Yelp Data Analysis for a new investor in Arizona

Qiqi Zeng

1. Introduction

As Yelp has collected a large number of business data, it can provide very insightful advice for business investors, especially for those new investors with limited experience. The aim of this research is to offer some data-based commercial guidance for a new investor in Arizona, trying to help them improve their customer satisfaction and get high ratings on Yelp.

In this research, we focus on the business in the Arizona, and there are five research questions included. *Q1*: What is the rating trend of the most popular industries in recent years? Which industries are the most worth invested in? *Q2*: Which industry attract the most influential reviewers in Arizona? *Q3*: Which attributes are greatly related to the user ratings for the top businesses in Arizona? *Q4*: Having defined some business features, what is the predicted rating for this specific business? *Q5*: What is the social network of Yelp users in Arizona? Who should be invited to the new business to help attract more customers after shop opening?

The dataset used in this research is collected from the Yelp Dataset Challenge. Most of the research questions are based on the dataset of version 8. The version round 13 is only applied to train the predictive model in Q4.

2. *Methodology*

Q1: What is the rating trend of the most popular industries in recent years?

This research question is designed to provide new investors an overview of the recent market trend and the changes of customers' attitudes towards different industries, helping them assess the industries that they want to invested in.

The tool used for this question is Python3. To simplify this problem, we only focus on the most significant and typical industries. Therefore, we counted the number of existing business in different 'categories' using the Business table, and then we picked out the top 6. To evaluate the rating changes over years, we merged the Review table to business table as only the review table records the rating times. We selected and matched these reviews in the Review table to the 6 categories picked out in the last step. Calculating the average ratings of each categories for each year, we plotted out the business rating trend from 2004 to 2016.

Q2: Which industry attract the most influential reviewers in Arizona?

The aim of this question is to find out which industry can attract the most influential reviewers, who can help them advertise. Investing in these industries, a new investor may easily gain a large number of customers in a short time and start his business more quickly.

As our research is aimed to provide information for Arizona investor, we needed to pick out the reviewers living in Arizona, using the assumption that a reviewer is Arizonian if most of the businesses he reviewed is in Arizona. Therefore, we firstly use Business table and Review table to select out those potential Arizonian who have ever reviewed for Arizonian business. For each of them, we calculated the proportion of his reviewed businesses in Arizona to check if he is exact Arizonian. In addition, for each Arizonian, we also checked if his friends live in Arizona, and removed the non-Arizonian friends so that we can exam his influence in Arizona. After that, using the processed data, we chose the top 1000 influential reviewers according to the number of their friends. With the Business table, we calculated the number of influential reviewers in each Arizonian industry and normalized them by dividing the total business number in that industry. Finally, we plotted out the average influential review counts for a single business in the top 10 popular industries.

Q3: Which attributes are greatly related to the user ratings for the top businesses in Arizona?

Once an investor has decided the investing industry, it is important to know what business attributes they need to consider to improve business performance. Based on the result of Q1 and Q2, we assume the investor decide to invest in Restaurant in this case.

From the Business table, we selected out the business of 'Restaurants' in Arizona among all business, extracted the 'attributes' and normalized the semi-structured JSON data into a flat table, and then we dummied these variables. According to Dubey (2018), the Random Forest, which is one of the most famous machine learning models, is easy to compute how much each variable is contributing to the decision. Random Forest consists of hundreds decision trees, which built over a random extraction of the observations from both the dataset and features. At each node of the decision tree, the dataset is divided into 2 buckets, and each of them get those observations with similar response. Thus, the feature importance can be calculated from how "pure" each of the buckets is. Therefore, we used the processed data to fit the Random Forest model by applying the package 'RandomForestClassifier'. Finally, we picked out the important attributes and visualized them. In addition to Restaurant, we also applied the same method to conclude the attributes importance for Beauty&Spas, which is a recommended industry in Q1.

Q4: What is the predicted rating for a specific business?

This research question aims to build a predictive model for new investors to predict customer feedback based on some known attribute so that they adjust their strategies before opening.

Different from Q3, we used the whole dataset, which include all business categories and all states, to train the Random Forest model. Thus, this model is available for various prediction cases. Similar to Q3, we applied the Random Forest Model, and we used the AUC-ROC to measure model performance. According to Narkhede (2018), higher the AUC, better the model is. The AUC of a perfect model is near to the 1. Having the trained model, we can put the known states, city, business category and attributes in to the model to get a predicted rating. In this case, as we research on Arizona, we would put the 'AZ', 'Restaurants'/'Beauty&Spas' and the important attributes got from Q3 into the models.

