Yelp – What makes a highly rated restaurant?

CSPB 4502, Spring 2023 – Data Mining Project

Eric Hanley
CSPB 4502, Team 3
University of Colorado
eric.hanley@colorado.edu

ABSTRACT

The work described in this document is a project for CSPB 4502 (Data Mining) for the Spring 2023 semester. This data mining project will analyze an open dataset of Yelp restaurant reviews to understand restaurant trends, identify investment opportunities, and find commonalities across successful restaurants. Additionally, reviewers will be evaluated to identify trend-setters – those reviewers that highlight subsequently popular restaurants first.

The project will include a full data workflow including ETL, analysis, and presentation. The work will be completed in a combination of open source and commercial software with findings presented in a report and all source code hosted on a GitHub repository.

1 Problem Statement/motivation

The motivation for this project is to derive insight from a large publicly available data source. The analysis seeks to understand the attributes that contribute to a high Yelp rating for restaurants. The author will consider the evaluation of restaurant ratings at several levels of aggregation (city, state, region, etc.) and along multiple dimensions (restaurant category, food type, etc.). The desired outcomes of this analysis include:

- Identification of potential investment opportunities in a given market
- Insights that current restaurant owners can use to improve their likelihood of success
- Probable trends in restaurant popularity in a given market

An additional project objective is to identify restaurant reviewers that identify trending restaurants. The author seeks to understand if there are reviewers who:

- Repeatedly submit positive reviews for new restaurants that subsequently become popular and highly rated
- Issue early negative reviews for restaurants that subsequently decline in ratings

2 Literature Survey

A literature survey of prior work based on the Yelp dataset yielded many project examples. Three such analyses are described in this section.

2.1 Kaggle Notebook – Exploratory Data Analysis

The Kaggle Notebook *What's in a review? - Yelp ratings EDA* [1] contains a thorough exploratory data analysis of the Yelp dataset. The notebook includes summary statistics of reviews, reviewers, locations, and other attributes and is not limited to only restaurants. The notebook also includes some interesting geospatial charts and analysis into the temporal aspects of reviewer check-ins. This is a Python notebook.

2.2 Kaggle Notebook – Detailed Review Analysis

The Kaggle Notebook *A Very Extensive Data Analysis of Yelp* [2] contains a deep dive into the reviews of a sampling of some of the highest rated restaurants in the dataset. The analytical approaches include world clouds, sentiment analysis, bigrams, and word relationships. This is an R notebook.

2.3 Stanford CS 229 Project Report

The report *CS 229 Final Project Report: Using Yelp Reviews to Improve Businesses* [3] details the findings of a machine learning project seeking to identify which aspects of a business are most important to customers based on review text sentiment analysis. The approach implemented and compared both Naïve Bayes and SVM approaches to evaluate review sentiment.

3 Proposed Work

3.1 ETL

The data provided by Yelp is clean in the sense that column types are honored, there are few nulls, and the data is nicely structured in a JSON format. However, the data includes reviews for many types of businesses and will have to be filtered to only include restaurants. The data will be filtered further to only include those restaurants with some minimal threshold number of reviews (cutoff to be determined via EDA).

3.2 Analysis

Several types of analysis will be performed. First, an in-depth exploratory data analysis (EDA) will be conducted to understand which subset of the data is suitable for the objectives of this project. EDA will include summary statistics of various review and reviewer attributes, scatter plot matrix to evaluate correlation across attributes, and various summaries by distinct levels of geographical aggregation.

Next, cluster analysis of restaurants by their attributes will be performed to understand and label restaurants with similar review ratings. The objective of this analysis is to identify a set of labels or groups of restaurants for additional comparison and insight into what common traits result in a high rating from reviewers. Additionally, a regression analysis may be conducted to identify which attributes most strongly correlate with restaurant ratings.

Finally, sentiment analysis will be conducted on a subset of reviews to understand review sentiment trends over time. The sentiment trends will be compared to reviewer ratings in an attempt identify reviewers who repeatedly review restaurants before a subsequent material change in review sentiment trend.

3.3 Presentation

Presentation of analysis findings will be completed on a suitable BI platform such as PowerBI or Tableau. The BI platform visualization will enable user interaction with the analysis outputs. Additionally, charts will be constructed in Jupyter notebooks for inclusion in the project final report.

4 Data Set

The Yelp Open Dataset [4] contains nearly seven million reviews of over 150,000 businesses. The data is in JSON format, and includes information on business, reviews, and reviewers. The data also includes information concerning customer check-ins at business as well as photos associated with businesses. This project will focus on three tables:

- Business.json
    - This table includes metadata and review counts for each business. The metadata includes name, location,

amenities, and hours of operation. Additionally, there are some derived metadata such as business category tags.

