CS396 Project Final Report
May Ninghe Cai, Sophia Chen, Romir Hysko
8/14/2021

Introduction and motivation

Good reputation often contributes to the success of businesses. In this Internet era, online reviews are playing an increasingly important role in the forming of a business's reputation. Existing reviews could influence and bias how future customers form their opinion or impression about a business. A glowing review might attract people who are not initially interested, while a negative review could deter potential clients from patronizing. With services such as "review booster" and review-management software, it seems intuitive that existing reviews have power over the future success of a business. Just how much is that power?

For our project, we aim to study trends in review count, review sentiment, and ratings over time, and then further build a machine learning model to evaluate the impact of a review on the future rating of businesses. We chose to focus on the review.json and business.json files and obtained a subset of the Yelp data for our analyses after some cleaning and preprocessing. The highlights of our project are 1) the approach to segment the reviews retrospectively on time and obtain insights on how different aspects of review change over time, and 2) modeling the reviews with a gradient boosting regressor to predict the future rating of a restaurant based on features of a given review and the previous review trend of the restaurant.

Dataset and Data Cleaning

Given that we need to find businesses with a sufficient amount of reviews left over a large enough period of time, we first chose to filter out businesses with more than 2 years of review history. By comparing the time elapsed between the most recent review and the earliest review for each business, we filtered out 141,613 businesses that fit our criteria of having more than 2 years of review history. We visualized the distribution of review history for all these businesses using a histogram (Fig 1).

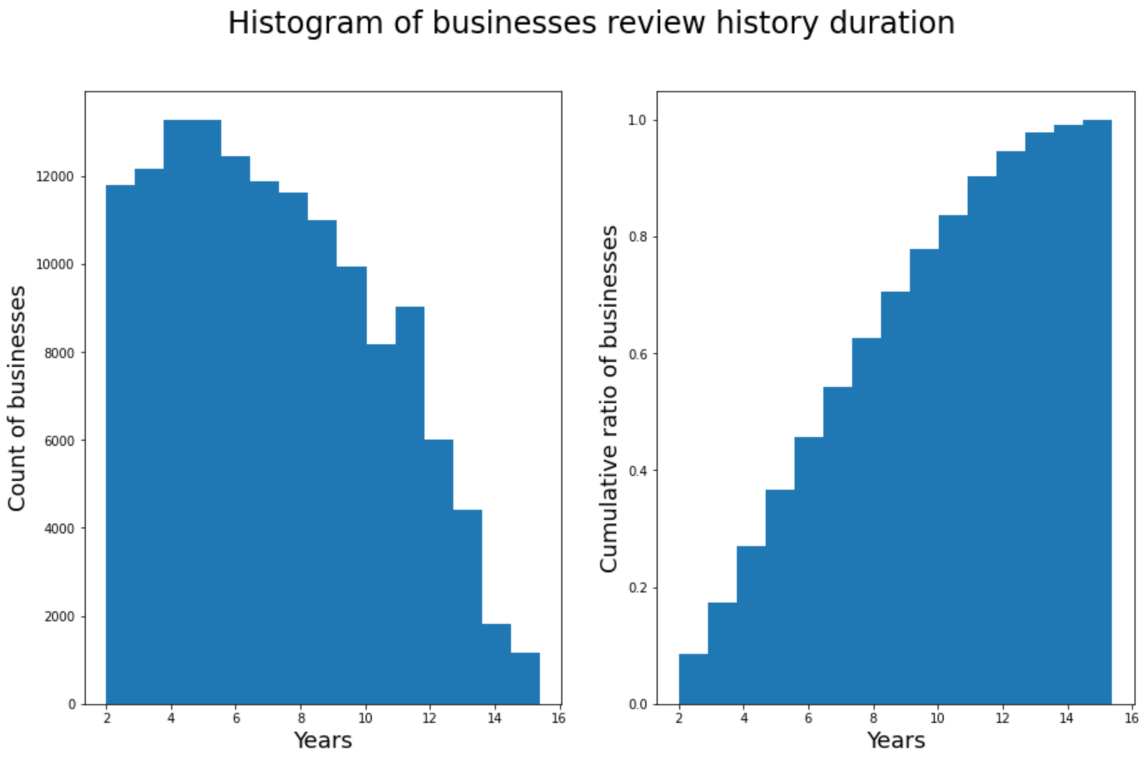


Fig 1. Histogram of the duration of review history for 141613 businesses with at least 2 years of review history. (Left) Count of businesses and (right) cumulative distribution of businesses binned by the duration of review history.

