Yelp Analysis Project

Yashica Na – 260945954
Nadine Hamra - 260666146
Yichen Wang – 260761601
Alice Liu – 261007356
Matt Buttler Ives – 261055281

1. Overview of the project scope

Our project's primary focus is to help Yelp construct a dashboard that will analyze the performance of certain franchised restaurants listed on their website. The tool will be an available resource for businesses to determine what improvements they can make in their service to increase their user ratings. The dashboard will provide a low-performing store with a list of factors that existed during a period of time when a restaurant began receiving consistently low user ratings. The purpose of discovering these factors is to assist each restaurant in determining the causation for the decline in reputation and to provide businesses the opportunity to improve the quality of their franchises. The dashboard will consist of essential visualization tools and a Random Forest(RF) model to determine which attributes likely contributed to the low user review scores. Yelp will then communicate this information to each restaurant so that they may improve their product offering.

For the scope of our project, we will focus on McDonald's as an example of implementation. We assume that any restaurant performing lower than the 25% percentile of user review scores is a target for recommended service improvement. This baseline will allow us to determine the difference between outliers in the dataset and which restaurants are performing adequately. For our project, we also assume that the disparity between review scores is directly related to the service between each restaurant. With our model, we will be able to observe overall trends in the scores of reviews. These trends will allow regional franchise managers to determine when a

particular store began receiving lower reviews. If they continuously have bad reviews, such as over the last 30 days, our model will show the disparity of attributes between the restaurants with low reviews and the restaurants with higher reviews. Illuminating which attributes existed during this period will assist restaurants when they make strategy-based decisions.

2. Benefit

Benefits to the businesses:

Our proposal aims to create value for two key stakeholders: Yelp and franchisee businesses listed on Yelp's website. Businesses invest a great deal of capital into opening franchises. With the data offered from this tool, they will be able to locate poor performing franchisees and take action before they are forced to shut down. For example, was there a chef change, WIFI problems, or a decrease in speed of service? If Mcdonalds were to use our analysis on the quality of their franchises, they would see that they received the highest review scores in 2012. They could then track changes in their service since that period to determine what caused the decline in user score. New information regarding how the restaurants serves its customers is especially important during the Covid-19 pandemic as businesses are struggling to maintain their market share and consumer retention. Hopefully, with these adjustments, the restaurant will be able to increase its user rating and satisfaction levels among its consumers.

Benefits to the Yelp

By implementing our proposal, businesses can have the opportunity to gain valuable insights on their franchises. This incentivization from Yelp’s platform will encourage more businesses to list

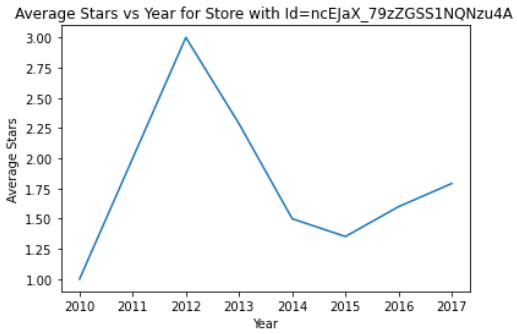
their service on the platform so they can access our model. An increase in listed businesses will result in an increase in user traffic to the website. As Yelp primarily generates revenue through advertising, the primary source of the company's revenue depends on the website's user traffic. This increase will assist Yelp's ability to earn money through advertising. Yelp would also be able to capitalize off the analytics report by offering it as an extra service tool for businesses to purchase. It could be charged to companies through a subscription or a one-time fee, both leading to an extra source of revenue for the company.

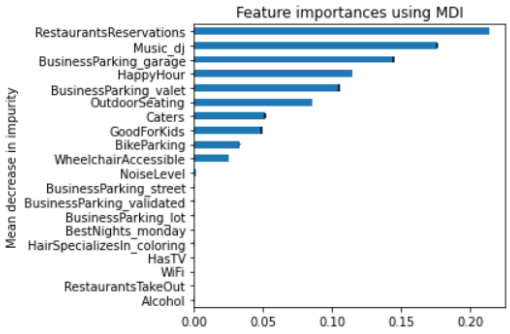
3. Implementation Strategy (more)

The first stage of the analysis would be to determine which stores are performing below expectations in user scores and what factors contributed to the low scores. For our analysis, we merged two data files on the variable "business_id": yelp_business.csv and yelp_business_atrributes. With the visual assistance of a boxplot, we can determine which stores are below the 25% percentile of user rating, (typically a rating less than 1.5 stars). From this information, we desired to know:

a. **Which are these stores? -** Using a simple filter in pandas data frame with a condition of stores with stars less than or equal to 1.5, we want to analyse only the stores that are currently open and can be improved upon. We added a dummy variable with the condition of is_open equal to 1. We found that there were 216 McDonald's stores performing in this area. For this specific analysis, we will focus on only 1 store that has the highest reviews.

b. **Analyse the timeline of the reviews and observe where it went wrong?** - We will be using the yelp_reviews.csv file to understand the store performance over the last 30 days. (In our demo, we are using the overall data from the years 2010-2017 since we do not have many observations for the last 30 days). The "date" column had to be modified as a datetime field and once the year was extracted, we wanted to group the reviews by year and find the average rating per year for plotting purposes. Using a simple line plot, we found that the store's user score performed well until 2013 when ratings fell sharply leading to the question of what occurred during this period. Perhaps there was a change in manager, or a new competitor opened next to it. With this information, businesses can dive deeper into the result with their internal data to understand what went wrong so that they can take correctional measures.





What went wrong and what can be improved? The last stage of the analysis displays a ranking of business attributes that directly affect the franchise store ratings. The implementation process of this stage is as follows:

1. Data cleaning is performed by transforming categorical attributes into dummy variables. Business attributes with null values are listed as false (0) and columns with the same values are removed.

2. We then proceed to run a RF regression with 100 trees with Y as review stars while keeping 30% of the data for testing to avoid overfitting. The reason why we selected an RF model is so we could present accurate estimates for variable importance, and to adapt to datasets with missing values and large numbers of variables.

Based on our results above, we see that Restaurant Reservation and Parking seems to be the most important services that can lead to a higher store rating. Due to the low quality of data for attributes, the result might not be accurate and hence need to be improved upon.

4. Post-Implementation Strategy (less)

To improve data quality for business attributes, Yelp can request the following information as a required field from businesses during their sign-up. Information that needs to be collected includes:

1. Location – city/suburbs area or population metric (can be collected with any Geographical API such as google maps)
2. Quantity and quality of staff members
3. Number and type of incidents that happened in the store (can be collected with a News API)

This data will add additional insights to the dashboard and provide further assistances so that businesses can improve their low user scores.