives. This means that my model favored predicting that a business was still open. I believe that if the Yelp Dataset Challenge contained more businesses that were closed, my model would not perform as successful as it did.

Test Set Total	Correct Predictions	False-Positive	False-Negative
778	600	57	121

There are also some hiccups in the general structure of my model. For one, my model did not factor in the random closes, such as the example of “Quiet Storm Vegetarian & Vegan Cafe” provided in the previous section. To support these types of closures, I believe that the model would need to include neighborhood demographics to make a prediction as to whether or not the community would support the given business. For the scale of this model, I do not think that that would be feasible.

Furthermore, the linear regression model derived from the business dataset never came close to a “false” prediction. Theta[0] started at 0.73. To alleviate this problem, I counted any businesses above 0.80 as successful, and any businesses below 0.78 as unsuccessful. I discovered some interesting observations on the features though.

Contrary to my belief, review count does not have a large impact on whether or not a business is still operating. I suspect that this is due to a large volume of businesses that are untouched by frequent Yelp users. So while they may not have a lot of reviews, the business still may be financially stable due to a steady flow of non-Yelp customers. On the other hand, the “good for” column aligned perfectly with my predictions. While not as large of an impact as star count, if a restaurant was not suitable for any time of the day (breakfast, lunch, dinner), then its theta valued dropped significantly. Lastly, star count held the largest impact of the three features. This is no surprise, as a business should be directly correlated to its star count.

When it came to the text analysis portion of my model, the model favored predicting closed for businesses that were labeled as unsuccessful. I am not sure why the model performed this way, but I believe that it had to do with the ambiguity of the phrases it selected. For example, the word “price” was ranked as one of the most positives phrases, while the word “prices” was ranked as one of the most negative phrases. The stark difference that one letter can make is definitely a sign for red flag. As such, I do not think that the text-analysis portion of the model functioned properly.

The analysis was not all in vain though. One conclusion that I deduced from the analysis gave me further insight into the psychology behind user reviews. This was that users tend to be more critical when writing negative reviews than when they write positive reviews. This is reinforced by the weights of phrases (unigrams and bigrams) extracted from the review dataset. For businesses that are closed, their user reviews consisted of specific phrases such as “the bar”, “the food”, “the prices”, or “the menu”. Successful businesses on the other hand received reviews that were more general. These consisted of phrases such as “it was great” or “I loved it”. While somewhat unrelated to my predictive task, I found this particular observation to be interesting.

Before constructing my model, and analyzing the Yelp dataset, I thought that predicting whether a business would close or not would be an easy task. After trudging through the data, however, this is certainly not the case. There are many other variables, other than a profile and a few reviews on Yelp, that factor into a business’s doom. In the end though, while my model may not have been successful, I was fascinated by some of the conclusions and observations that I discovered along the way.