Q5: Who should be invited to the new business to help attract more customers after shop opening?

After opening the store, this research question can help investors attract customers quickly by inviting the famous reviewers to come to his store and advertise for him.

To find those famous reviewers in Arizona, we firstly use the business table to pick out the Arizonian reviewers with Arizonian friends, using the same method in Q2. Then we built up a network graph based on this relationship. To explore this relationship, we calculated the degree distribution, which test how many friends a single reviewer have. Closeness centrality was also calculated to evaluate how fast information travel, using the formula: $C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}$, where d(v,u) is the shortest-path distance between v and u, and n is the number of nodes in the graph. In addition, we used betweenness centrality $B_i = \sum_{k>j\neq i} \frac{\#\ of\ geodesic\ paths\ between\ k\ and\ j\ passing\ through\ i}{\#\ of\ geodesic\ paths\ between\ k\ and\ j}$ to measure the centrality and

relevance of a single reviewer. Furthermore, we used the eigenvectors $x_i = \lambda^{-1} \sum_j A_{i,j} x_j$ to evaluate the importance of a reviewer ($A_{i,j}$ is an adjacent matrix, x_j is the importance of reviewer j). All of these four measures are calculated by the python package networkx, and was visualized by Gephi.

3. *Results*

Q1: What is the rating trend of the most popular industries in recent years?

The top 6 popular industries in Arizona are Restaurants, shopping, Home Service, Automotive, Beaty&Spas and Health Service. As these industries are most concerned by business investors, the rating trend analysis was only focused on them.

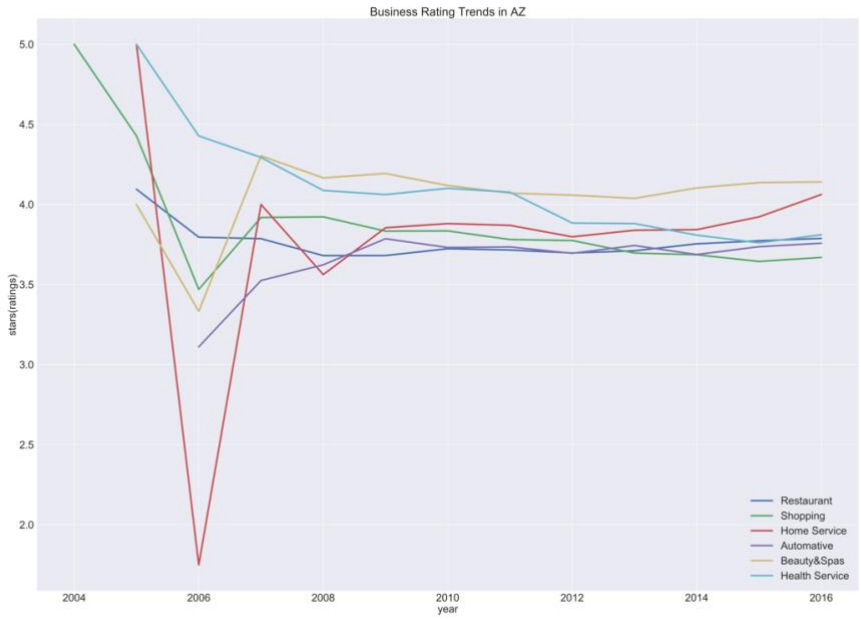

Figure1: Business Rating Trends in AZ

As this figure illustrated, the rating trend for Beauty&Spas remains in the highest level of average rating since 2007 and it increases slightly from 2013 to 2016. The average rating of Home Service drops radically in 2006 and rises back to the normal level in 2007. Then, after 2008, it shows a sustained upward trend, becoming the most prominent industries out of 6 industries in Arizona. The rating trend of Restaurant and Automotive grow marginally after 2008. Only 2 industries, Health Service and Shopping, showed a downward rating trend.

Q2: Which industry attract the most influential reviewers in Arizona?

As the figure below shows, a business in Nightlife and Bars receives a highest average reviews count (around 9) from influential reviewers in Arizona. The business in Restaurants also receives a quite high count at around 7. All the businesses in the other industries received less than 4.