Based on this, we see that about 50% of the businesses we filtered out have review histories of 2-7 years, and the other 50% have review histories longer than 8 years. This is rather encouraging given that we initially were afraid that only a few businesses in this Yelp database would have a lengthy history.

Following, we plotted the histogram of total review counts for these 141613 businesses (Fig 2).

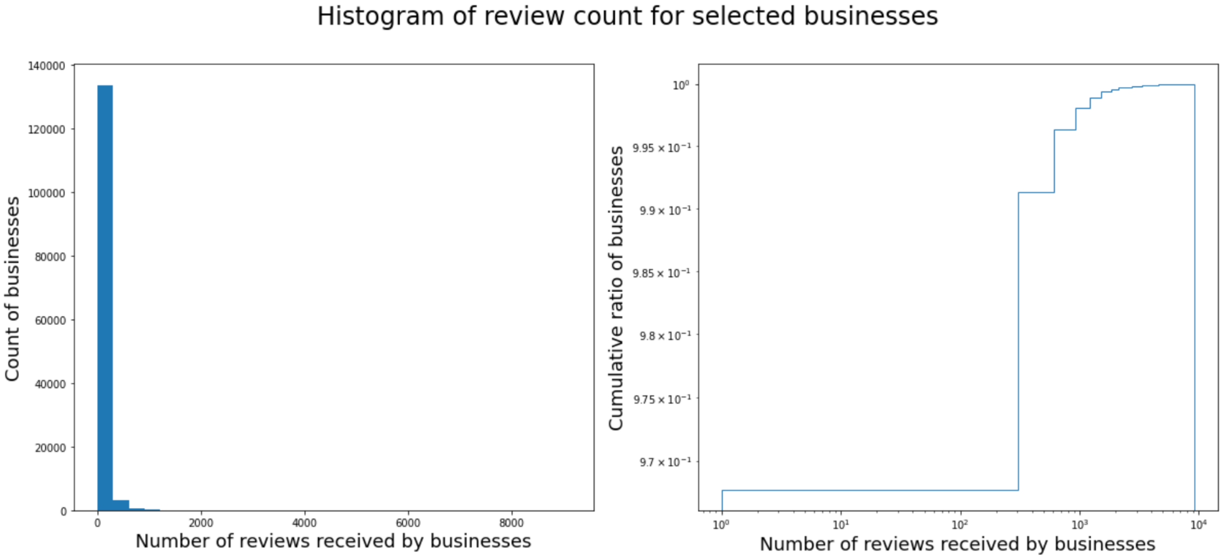


Fig 2. Histogram of the total review counts for the 141613 businesses with a review history of at least 2 years. (Left) Count of businesses, (right) cumulative distribution on a log-log scale, of businesses binned by the total review counts.

Based on this, we see that more than 97% of these 141613 businesses have accumulated less than 200 review counts over the course of their entire Yelp history. Given that our goal is to analyze the trend of reviews by time segments, a business with 8+ years of review history but only 10 reviews would not be ideal and would limit the type of analyses we can do. As a result, we further filtered and selected businesses based on their review count. This distribution of the total review counts follows the power law distribution - only a small percentage of businesses have received a large number of reviews while most businesses only have a handful of reviews accumulated over years. We actually fitted a power law model on this data and found that x-min of the fitted model to be 424. If we select only businesses with more than 424 reviews received over its entire history, then we would have 2544 businesses to work with, about 1.8% of the 141613 businesses. This approach may be a bit too aggressive. Therefore, we decided to select businesses whose total review count is above the 90th percentile of the review count distribution, which means we will select businesses with more than 129 reviews. This left us with 13,814 businesses, and 4,438,241 reviews to work with. We further cleaned the reviews to exclude non-English reviews. After this step, we used the data set containing 13,814 businesses and 4,337,534 reviews for our analysis.

