Modeling Restaurant Success and Failure Using Yelp and Median and Mean Household Income Data

Taylor Gunter
University of Colorado- Boulder
taylor.gunter@colorado.edu

Keaton Whitehead
University of Colorado- Boulder
keaton.whitehead@colorado.edu

Matthew Coker
University of Colorado- Boulder
maco9370@colorado.edu

Devin Arnold
University of Colorado- Boulder
dear0350@colorado.edu

ABSTRACT

This paper is a Project Proposal for group 18-- Yelp Us! for Spring 2018 CSCI 4502 Data Mining. The paper outlines the semester project, and includes sections describing the project motivation and objective, information about the data sets, and a summary of the previous work done on the proposed topic. This paper will also summarize the proposed evaluation methods, the proposed tools that will be used to implement the methods, proposed milestones for project completion, and a summary of the peer review session which took place during the week of February 26th – March 2nd in class.

KEYWORDS

Data Mining, Yelp, Data, Income, Mean, Median, Modeling, Regression, Classification, Clustering, Cleaning, Processing, Python, Visualization, Failure, Success.

MOTIVATION AND OBJECTIVE

The restaurant business is notoriously difficult. Margins are slim, risk is high, and people are picky about their food, but over time many restaurants do find success. We are setting out to determine what makes a restaurant successful, and conversely not successful. Are there objectively better ways to run a restaurant to increase the probability of success or is the secret just in the sauce?

Our objective for this project is to prove that successful restaurants have more in common than just good food, and unsuccessful restaurants have more in common than just bad food. In order to accomplish this, we must determine what successful restaurants have in common with other successful restaurants, likewise we must do the same for unsuccessful restaurants. Additionally, we will integrate mean and median household income for all zip codes in the United States and use that data as an attribute to help in our discovery process.

DATASET SUMMARY

The primary data sets we are using are provided by Yelp and Kaggle. The data set that comes directly from Yelp is downloaded to Taylor Gunter's computer as a set of SQL tables. The data set from Kaggle is a set of .csv files, and is also downloaded to Taylor Gunter's computer. Yelp also provides the data in a JSON format, which we will use to implement map visualizations if the project timeline permits.

We will be working with the .csv version of the data from Kaggle. There are seven files in the set, they include: business info, business hours, reviews, tips, user data, check-in values, and a summary table that contains Boolean values of the attributes that each tuple contains. We will be using the business info table as our primary source of data, and the rest of the tables will be supplemental. The business info table has thirteen attributes. Six of the thirteen attributes contain location information; one attribute is a numerical rating; one attribute is the number of reviews; one attribute is a Boolean open or closed value, where 1 maps to open, and 0 maps to closed; one attribute is the

Business name; one attribute is the category, which is a brief description of the type of business; and one attribute is a unique identification key, which will allow us to join businesses across tables.

The secondary data set we are using is provided by the University of Michigan, and is a three attribute .csv file. This dataset can easily be integrated with the Yelp .csv files since both tables share 'zip code' as an attribute. The other two attributes are 'median income', and 'mean income'. The data sets can be found at:

- https://www.kaggle.com/yelp-dataset/yelp-dataset
- https://www.yelp.com/dataset
- https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/

PREVIOUS WORK

Since a source for our data set is Kaggle, previous work is abundant. I will summarize two articles in this proposal. The first article is *Using Yelp Data to Predict Restaurant Closure* by Michail Alifierakis, who is a PhD Candidate at Princeton University. The second article is from Stanford University. It is titled *Predicting New Restaurant Success and Rating Using Yelp,* and is written by Aileen Wang, William Zeng, and Jessica Zhang.

Using Yelp Data to Predict Restaurant Closure was written with the motivation to provide banks and investors with a model to predict whether loans and investments in a particular restaurant are a good idea. The author initially attempted to use the Yelp rating system to predict restaurant success. Yelp's rating system proved to not be useful however, because the data was not easily normalized. For this reason, the authors of the paper chose attributes including whether a restaurant was a part of a chain, the density of local restaurants in the area, the number of reviews, star rating, price relative to other restaurants in the area, and finally the age of the restaurant. The primary issue the author faced in performing his analysis was strong (accurate) prediction of a restaurant staying open, and weak prediction of a restaurant closing. The author found the most important attribute leading to success was whether a restaurant was a part of a chain, and the most important attribute leading to failure was restaurant density in an area, where restaurants in restaurant-dense areas were more likely to close. The attribute the author wishes he would have used is population demographic, which is something we will be able to determine with our secondary data set describing mean and median household incomes.

The second article is *Predicting New Restaurant Success and Rating Using Yelp* by Wang et al. This paper is by far the most academic, and describes many machine learning methods we had never heard of, but they do use chi-squared testing which was introduced in class. An unexpected benefit of this paper is it's analysis and use of previous work, so we multiple sources we were unaware of. Overall, this paper is very impressive, and will make sense of multi-characteristic attributes, and how to concatenate them. For example, there are attributes that can be converted to binary, then converted to decimal numbers to allow for east encoding. This paper found parking, attire, and ambience to be the most heavily weighted attributes in their chi-squared testing, and the one's that deserved the most attention.

It is interesting to see the differing methods both papers employ, how they approach the problem, and how their results are affected. The most significant commonality between the papers is that the Yelp data set is a poor training platform, and classification (open or close) accuracy is low for both papers.