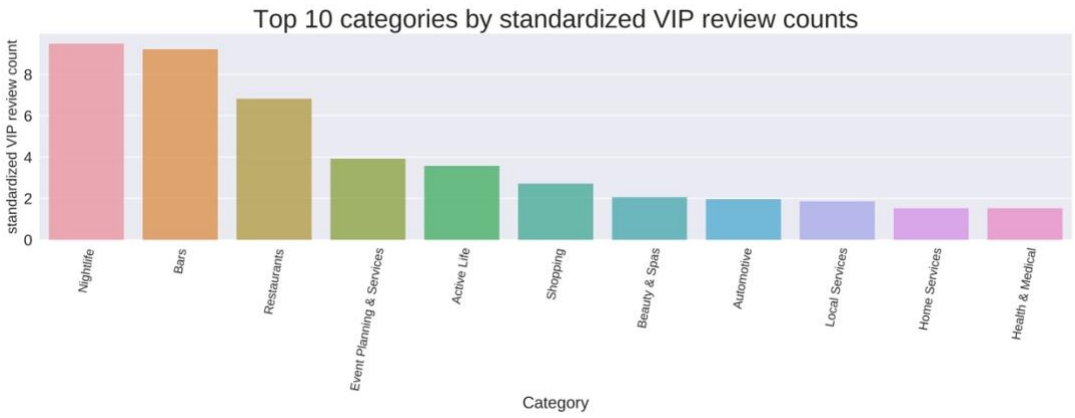

Figure 2: Top 10 industries of receiving reviews from influential reviewers

Q3: Which attributes are greatly related to the user ratings for the top businesses in Arizona?

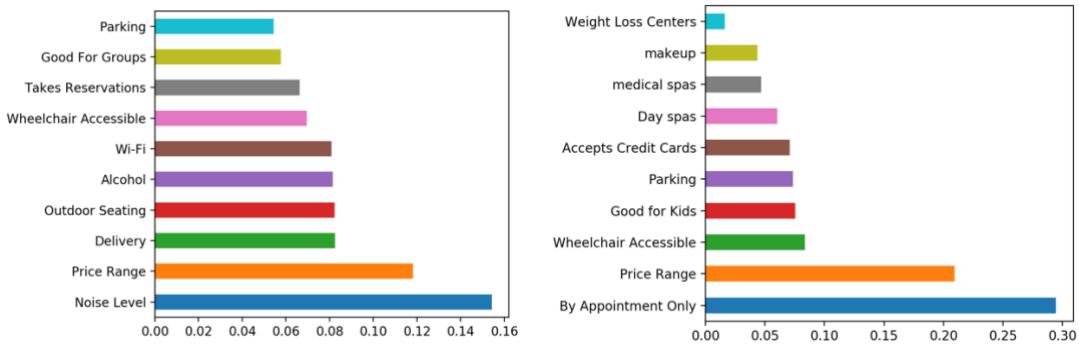

Figure 3: Feature Importance for Restaurants and Beaty&Spas

The left figure shows the feature importance for Restaurants, and it shows that the most important features are 'Noise Level', 'Price Range' and 'Delivery'. The feature importance for Beauty&Spas showed on the right illustrates that the attribute 'By Appointment Only', 'Price range' and 'Wheelchair Accessible' related most to the rating.

Q4: What is the predicted rating for a specific business?

The AUC of ROC for this Random Forest model is 0.56. Based on the result of Q3, we assume the investor would open a restaurant in Arizona, and he would consider providing a quiet place, a reasonable price range, a delivery service, outdoor seating, a full bar and free Wi-Fi. Then the predicted rating for his restaurant is 4.0. However, if he provides none of the attributes listed above, he might only get a rating of 2.5. In another case, an Arizonian investor would like to open a Beauty&Spas salon, and he would serve his customers by appointment only with appropriate service charges. He would provide wheelchairs and kids watching service. The predicted rating for him is 5.0. If he provides none of these attributes he might get an predicted rating of only 3.5.

Q5: Who should be invited to the new business to help attract more customers after shop opening?