Exploratory data analysis

To explore our data, we first plotted the histogram of review duration and the total review count for the 13,814 businesses we have after pre-processing (Fig 3).

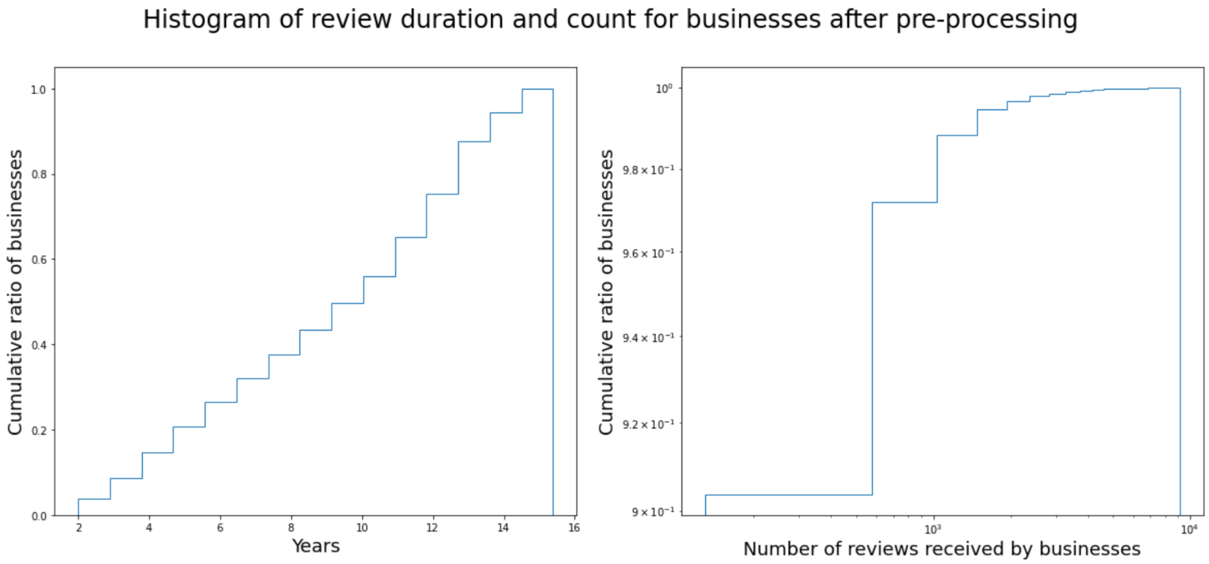
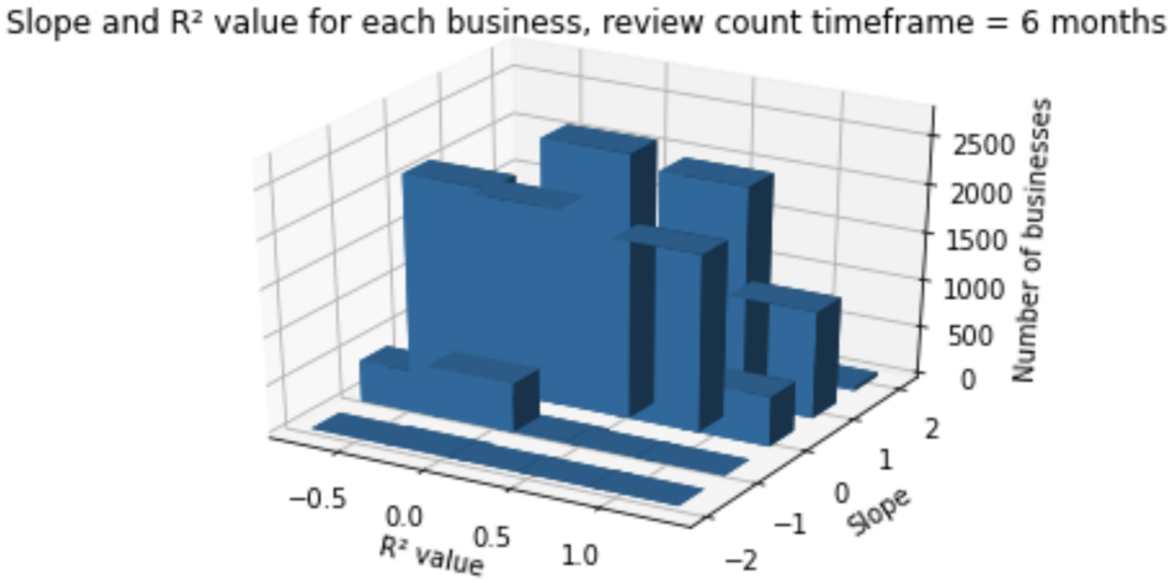


