Lukas Geiger (lg2960)
Zac Macintyre (Zim2103)
Github: https://github.com/z-data/Yelp-Project.git

April 19, 2021

Keeping Businesses Open in COVID 19

Abstract:

Our motivation for writing this paper is to help business owners and small business shoppers identify steps they should take to help keep their businesses open through and beyond the COVID 19 pandemic. Using the Yelp API, Yelp COVID 19 Addendum, and COVID 19 vaccination data we were able to identify features that would lead to success. Success is defined as an open or closed business. After a few merges and data cleaning we had information on ~8000 businesses in the United States. We identified review count, rating, price, grubhub enabled, and percent of population vaccinated as usable features for a logistic regression model. Our conclusion was that review count, business price, grubhub enabled and a self defined scaled feature (on COVID 19 vaccination density) were the most important features to small businesses staying open. We would have liked to collect more data from Yelp but were limited by API restrictions. For future studies, it may be beneficial to verify the data over multiple years as Yelp continues to collect new data points within the COVID 19 Addendum.

Introduction:

US local businesses are some of the hardest-hit industries in the COVID-19 pandemic. Many large businesses are able to wait out the pandemic until they are able to reopen stores to the public again. Small businesses, however, do not have this luxury. In this paper, we look to get a base level understanding of how hard US businesses were hit and generate a model that predicts which features would most likely keep the business open. We believed this was important for small business owners to know increasing their chances to stay open. Additionally, there might be actions customers can take to help keep their favorite local businesses doors open too.

Overview of Practices and the Problem:

We choose business closure as the metric for success in our problem. Traditionally one might measure business success using metrics, such as profit, number of employees, customers, or repeat customers. Given the COVID 19 environment of the country and world it is hard for small businesses to remain open, and we felt this was an appropriate metric. Additionally, business closure allows us to frame the problem as a classification type problem, and hence why we used logistic regression. The Yelp API provides the necessary labeling to run a model. Without Yelp this problem would have become much more complicated as there is no other widely accessible source of information for local businesses other than manually gathering data.

Throughout the course of the project, we had a number of iterations. First, planned to use a Kaggle dataset which had pre-cleaned data. However, we realized this was not a reliable dataset to use, as it was not managed by a governing body. In addition, the Kaggle dataset was outdated (from 2013). Then, we decided to pull data from Yelp, but did not yet know how to combine this with other datasets. After looking through Yelp API data, we realized we could bring value to business owners fighting closures during COVID 19. We first looked if Yelp had a complete dataset for COVID 19 use case. While they did have an additional COVID 19 addendum, this was not complete. It missed key information like business name, location, rating, etc. We made API calls on the business id to gather all essential information. More on how we gathered, combined, and cleaned our dataset below.

Data:

Reliability:

To prevent data snooping, the practice of significance chasing during statistical inference, we followed common data science guidelines. We clearly defined the questions to answer before collecting Yelp business data. We knew this data source was reliable and maintained. We are aware of potential Yelp business interest interference, but when looking at similar projects across different years we notice similar results.

Collection:

We collected data from Yelp on businesses through a combination of Yelp COVID-19 addendum and the Yelp API. The Yelp COVID-19 addendum contained information on the digital transformation each business took (if any). This dataset had information like if the business did delivery or takeout, if they had Grubhub enabled, if they had a COVID informational banner, etc. However, this dataset did not have information on basic business information such as the rating, price, location, and name. So, we queried the API for this basic business information. The process was slowed down, as we were limited to 5000 API calls on any given day. Additionally, the process seemed to be slowed because the Yelp API also has restrictions for how much information it will return in a day when making certain calls. For example, when using yelp business search the total number of new information Yelp returns is only 1000 entries a day. In total, we collected around 11000 rows in our data frame (DF).

We also mined data off of google's main page when you google “COVID statistics”. We got the data of total cases for each state, as well as percent vaccinations in each state. This data was easy to gather, and use. As discussed later, we used this data for feature engineering and also merged it together with our main DF using state as a merging key.

