Yelp Business Reviews Group 5

Big Data Management

Sanoj Anjamkudy

Bala Ramalingam

Andrew Porter

**Project Overview:**

Group 5 analyzed two data sets from the Yelp application. Yelp is a mobile app that lets customers review business and share their review with the Yelp community. Yelp is a popular tool used all around the world for visitors in new locations that may want to search for something specific or general.

**Problem Statement:**

With available facilities/features for a business we will predict the review that we might get for any business we are going to start. Build a model to suggest the parameters which helps to improve the review of any business.

**Data at a ‘Glance’:**

The first data set that was analyzed was the business data set. This data set contains 27 variables and 174,567 rows or businesses. This data set contains attributes for each business for example, type of business, location data, does the business have a parking lot, does the business accept credit card and the overall average rating of that business along with specific attributes each rows also contains an unique business id that will be used to join on the yelp reviews data set. The second data set is individual reviews for each business. The data set contains features like was the review useful, cool, or funny and the review text. The group partitioned the business data set into a restaurant in USA, non-restaurants in USA, restaurants outside of USA, and non-restaurants outside of the USA.

**Exploratory Data Analysis:**

The Group first looked at all business locations and discovered that the major of the businesses was within the United States. Inside the US, most of the business locations are in Arizona and Nevada. After looking at the locations, the group focused on the type of business in the data set. The business with the most reviews is restaurants. The locations of these restaurants are primarily in Nevada and Arizona. Nevada has an average rating slightly higher than 3.4, while Arizona is just lower. The lowest rating is in Arkansas and the highest rating is in Wisconsin. The restaurant with the most 5-star ratings is Mon Ami Gabi while the restaurant with the most 1-star ratings is the MGM grand hotel. Next the group looked at the review data set to understand what restaurants received the most useful, cool, and funny reviews. Hash House A Go Go is consistently at the top, having the most useful and cool ratings. McDonald’s was at the top of having the funniest reviews.

The team calculated the correlation coefficient of features as they relate to the stars. Parking lot has the highest correlation at .127 while Ambience has the lowest at .01.

**Models:**

**Unsupervised Models:**

The group conducted a series of clusters models to understand if there were any like groups of data.

K-means clusters, Bisecting K-means were executed. The K-means cluster created 6 clusters ranging in size from 11,413 to 22,505. Based on the summary a cluster was created for businesses with low ratings, low chance of a parking lot, low price range and another cluster was creating for businesses with high ratings, mid-level pricing, over relatively good chance of a parking lot. The Bisecting K-means algorithm created 6 clusters ranging from 12,993 to 25,320. A cluster exists for low stars, low price, and low chance of a parking lot, similar to the K-means cluster. The Bisecting cluster also decided that for a grouping of above average stars ratings, the business needed to have mid-level pricing, but did hold as much weight to the parking lot as the k-means.

**Supervised Models:**

The team decided to take a couple different approaches for the predictive section of the project. First, the team ran a series of regression models to predict the reviews and then ran a series of classification models to predict if the review was good or bad. The team also evaluated a predictive classification model to determine if a restaurant was open or closed.

For the regression models, the team made use of both datasets provided. The restaurant statistics and the review statistics. The result was 18 variables to predict number of stars given to the restaurant. All the data was converted into integers and most are binary (1 – yes or 0 – no). For example, if a restaurant has restaurant hours on Tuesday, the team turned that data element into a 1, if there were no hours provided for the restaurant on Tuesday, the team turned that data element into a 0. From here, the data was prepped and ready to split into a 70% training (2,098,145 rows) and 30% testing (898,787 rows) data set.

4 Regression models were executed and evaluated (Linear Regression, Decision Tree, Random Forest, and Gradient Boosted Tree). All models performed relatively the same, the GBT had the lower RMSE, but if the team had to decide on the best model. The team would lean towards the less complex.

4 Classification models were executed and evaluated to predict if a restaurant would get a good or a bad review (Logistic, Decision Tree, Random Forest, and Gradient Boosted Tree). Of all the models, the best performing in terms of accuracy is the GBT, but between the best performing and worst performing there is only a 5% difference.

4 Classification models were executed and evaluated to predict if a restaurant is open or closed (Logistic, Decision Tree, Random Forest, and Gradient Boosted Tree). Of all the models, the best performing in terms of accuracy is the Decision Tree at 87.1 %. There is a significant drop off for the more complex Random Forest and GBT models.

**Association rule mining:**

Note: The Association rule was performed on just restaurants initially and did not yield useful results, so the team opened the analysis to all businesses to see if there was any cross categorical trends.

An association rule model was performed to understand if there are trends in if a business is reviewed, what is the chance of another business reviewed? Based on the results (assuming a review is similar to a visit) it seems like the business that individuals frequent often raise to the top for example, target, post office, whole foods and the consequent is Starbucks. Which makes sense because there many reasons to go to these stores and a lot of individuals like to get their coffee before running errands.

**Collaborative filtering:**

The group combined both data sets and transformed the data into users as rows and business as columns in prep for the collaborative filter analysis. The collaborative filter needs integers or numbers for rows and columns, so the group had to get creative with the transformation. But the result was predicting the likelihood of one user reviewing another business because of having something in common with the user that reviewed that second business. The next steps for this analysis would be to re-map the integers assigned to the user ids and the business ids to understand where the trends are.

**Summary:**

The group analyzed over 5 million review from 175k businesses by using HIVE portioned tables, spark data frames, exploratory data analysis, unsupervised and supervised learning models. The group decided to focus on US based restaurants for most of the analysis and try to determine if give a set of attributes could the group predict the average number of stars. Based on the analysis, if a business has a mid-level priced menu, parking lot, counter service the like hood of the review will be good.