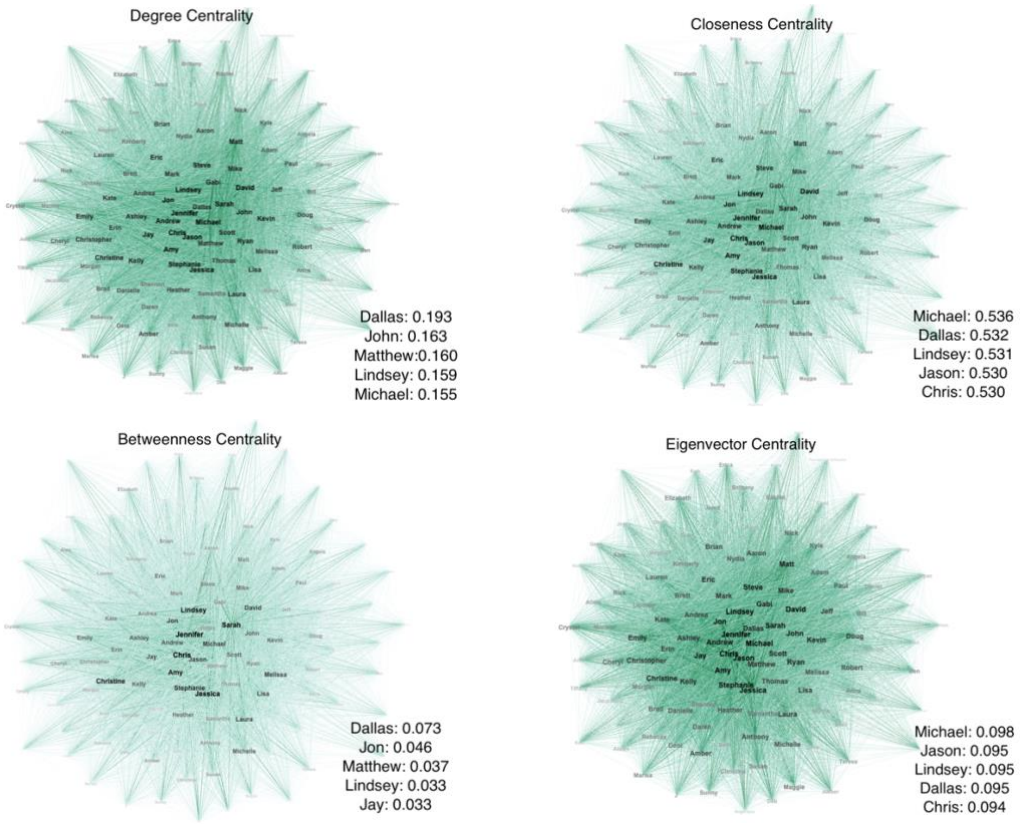

Figure 4: Social Network

From this figure, it can be concluded that Dallas and Lindsey are the most influential persons in Arizona as they appeared on four measures at the same time. John has a very high degree centrality, but he is not prominent in the other three measures. Similarly, Jon and Jay are only outstanding in betweenness centrality.

4. Discussion & Conclusion

By analyzing the Yelp data, we concluded several advices for a new investor in Arizona.

From the result of Q1, it can be shown that Beauty&Spas is the most stable and promising industry since its customer feedback is the most positive for most of the times in the past few years. Though the average rating of Home Service grew fastest quickly in last few years, it is possible to dramatically fluctuate again like that in 2006. Therefore, if an investor expects a long-term revenue, the investment in Home Service should be careful. In addition, investing in Restaurant and Automotive can also be a safe choice for investor in Arizona. The result of Q2 shows that Nightlife and Bars industry can attract the most influential reviewers, and Restaurants can also easily attract these famous reviewers. Therefore, we speculate that an investor can consider these industries if he wants to gain a large number of customers in a short time.

The results of Q3 suggest the attributes that a Restaurant investor should consider, such as 'Noise Level', 'Price Range', 'Delivery' and so on. Once he applied those attributes to his restaurants, he is more likely to get a high rating (4.0 using the predictive model in Q4). However, if he ignores those attributes and applies none of them, he might get a lower rating (2.5 in Q4), which might be a bad signal in terms of user feedback and it is not conducive to long-term revenue. The same methodology can also be applied to analysis the important attributes for other industries.

The Q5 analyses the network of Yelp users and concludes who should be invited to the new business to help attract more customers. All the reviewers appeared on figure 4 can be their choices. Among them, Dallas and Lindsey would be their first choice as they both have good performances on four measures. Furthermore, if investors want to disseminate business news quickly, they should consider the high closeness centrality group. To spread ads among famous Arizonina reviewers, they shall think about the high eigenvector centrality group. If investors just want more people get to know their businesses, they may invite the reviewers with high degree centrality, such as John.

Word Count: 1945 (exclude headings)

5. Reference

Dubey, A. (2018) “Feature Selection Using Random Forest”, *Towards Data Science* [online] Available from https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f (Accessed: 28th Mar 2019)

Narkhede, N. (2018) “Understanding AUC - ROC Curve”, *Towards Data Science* [online] Available from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Accessed: 28th Mar 2019)