Cleaning:

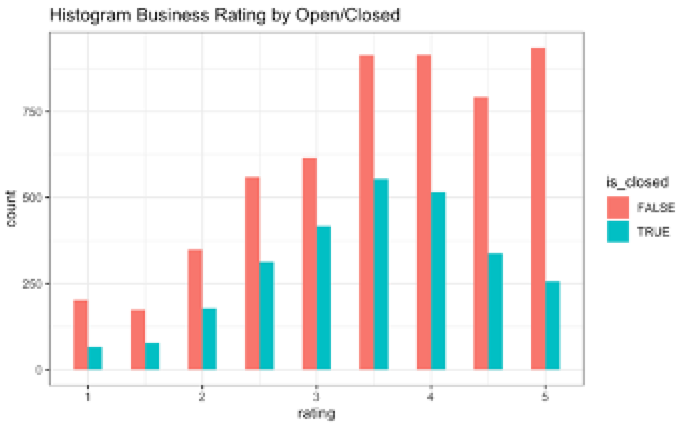
Much of the data was errors, Canadian businesses, or missing values. After removing bad data, such as the Canadian business or error terms we were left with ~8000 businesses. However, this data was not actually cleaned yet. Some of the cleaning steps were to transform empty strings into NA values. This would allow us to accurately get the percentage of missing data, which turns out to be 14% missing values. Given this data was mined on Yelp, this seemed to be a small number of missing values, given it is actually real data. Additional cleaning had to be done to turn characters into numerical values, as well as characters into Boolean values.

Now, on to the COVID statistics data. Overall, this data was very clean, the only thing we had to do was remove some empty rows and properly map the state to the right information. In other words, we had 2 smaller DFs and we combined them into one overall DF that had all of the COVID statistics in one DF mapped by state. The final “cleaning” step was to combine all of this information into one DF. This final DF had all of the yelp data and we merged the COVID statistics data with it by using the state a business was in and merged the COVID statistics state to it.

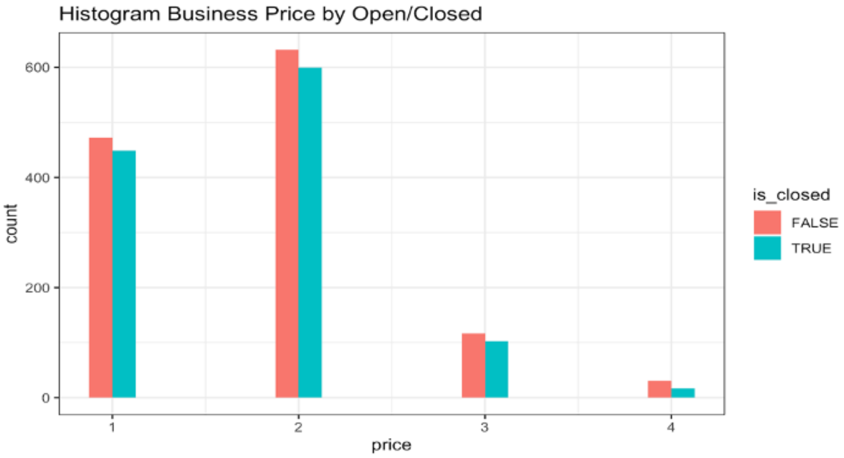
Distributions and Visualization:

We were worried that the distribution of closed to open businesses might be very skewed. If this distribution was not similar then our model would be inherently biased and not give accurate results. After looking at the data, we had a total of 5455 open businesses, and 2722 closed businesses. The results showed this data was not unevenly distributed to a concerning level. Closed businesses made up well more than 20% of the total data. Additionally, the finding seemed to make sense, as we expected more open businesses than closed businesses, but still anticipated a large number of closed businesses.