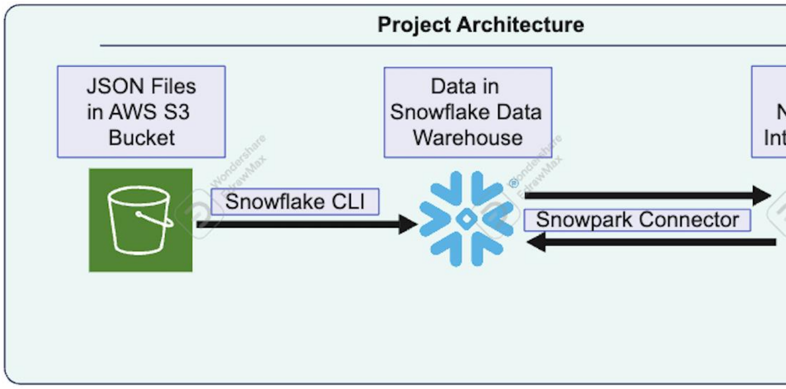
- Review.json
    - Each row in this table represents an individual review of a business. Each review includes a reviewer id, date, star rating, review text, and counts of any review accolades ("funny," "useful," or "cool").
- User.json
    - Each row in this table describes a reviewer. The attributes include a reviewer id, first name, review count, years in which the user was in the "Elite" category, number of fans, and summary information about their reviews. The review summary information lists and average star rating as well as counts of accolades received by their reviews.

5 Evaluation Methods

The evaluation of this project analysis will be via comparison of a train/test split in the dataset. The test set will be evaluated with the clustering and regression algorithms trained on the training set. Additionally, the output of the sentiment analysis will be compared to the similar methods analyses described in the Literature survey. Finally, this project will be peer evaluated by other teams and class members in CSPB 4502.

6 Tools

The tools for this project differed significantly from the original proposal. Some of this was driven by compute necessity and some by the author's desire to learn new tooling. Ultimately, the raw data in JSON format was placed in an AWS S3 bucket. From there it was ingested into tables in Snowflake via CLI. Finally, the analysis and app was completed in HEX. An architecture diagram of the project is as follows:



7 Milestones

- April 1, 2023: tools running and tested on dataset sample
- April 8, 2023: initial EDA completed, and filtering review count thresholds selected
- April 15, 2023: clustering and regression models running
- April 22, 2023: sentiment analysis tested and operable; clustering and regression complete/evaluated
- May 1, 2023: sentiment analysis completed; initial visualization completed
- May 4, 2023: draft report completed; code shared
- May 7, 2023: presentation complete

8 Final Project Summary

The project did not fully accomplish the proposed objectives. Both sentiment analysis and clustering was completed but both approaches could use additional refinement. I would consider them both proof-of-concept level with neither demonstrating particularly valuable signal.

8.1 Sentiment Analysis

The project relies on a pre-trained model called TextBlob. The performance of the model on the subject dataset is directionally correct. The scaling seems a little off, but correlation between star rating and sentiment is intuitively correct and highly correlated.

8.2 Clustering

The clustering analysis was K-Means. An elbow analysis of SSE vs number of clusters was conducted to arrive at the decision to proceed with three clusters. After completed the clusters, PCA analysis was conducted to build a 2D plot of clusters vs PCA1 and PCA2. This isn't easily interpretable but the clusters are visible in what would otherwise be too many dimensions to display.

8.3 Application

All of the analysis outputs are hosted in a HEX app. The app takes a non-trivial amount of time to initially load (~15-20 minutes), but is fully interactive afterwards.

References

[1] Jagan Gupta, 2018. *What's in a review? - Yelp ratings EDA, version 22.* https://www.kaggle.com/code/jagangupta/what-s-in-a-review-yelp-ratings-eda

[2] Bukun, 2018. *A Very Extensive Data Analysis of Yelp, version 147.* https://www.kaggle.com/code/ambarish/a-very-extensive-data-analysis-of-yelp

[3] Bonnie Nortz, Stephanie Mallard, 2016. *CS 229 Final Project Report: Using Yelp Reviews to Improve Businesses.* http://cs229.stanford.edu/proj2016/report/NortzMallard-UsingYelp%20ReviewsToImproveBusinesses-report.pdf

[4] Yelp Open Dataset. https://www.yelp.com/dataset

[5] DuckDB https://duckdb.org/

[6] Polars https://www.pola.rs/

[7] Hex https://hex.tech/

[8] Palantir Foundry https://www.palantir.com/platforms/foundry/