Fig 3. Cumulative distribution of review history duration binned by years (Left) and cumulative distribution of businesses binned by the number of total reviews received on log-log scale (Right).

Most of the businesses among the thirteen-thousand selected after our preprocessing have <1000 total reviews and over 50% of them have a review history of more than 8 years. The distribution of the review duration based on these 13,814 businesses is similar to that obtained in Fig 1 based on 141,613 businesses. The distribution of the review counts based on these 13,814 businesses is still a bit skewed, given that the data still follows the power law since we chose a cutoff value of 129 reviews, which is lower than the xmin obtained with our fitted power law model.

We also analyzed the categories they belong to and realized that the 13,814 businesses we obtained after preprocessing are all restaurants. This is important for us to keep in mind when thinking about the impact and applicability of our results.

We then segmented the reviews by a 6-month period, retrospectively summarizing their review count and average stars ratings received over that 6-month time window. We then used least-squares linear regression to get the overall trend of the review count for each business. Below is a 3D plot of the slope and R² value of the overall trend of review count for each business.



We then created two groups of businesses loosely based on the coefficient of the regression; one group consists of businesses whose review counts had a linear coefficient that was less than or equal to -0.5, and the other group consists of businesses with a linear coefficient of greater than or equal to 0.5. In other words, this loose grouping allows us to examine businesses that receive fewer and fewer reviews and the ones receiving more and more reviews; respectively, we name them group ‘decline’ and group ‘increase’.

There are 3262 businesses belonging to the group ‘decline’ , with an average slope of -0.28, suggesting that these businesses mostly are slowly declining in their review count. Among them, 867 businesses are closed. On the other hand, there are 2895 businesses belonging to the group ‘increase’, with an average slope of 0.65. If this result is not skewed by outliers, then this suggests that a lot of them are doing quite well and have a relatively quick increase in their review count. Among the 2895 businesses in the group ‘increase’, 301 of them are closed. The table below summarizes this result.

	Group ‘Decline’	Group ‘Increase’
# of open businesses	2395	2594
# of closed businesses	867	301