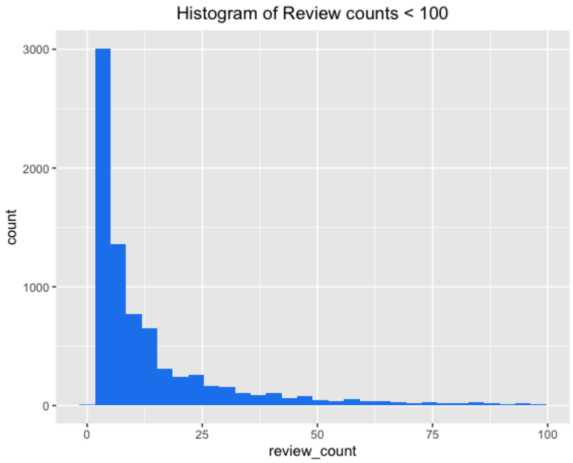
Some of the features were predictive in nature, such as rating. It makes sense the higher the rating the less you expect closures. If we take a closer look at the distribution of ratings split by open and closed businesses (figure to the right), we can see slightly meaningful results. We notice that businesses with higher ratings are more likely to stay open. After the 3.5 ratings, the number of restaurants closing trends downward while the number of open restaurants stays constant. When we run our model, we expect to see this feature as significant.

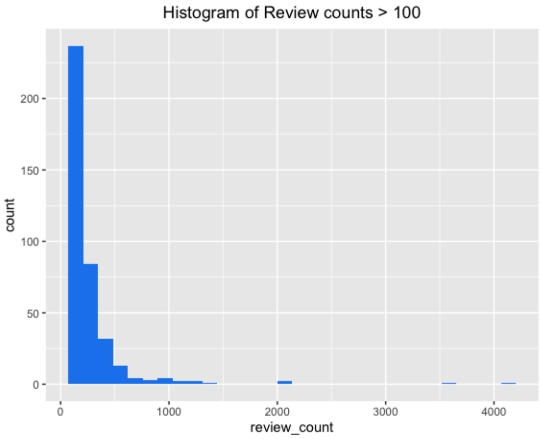


If we look at a similar histogram graph based on the price of the business (figure below), we now see less of a correlation. As the price of a business increases the delta between number of closures and number stays around the same. We expect this feature to have less significance when we run our model.



Review count is another feature we would like to use in our model. However, a majority of businesses have very few reviews on yelp. From the figures below we can see that over 2500 of our businesses have zero reviews. We do have a few hundred businesses with reviews larger than one hundred. However, we do not believe this will lead to a significant feature in our model because of the lack of reviews for most businesses.





Model:

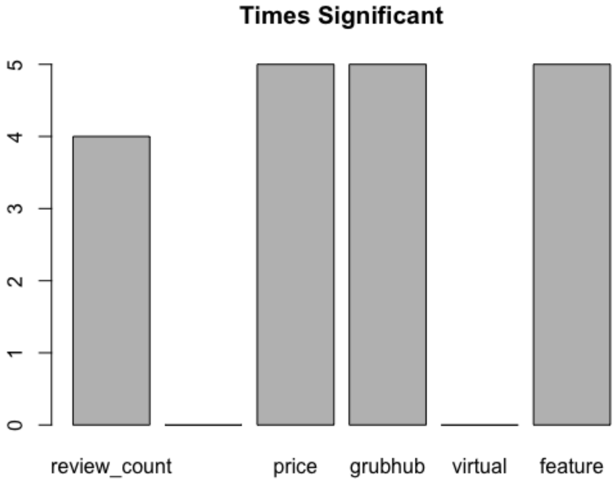
Feature Engineering:

For this project, we felt it necessary to engineer a feature to help with our model. Since we were dealing with COVID 19 data, and how it affects businesses, we engineered a feature to help capture some of the key characteristics. The feature we created was for every state. We took the total number of COVID cases and multiplied it by (1 - percent of people fully vaccinated). So, **New Feature = COVID Cases per State x (1- % of people vaccinated in state).** We believe COVID 19 cases help identify two things. First, it should be a decent representation of population density, and second, it should reflect how COVID conscious these states were. For example, New York has had one of the highest rates of cases because of New York City. Even though they follow guidelines quite well, the very dense population makes it extremely easy for the virus to spread. Whereas, Maine or Vermont are more sparsely populated and have fewer cases, even though they might not follow guidelines as well. Then, the (1 - percent of people fully vaccinated) could be thought of as a penalty, or maybe reward depending on the circumstances. It allows for places to counteract how well or poorly they handled COVID. For example, Texas had the highest count of COVID cases in America, but they are also one of the top three states for percent of people vaccinated. So, while they might not have handled COVID well at first, they are taking COVID more seriously now. In the sense that people are getting vaccinated. Ultimately, we thought this was a good way to represent population density and COVID conscious people, which we believe will help reflect whether or not a business will be open in these conditions.

