Project Proposal

March 17, 2024

Course code: ORIE 4741

Github link: https://github.com/lincychen/ORIE4741-Project

Research Question: Can we predict the success of a business (open/close) based on Yelp review information, and if so, what factors extracted from the reviews are most predictive of business outcomes?

Dataset: Yelp Dataset: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset/data

Significance: As we grow increasingly more reliant on technology, the ability to share opinions has offered businesses the ability to gain valuable insights in understanding customer choice. Several studies have explored the connection between these factors, highlighting the importance of word-of-mouth on popularity of a business (Zhang, Ye, Law, & Li, 2010). Our project should reveal to prospective and current business owners whether online platforms have a significant influence on their success and provide actionable strategies for leveraging this influence to improve their outcomes. For example, if we discover that businesses operating in the restaurant industry are less prone to closure when there are more reviews published to their Yelp page overall, this would indicate to restaurant owners that encouraging customers to post a review after being serviced would be beneficial when it comes to maintaining operation. Furthermore, controlling for metrics that may influence business performance is a vital step in operations management; our findings will showcase how businesses can prioritize certain operations, or factors, over others in relation to their online reputation.

Not only is our research question important to business owners, but it could also signal to consumers the value of their Yelp contributions on the community. Ribeiro-Soriano's 2017 study showed that small businesses shape the regional identity and develop local economies. If Yelp reviews play a role in sustaining businesses, encouraging the local community to visit stores and publish reviews could be a crucial factor in gaining the traction that small startups often need.

Relationship Between RQ and Dataset: The dataset contains 150346 rows of relevant business-level information such as location, number of reviews, ratings, business attributes, and status. With this volume of data, we expect our machine learning model to have sufficient information for making accurate predictions. We aim to use operating status as our outcome variable and a measure for the success of a business. Operating status is a binary variable with values ranging from either 0 or 1, with 0 indicating closure and 1 indicating continued operations. Location is measured through a variety of features such as `city` (string), `state` (string), `postal_code` (int), `latitude` (float) and `longitude` (float). Including location in our analysis as a potential important factor would tell us whether a business's location impacts business operations. Further analysis could show whether an area with many businesses clustered together flourish or fail due to the competitive landscape. Another potentially important variable is `stars`, or rating. Yelp permits users to rate businesses on a scale from 1–5, with half step increments in between each (1, 1.5, 2, 2.5, … , 4.5, 5). The ratings serve as a quantitative measure of satisfaction with the business and if found to be significant, could link customer satisfaction to business

performance as well. Tangentially related is `review_count`, another quantitative measure on the number of reviews a business has received. If a business is performing well, we expect it to have more reviews, thanks to it being more popular. To reiterate, this information could help small startups revise their business strategy in making employees encourage customers to leave reviews on Yelp. Lastly, we can use the `categories` column to understand consumer preferences. This qualitative feature should reveal whether closures are more prevalent in certain industries than others. For example, if we found that the industry a business is involved in is more predictive of closure, then the implication is that consumers are more influenced by reviews in certain industries than others. Further research could explore at location-level whether different industries are more likely to receive higher reviews than others.

Hence we are likely to be successful in development of this project thanks to its intuitive, well-structured fields, clear target variable, and large volume of data. The features described above are of relevance to real-world business decisions, making our prospective model useful and generalizable when it comes to working with new data.

Project Value: This project would be helpful for Yelp to determine how best to tailor its recommendation algorithm by assessing the performance of restaurants, ranging from those that have excelled to those that have not. This can be evaluated through star ratings. Moreover, it would greatly benefit consumers and Yelp alike in keeping track of current trends, thereby increasing visibility for these restaurants. By monitoring consumer dining preferences, Yelp enables consumers to easily refine their interests in potential new restaurants and cuisines, such as plant-based meat for example. Not only on a consumer level, but also by understanding the factors contributing to positive or negative ratings, Yelp can provide feedback to restaurants within its platform. This could ultimately result in higher customer satisfaction and ratings.

Outcomes: The outcome of this study will provide specific factors and insights that will better Yelp's recommendation algorithm with the following expected outcomes:

1. **Study:** Conduct a regression analysis to estimate which restaurants or cuisines do well and are rated highly along with assessing their performance. Furthermore, we can analyze which factors will be the most important or predictive in a restaurants' success such as trends and categories.
2. **Recommendations:** Based on the findings, this will suggest what Yelp should do to improve their presence and better give advice as to how businesses can improve their success and customer satisfaction including service quality, managing reviews, etc.. Moreover, we can provide factors as to how we can best improve users' experiences on the app such as offering a more organized categorical feature.
3. **Implications:** The following insights will offer Yelp in how they can offer specific tools to help businesses like alerting businesses of specific trends. Moreover, discuss how these newly implemented features will help increase Yelp's user base.
4. **Limitations:** Although we have conducted the previous analyses, we will recognize the limitations of the research such as how the following reviews don't encompass all ratings so critical data may be left out. Moreover, there could be potential biases as to the kind of people who are leaving the reviews which might sway the types of reviews or analyses we conduct.