PROPOSED WORK

The proposed work for our project will be broken into four steps: Data Cleaning, Data Preprocessing, Data Integration, and Data Evaluation. The Data Cleaning process will consist of discarding sparse tuples, discarding non-restaurant tuples, and discarding restaurants with an insufficient number of reviews. Originally, we wanted to set the required number of reviews at 30 with hopes that we could appeal to the *Central Limit Theorem* when looking at distribution models and confidence intervals, but upon cursory investigation of the data it appears many businesses would not meet the 30 review limit, so at this point we will consider all restaurants in the data set, but we are open to setting a minimum number of reviews further on in the process.

After data cleaning is complete we will begin data preprocessing. The primary objective in this portion of the work will be to create consistency rules, fill in missing attributes, delete non-essential attributes, and convert data points that are not type consistent to the correct type. An example of a consistency rule is if a

tuple has attributes that say 'Cash Only' = True and 'Credit Cards Accepted' = True, then the tuple is inconsistent, and will need to be processed. An example of deleting non-essential attributes is to eliminate all location data except zip code. This will make the data easier to process and eliminate the potential for computation mistakes, or coding errors. An example of processing an attribute to the correct type is in price—some tuples may list price in a dollar range, and some my list price symbolically with a varying number of '$' symbols. That symbol would need to be mapped to an integer value price range, or a string qualifier such as 'low', 'medium', or 'high'. The final preprocessing component will be in more obvious type conversions i.e., true or false to 1 or 0.

The data integration portion of the project will be to combine our primary and secondary data sets. As stated previously, the integration process will not be complex because the sets share zip code as a common attribute. The primary drawback of mapping the zip code from the secondary data set to the zip code attribute of the primary data set is that there will a lot of redundancy i.e., the data set will not be normalized, but it will be easier to process, and model restaurant success and failure.

The final step in our proposed work is data evaluation. We will begin by setting our dependent variable as the open or closed Boolean value 0 or 1. We will then perform several methods on the data. We will perform decision tree induction, which we learned about in class on March 5th. This will help with classification, where we can build out a feature set that gives us the highest accuracy of labeling before returns diminish. Another method we can use is support and correlation, which was a primary component in homework 3. If we treat restaurant attributes as market basket items, then the translation of the methods from homework to project is more straight forward. As we learn more about clustering we think that it will be a useful way to model the restaurant data. Our discovery may not result in black and white answers, rather the data mining process may result in a spectrum of probabilities. For example, Tibetan restaurants may be successful in high-income and low-income neighborhoods, and not successful in middle-income neighborhoods. From that we can develop hypotheses, and test the hypotheses using Null Hypotheses tests and confidence interval--"I am 95% confident that the median income where successful Tibetan restaurants are located is high, or low, but not middle".

The proposed work on the project should be split 15% cleaning, 30% preprocessing, 5% integration, and 50% evaluation. We expect the statistical analysis methods to be the most difficult to perform correctly, and the classification methods to be the most difficult to implement, and interpret correctly.

TOOLS

We will be using common data mining tools as we are setting out to determine what makes a restaurant successful. Our primary programming language will be Python 3, and we will use Jupyter Notebooks for quick data proto-typing. We will use the Pandas library for our data cleaning, preprocessing, and integration. We will also use the Pandas library for easy statistical summary measures such as mean, median, max, min, etc. All of our statistical analysis will be aided with the SciPy library. We would like to include visualizations in our project, and the Matplotlib library provides good support for scatterplot visualizations, histograms, linear visualizations and regression modeling. We are also considering using d3, which is a javascript language that makes interactive visualizations relatively easy to create. If we do not use d3, then we will use Tableau to collect visualizations and produce a summary dashboard that will help us tell a clear story about the data and what we discovered.

MILESTONES

We are currently in the eighth week of the semester, and we will present the project in the sixteenth week, so that gives us eight weeks to complete the project. We are projecting cleaning, preprocessing, and integration to be completed before Spring Break, which is March 23rd. This gives us three weeks to complete those three steps. We expect preprocessing to take more time than cleaning and integration, so it should get the bulk of the attention. Based on this, cleaning should be done by the end of this week – March 11th, preprocessing should be done by Tuesday March 20th, and integration should be done by Friday March 23rd. That leaves four weeks when we return from Spring Break to complete evaluation, visualization, and write-up. Evaluation will require the most amount of time in that frame, so it should be finished by April 15th. Visualization should be

completed by April 21st, and the project should be polished and ready to present by Monday April 29th.

PEER REVIEW FEEDBACK

We presented on Friday, and had to rush through the presentation, and did not receive much feedback, but we did learn a lot from the feedback other groups received. Our primary takeaway is that we really need to dive into the previous work done on this topic and use it as a litmus test against our process and conclusions. We are lucky in that our questions have specific answers from reputable sources. Restaurant success is written about in an academic article from Stanford, and restaurant failure is written about from a PhD candidate at Princeton. We can dive deeper into technique, and learn from what other people have already done.

Another piece of feedback that was common to the class was that questions were too broad. We're not writing a PhD thesis, and spending years on research. In order to produce coherent results we need to ask very clear, and focused questions. This will be beneficial as a time saver and help to keep the project on track. The final piece of feedback that we will apply is to communicate with other groups who are using the same data as we are. Luckily, there is one other group who is using the same data set, so we will need to connect with them and discuss shared problems, and ways to solve them. Fortunately, our data is supplemented with a secondary data source which will narrow the scope of our project and differentiate from theirs, so too much influence over each other's projects should not be an issue.