Model Selection:

First, we needed to select features for our model. Out of all the data the following seemed to give us what we were looking for: [**review_count, rating, price, delivery.or.takeout, Grubhub.enabled, Virtual.Services.Offered, scaled_feature**]. Also, they were all the features that were not characters. Using a correlation matrix, we saw that “delivery or takeout” had a correlation of around .3 with 3 other features. So, we decided to remove it. The rationale being that we wanted to remove any potential situations for multicollinearity. Also, on an intuitive level this makes some sense to remove something that seemingly maps so easily to other variables. In other words grubhub enabled and delivery or takeout are almost exactly the same thing.

Next, for our actual model we used logistic regression to predict if a business would be open or not. However, we were most interested in what factors actually determined whether a business would be open or not. This is because it could help other businesses, and maybe help customers keep businesses open. The model was, “was the business open” and we regressed it on [**review_count, rating, price, Grubhub.enabled, Virtual.Services.Offered, scaled_feature**]. This is the feature we engineered, and then scaled it to have values between 0 and 1. Next, after running a few iterations over randomly determined data, the features that were most predictive were review count, business price, grubhub enabled and scaled feature. It is also important to note that all the values were negative while the intercept was positive. This means that a business starts with a higher log odds of being open, and as you factor in the key features, review count, business price, grubhub enabled and scaled feature, your chances of being open go down.



Model Outputs/Interpretation:

Given the above, that the key features decrease your odds of being open is rather interesting. Now, we will try to explain some intuition behind the features and the results. The first one we will mention is the scaled feature, COVID cases * (1 - percent of people fully vaccinated). This being negative makes some sense. It implies the more COVID cases you have untreated the less likely you are to be open. The next feature is price, and this also makes some sense. The higher the price the less likely a business will be able to stay open. This can be thought about as such, expensive restaurants cannot stay open because those restaurants specialize in gourmet food and hospitality. These things and ideas do not transfer to delivery. The same logic goes for really any other business, the higher the price, the more you expect from the business. The next two features review count and grubhub do not seem to make sense. However, we believe that the data has some COVID bias in it. So when adding more features to the model the better it captures COVID circumstances. This is what leads to the features being negative. Gubhub might be best explained by this, businesses that did not need a take out service before need it now.

A further investigation:

While the findings above had some interesting information, we believed that we could bypass some of the biases in the data. When looking at the data findings above, the one thing that stood out was that the review count was a small negative number, close to 0. We hypothesized that maybe if we control for business with lots of reviews we might find something different. That turned out to be true, when controlling for businesses with over 200 hundred reviews, review count and rating turned into positive log odd values. Meaning, that reviews and ratings to a business were actually important for a business staying open during covid. The intuition behind this is that people are checking businesses more and more online to get a feel for them.

When COVID hit, the businesses with the most reviews and highest ratings were the ones people were still willing to go to.

Conclusion:

Given our model selection we found interesting results about businesses ability to stay open. All of our features had negative values, meaning they all decreased the odds a business would be open. As mentioned before, the data suggests that any business should be open, but upon giving it information about Covid, all the features make the odds less likely. However, we were able to find out more, which is that when controlling for review counts we got better odds for a business to stay open. In other words, as your review count increases and your business rating increases you are more likely to stay open. This is an important actionable takeaway. It might be too late to save some small businesses now, and that is unfortunate. However, going forward customers should give good reviews of the businesses they like and businesses should push to have customers review their business. This hopefully will help current businesses stay open, and in the future can prevent shutdowns. Ultimately, whether you are the customer or the business owner do your part in making sure good reviews happen to places you want to see stay open.