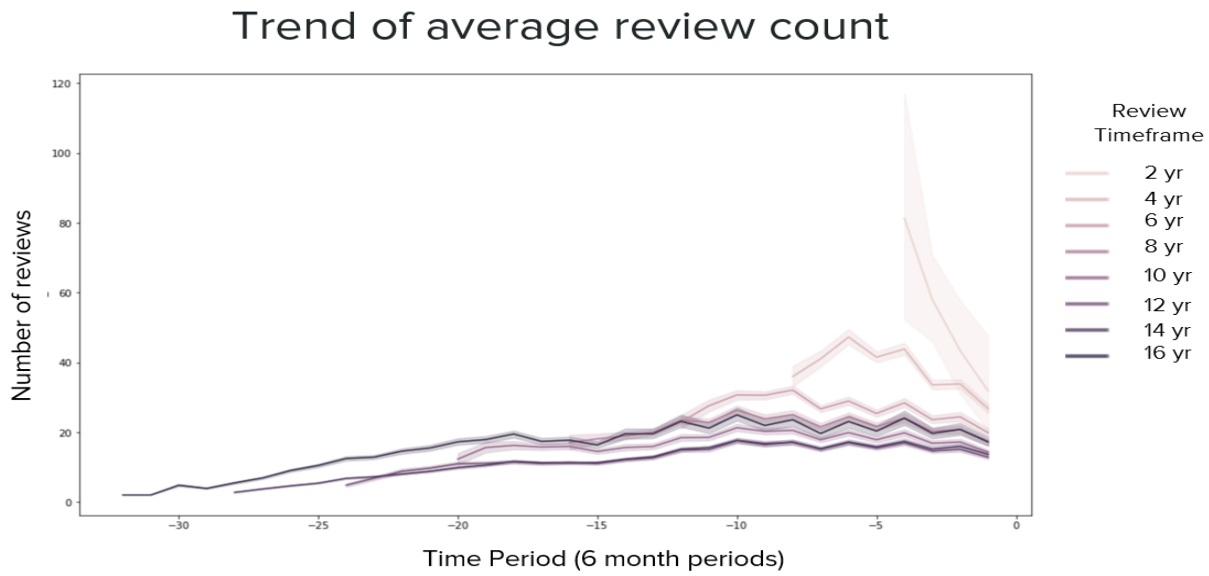
We were curious about whether the slope of the review count had anything to do with whether the business is still open or not. As a result, we performed a chi-square test to see whether these two things are independent. Chi-square value was 261 and p-values was 0.000, indicating that the slope of review count and whether the businesses are still open are not independent.

Data Modeling

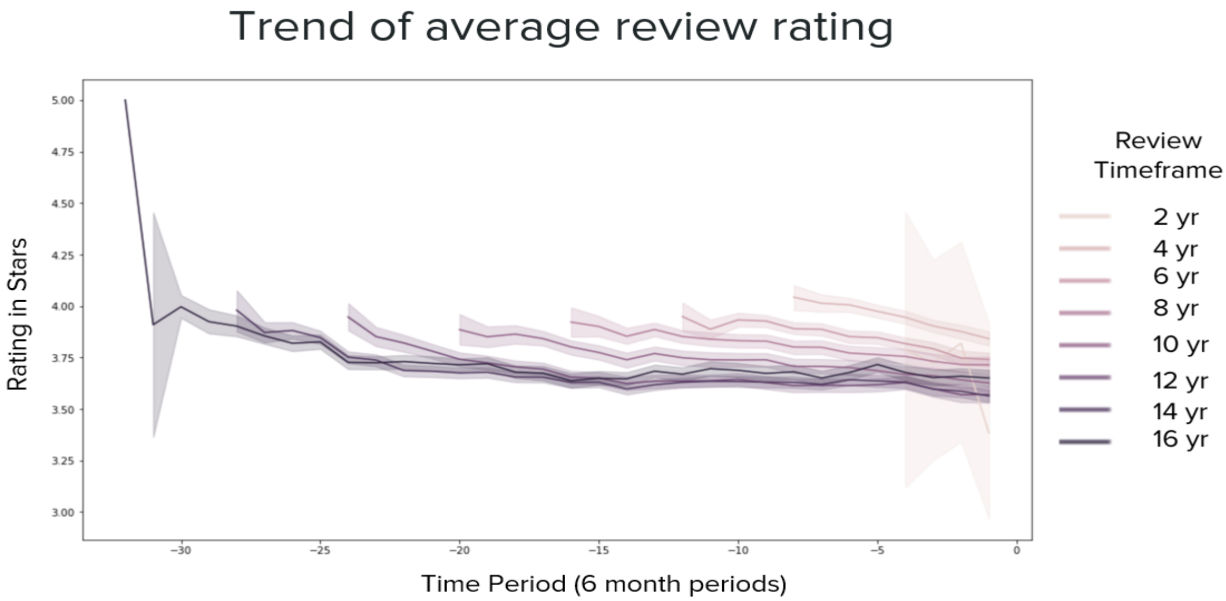
After determining our business category of focus would be the “Restaurant” category, we wanted to see if we could figure out which parameter changes over time have the most effect on the star rating of a restaurant, and if we could use these parameters to predict future ratings. We grouped information about individual reviews and business data into one dataframe. For individual reviews we added how many people found a specific review by a user “useful”, “funny”, and “cool” into one summation statistic for each review, called “total votes”. We then performed sentiment analysis using TextBlob over the text of each review to determine their polarity and separated this data into “negative”, “positive”, as well as “overall compounded” which gave us good data about review sentiment. Furthermore, we also integrated the data about total review counts and review ratings in our main dataframe.

