INSY-446-001 Data Mining for Business Analytics

Executive Summary Report: Yelp Star Rating Initiative

Ruo-Ying Qi: 261044984

Yudi Su: 261110250

Keane Dylan Yennoto: 261194825

Claire Zhao: 261194054

Kaibo Zhang: 261110409

Desautels Faculty of Management

McGill University

March 26th, 2024

1. Introduction

Yelp is a widely used online platform that connects people with local businesses through reviews, ratings, and user-generated photos. For individual businesses, Yelp serves as a window to increase their public exposure while also being an essential source for gathering customer feedback and dynamic market trends. By utilizing this platform, businesses can directly connect with their customers to enhance their reputation and foster business development. In this project, we aim to offer Yelp strategic recommendations on supporting its affiliated businesses. The analysis was based on a universal scale, encompassing all types of businesses, rather than focusing on one or a few specific categories.

2. Data and Preprocessing

We relied on the Yelp dataset comprising five distinct files to extract valuable insights about the company. Given our objective is to examine the influence of various attributes of businesses and customer reviews on star ratings, we choose to leave the user file untouched, as the data it contains is irrelevant to our research focus.

Our priority in the initial stages of data preprocessing was to comprehend the natural classification of businesses in the dataset. It became evident that Yelp's reach extended globally, with businesses located in 1,093 cities across 67 states and a remarkable diversity of 76,419 unique business types. However, such diversity also poses challenges for analysis on a universal scale as overly delicate classifications hinder the generalizability of findings and their applicability. Streamlining classification becomes crucial for clarity. To address this, we followed a common practice of multinational businesses and group businesses by continent,

utilizing the columns that document the geographical location of individual businesses by longitude and latitude coordinates. To ensure the reliability of our data, we removed rows with missing values in these two columns. The coordinate range for each continent is extracted from the world map[1] sourced from gisgeography.com. Rows that did not fall into any specified category were manually assigned to their designated continents based on their city.

To glean insights from textual data such as reviews and tips, we adopted the TextBlob[2] package for natural language processing (NLP). Transforming text data into numerical metrics: polarity, ranging from -1.0 (very negative) to 1.0 (very positive), and subjectivity, ranging from 0.0 (very objective) to 1.0 (very subjective), we then averaged these processed columns to determine the overall scores for each business.

Aligning with the scope of our study, we focused solely on universal attributes applicable across all business types. Consequently, we selected 'stars', 'review_count', 'check-in' (aggregated by sum), and 'label' (representing assigned continent labels), along with sentiment and subjectivity scores extracted from the "reviews" and "tips" files. For the attribute file, columns specific to certain business types were excluded. 'AcceptsInsurance' was also dropped as it only contains unitary values that have no variability or informative content. Over 90% of the rows in the attribute file were stored as 'Na', so removing them would significantly diminish the sample size. Instead, we retained these rows, treating blank values as a feature, and created dummy variables for the remaining columns.

Several strategies were implemented to address the presence of null values when we were merging data frames. Rows with null values in columns related to check-ins and columns of

'{attribute}_True/False' were filled with zeros, as they can be interpreted as indicating no corresponding records. We then filled the corresponding rows for columns related to '{attribute}_Na' with 1 to denote their logical definition, providing clarity in the dataset's representation. For sentiment and subjectivity columns, we assigned the mean value within their respective ranges to null rows to signify "no polarity" and "no subjectivity". By systematically handling null values using these methods, we ensure the integrity and completeness of the dataset for subsequent analysis.

3. Hypothesis Testing

We explored potential regional differences in businesses across continents by selecting distinguishing features and applying k-means clustering with an optimal number of clusters (k=2). Subsequently, a contingency table of sample proportions was generated for a chi-squared test to assess if one cluster disproportionately favored businesses from a particular continent. The resulting chi-squared statistic with 2 degrees of freedom was 0.0032, yielding a p-value of 0.9984. At a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis and conclude that there are no significant differences in businesses' distribution across continents.

4. Logistic Regression Modelling

Based on our project idea of inference, we selected logistic regression over more complex prediction models, prioritizing the interpretability of feature coefficients in alignment with our goal, rather than focusing solely on accuracy. To exploit the effectiveness of logistic regression in handling binary tasks, we partitioned our dataset into two classes based on star ratings, using a threshold of 4. The abovementioned k-means clustering analysis with k=2

supported this decision as the cluster with higher rating businesses has centroid values exhibiting star ratings of 4.23. Additionally, there is no class imbalance concern when using this threshold to classify businesses as "positive". All inputs were scaled by standardization.

Due to the prevalence of "Na" values across most '{attribute}_' variables, columns representing Na values gain undue significance. To address this, we assumed that if the data is "Na", it is false. However, these resulting almost unary variables led to convergence issues in computing coefficients. As a step further, we dropped variables with minority class ("True") frequency below 0.05. After this step, we found that all the coefficients were statistically significant at the 0.05 alpha level, except for "WheelchairAccessible_True". Therefore, we dropped this predictor and retrained the model with the remaining variables. Conducting a 10-fold cross-validation on the dataset also yielded an average accuracy of 80.96%.

Based on the regression output, we found that "sentiment_comment" notably influences businesses' likelihood of achieving a rating above four (what we define as being a "positive" business). The coefficient indicates that for every standard deviation increase, businesses are approximately 12.94 times more likely to be classified as positive. However, counter-intuitively, we discovered the number of check-ins has negligible impacts on rating performance, with an odds ratio close to 1. In addition, having valet parking decreases the odds of a business being considered favorable. These insights offer valuable guidance for decision-making and strategy development related to businesses and their ratings.

5. Proposed Initiative

In line with our findings, we propose an advisory service available to all businesses

registered on Yelp seeking to enhance their ratings. Leveraging our developed model as a guideline, this service will assist businesses in comprehending and forecasting their probability of achieving top ratings. As part of additional support, we used WordCloud[3] to analyze frequently appearing keywords among reviews in the top and bottom 25% of sentiment scores. Figures 1 and 2 highlight the recurring positive and negative review keywords. Through this initiative, Yelp can potentially attract more businesses to sign up on its platform and aid merchants in elevating their ratings by reinforcing frequently mentioned positive aspects. Conversely, steering clear of actions associated with negative keywords in reviews may assist merchants in mitigating the likelihood of receiving lower ratings.

Figure 1: Recurring Keywords in Reviews with Top 25% Sentiment Score

Figure 2: Recurring Keywords in Reviews with Bottom 25% Sentiment Score

6. Post-Implementation Strategy

Since the scale of the current analysis is universal, many of the regional and business-types variations are neglected. Thus, plans for more detailed exploration can be beneficial. Likewise, substantial null rows in attributes data limited our capacity to draw meaningful conclusions from them. Moving forward, we plan to persist in collecting data on these attributes. The dataset can also be enhanced by closely monitoring businesses that actively adopt the advisory's suggestions. This proactive approach will allow us to gauge whether tangible improvements in star ratings occur as business owners integrate features advised by the initiative. Through this feedback loop, we can iteratively refine our strategies accordingly.

Reference

[1] GISGeography. “World Map With Latitudes and Longitudes.” *GIS Geography*, 18 Nov. 2023, gisgeography.com/world-map-with-latitudes-and-longitudes.

[2] *TextBlob: Simplified Text Processing — TextBlob 0.18.0.post0 Documentation*. textblob.readthedocs.io/en/dev.

[3] Sidlaurens. “GitHub - Sidlaurens/Wordcloud: Lists the Most Frequent Words in a Text File.” *GitHub*, github.com/sidlaurens/wordcloud.