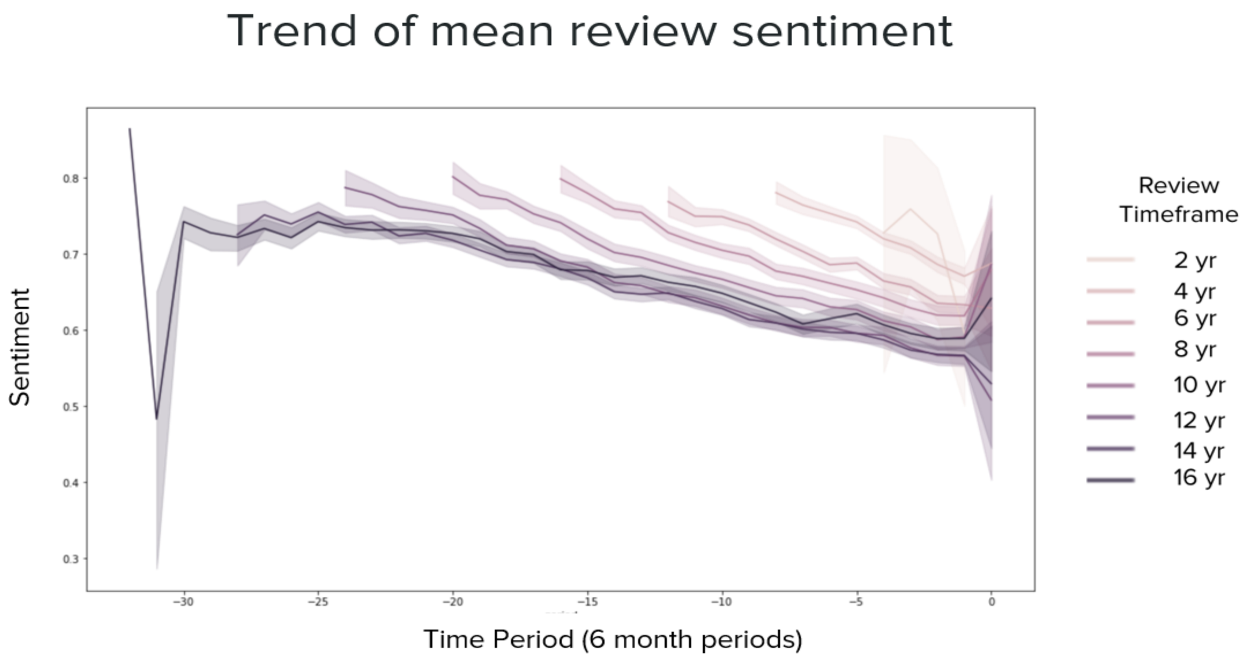
We decided to explore the overall trend of average review counts, average review ratings, and mean review sentiment. However, plotting the data for each separate business proved to be quite unsightly and hard to make sense of. In order to separate the data for better visualization we decided to separate them by a common feature, which we determined as the review time-frame of a business. While some newer businesses have reviews for only a few recent years, others go back to as many as 16 years, so we split the businesses by groups which had 2 years, 4 years, and so on until 16 years, which made the data much easier to model. Looking at the graph of the trend of average review counts, we see on the Y axis the number of reviews these different groups of businesses, as separated by review timeframes, received from their inception to today. The X axis depicts the time lapse with segments of 6 month periods, so for example -4 would be 2 years ago and 0 would be the date of the most recent review.

We can see that in general, the review count increases for businesses and then slowly plateaus towards a general decrease. There are some differences that we can see. For example, businesses that started about 16 years ago had a slow incline which most probably corresponds with the slow increase of Yelp usage by users. More recent businesses have a somewhat stronger increase which could be because most new restaurants get a large influx of customers in their opening phase, with many feeling inclined to try out a new place and share their thoughts about it. The eventual decrease we see for all businesses could be a part of the fluctuating trend that averages out over time.



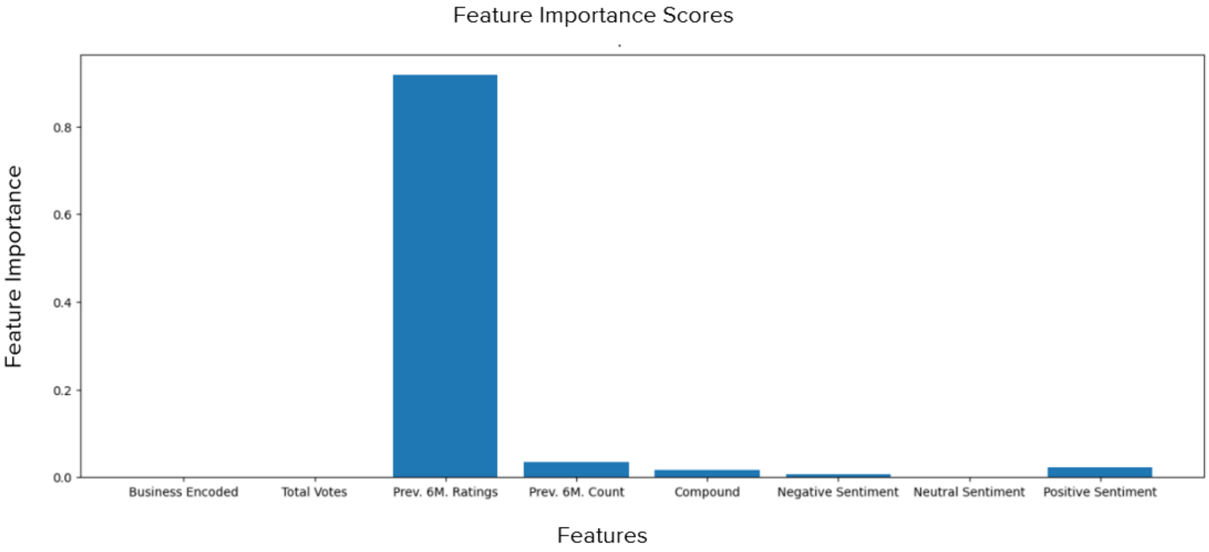
The same timescale and business grouping were used to model the trend of average review ratings and the trend of mean review sentiment. Looking at the trend of average review ratings we can see that most businesses converge to an average that ranges from 3.5 to 4.5 stars. It's interesting to see that almost all groups start at a rating of about 4 stars, and slowly go down to follow the general average. It could be the case that businesses receive better review ratings close to their initial opening because they're more dedicated to go above and beyond in order to capture a market share in the beginning, or people are just excited to see a new place and so tend to think better of it. Either way, they all seem to follow the same trend of a stabilizing average.





The idea of customers having a more positive outlook in the beginning can somewhat be supported by the graph above which showcases the trend of mean review sentiment. We can see that for businesses of all time-frames evaluated, they initially start with a more positive outlook from customers which slowly goes down and converges. We can also see that at around -1, which would be 6 months ago, most businesses have received a boost in positive sentiment which could be due to people going out again and re-enjoying restaurant experiences after the pandemic measures were somewhat mediated.

After modeling the trend analysis we wanted to see which features were most important in determining the rating of a business and if we could use these features to predict their ratings in the future. We decided to use a Gradient Boosting Regression since the rating is numerical, and a tree based model in order to rank our features by importance. The features we focused on were total votes of individual reviews, the aggregated previous 6 months rating and previous 6 months count for each business, general sentiment as positive, negative, or neutral, as well as compound sentiment. In addition, we decided to add business id in the model, depicted in the graph below as business encoded, in case certain businesses get better ratings just because of their reputation, which would be tied to their name.



From the model, it seems the number of total votes individual reviews receive don't matter much overall, and neither does the id of a business. In terms of sentiment, neutral doesn't seem to really influence the rating, which is to be expected as most people are moved by salient reviews. It's interesting to see that a positive sentiment has more of an impact than a negative one, however. The two features that seem to most affect business rating are the review count of the previous 6 months, and the star ratings of the previous 6 months, the latter being by far the most important feature in the model. Given that as was seen in the trend of average review ratings, ratings tend to converge over time, it's to be expected that previous ratings would be the most important feature in determining future ones. In terms of the machine learning algorithm, this is because the data has been established towards convergence over time and as such isn't expected to deviate much. We could also claim that in terms of real life expectations, since these businesses have been established in a convergent range over time, a decentralized mass of people has decided on a rating which accurately reflects the rating of a business, and given that the service a business isn't expected to change too much over time, future clients will share a similar experience and determine a similar star rating.

Summary of Findings

In summary, what we found from our data modeling is that the history of a restaurant can be impactful in determining their future. We saw how initially restaurants receive a lot of attention, seeing an increase in review counts after their first review, which could be due to the initial interest of people in a new restaurant being opened and wanting to try it out. The number of reviews stabilizes over time, which could be either due to people spreading out to other different establishments, averaging out even potentially amongst currently operating restaurants, and finally starts decreasing with less people feeling inclined to leave new reviews. This could be because it’s to be expected that most people that visit a restaurant are mainly local, and after most visitors have left a review, they don’t feel the need to do so again later. As the recurring locals don’t leave reviews, the number of new reviews is hence bound to decrease. Another part of this story is the trend of average review ratings, which seems to start at a high point for all groups of restaurants no matter when they were established, and converges to a range from 3.5 to 4.5 stars. Since most visiting locals again aren’t expected to provide too many repeat ratings, it’s expected their general sentiment won’t change too much and thus their ratings will either be similar to their previous ones or there won’t be any updates. On the other side of the story, businesses could also be more eager to go above and beyond to give satisfaction to their customers initially, but are less likely to be as enthusiastic as they establish themselves. This means their services aren’t expected to change too much, which means new customers will probably have a similar experience to others on average, which can be seen in the converging trend. The trend of review sentiment also showcases this, as the average review sentiment for businesses of all time periods seems to start high and converge to a lower one over time.

Performing 3-fold Cross Validation, we found our Gradient Boosting Regression model to have a mean R^2 of 0.59, which means that our regression model is good enough to explain 59% of the variance, but there is still room for improvement. Nonetheless, this result still showcases how important the previous overview of a business is in determining their future success.

Limitation and future directions

We selected our data based on the Yelp dataset available, and filtered out a lot of businesses that did not have enough reviews. Therefore, our analysis is limited by its scope, especially since we only focused on restaurants.

We built the gradient boosting regressor to predict the average rating of a restaurant in the next 6-month based on a given review and the restaurant’s past review trend. One limitation of this model is that the target feature (average rating of next 6 months of reviews) of the training data might not be super accurate, especially for restaurants with a shorter review history. Another limitation is that we modeled previous rating features and the future rating target feature using a time window of 6 months, which is relatively long. It is likely that our model might perform better if we segmented our data using a shorter time window like 3 months. In the future, we can further improve the accuracy of our model by incorporating features related to the reviewer, the subtype of the businesses, and other more detailed information of the review such as if the review is made in summer or winter.