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

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

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

Matthew Coker
University of Colorado- Boulder
maco9370@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 tools that will be used to implement the methods, milestones for project completion, and sumamry of the results so far.

KEYWORDS

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

MOTIVATION AND OBJECTIVE

Client-service businesses are notoriously difficult. Margins are slim, risk is high, and customers can always find another establishment that offers the same service, but over time many businesses do find success. We are setting out to determine what makes a business successful, and conversely unsuccessful. Are there objectively better ways to run a business to increase the probability of success or is the secret just in the service?

Our objective for this project is to prove that successful businesses [1] have more in common than just good service, and unsuccessful business have more in common than just bad service. In order to accomplish this, we must determine what successful businesses have in common with other successful businesses; likewise we must do the same for unsuccessful businesses. 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.

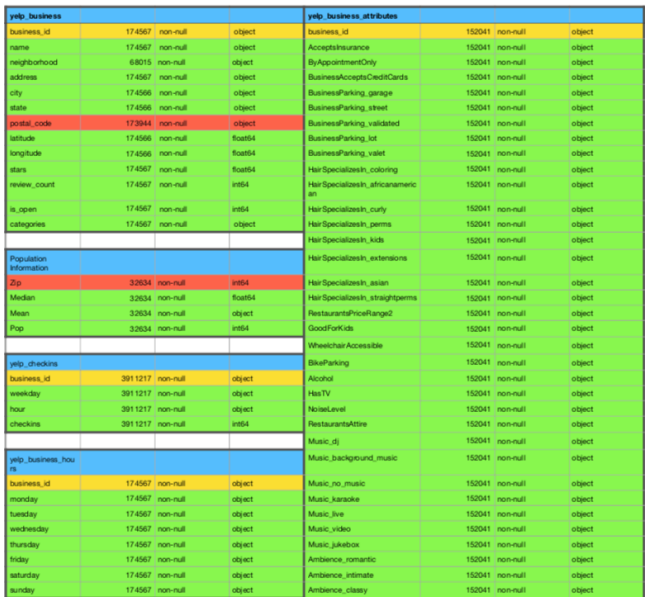
DATASET SUMMARY

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. Our primary data will come from the following files: yelp_business.csv, yelp_business_attributes.csv, yelp_checkins.csv, and yelp_business_hours.csv. The secondary data set we are using is provided by the University of Michigan and is titled MedianZip-3.xlsx. The MedianZip-3.xlsx file has been integrated with the yelp_business.csv file to create a file called yelp_money_merge.csv. The

[1] The scope of our project has changed because the number of restaurants in the dataset will not be sufficient for the project requirements.

following is a screenshot of the data organization, which illustrates the data attributes and file sources for this project. In this visualization the yellow highlighted rows are matching keys, and red highlighted rows are matching keys. The yelp_business_attributes.csv table is truncated to save space.



yelp_business			
business_id	174567	non-null	object
name	174567	non-null	object
neighborhood	68015	non-null	object
address	174567	non-null	object
city	174566	non-null	object
state	174566	non-null	object
postal_code	173944	non-null	object
latitude	174566	non-null	float64
longitude	174566	non-null	float64
stars	174567	non-null	float64
review_count	174567	non-null	int64
is_open	174567	non-null	int64
categories	174567	non-null	object
Population Information			
Zip	32634	non-null	int64
Median	32634	non-null	float64
Mean	32634	non-null	object
Pop	32634	non-null	int64
yelp_checkins			
business_id	3911217	non-null	object
weekday	3911217	non-null	object
hour	3911217	non-null	object
checkins	3911217	non-null	int64
yelp_business_hours			
business_id	174567	non-null	object
monday	174567	non-null	object
tuesday	174567	non-null	object
wednesday	174567	non-null	object
thursday	174567	non-null	object
friday	174567	non-null	object
saturday	174567	non-null	object
sunday	174567	non-null	object

yelp_business_attributes			
business_id	152041	non-null	object
AcceptsInsurance	152041	non-null	object
ByAppointmentOnly	152041	non-null	object
BusinessAcceptsCreditCards	152041	non-null	object
BusinessParking_garage	152041	non-null	object
BusinessParking_street	152041	non-null	object
BusinessParking_validated	152041	non-null	object
BusinessParking_lot	152041	non-null	object
BusinessParking_valet	152041	non-null	object
HairSpecializesIn_coloring	152041	non-null	object
HairSpecializesIn_africanamerican	152041	non-null	object
HairSpecializesIn_curly	152041	non-null	object
HairSpecializesIn_perms	152041	non-null	object
HairSpecializesIn_kids	152041	non-null	object
HairSpecializesIn_extensions	152041	non-null	object
HairSpecializesIn_asian	152041	non-null	object
HairSpecializesIn_straightperms	152041	non-null	object
RestaurantsPriceRange2	152041	non-null	object
GoodForKids	152041	non-null	object
WheelchairAccessible	152041	non-null	object
BikeParking	152041	non-null	object
Alcohol	152041	non-null	object
HasTV	152041	non-null	object
NoiseLevel	152041	non-null	object
RestaurantsAttire	152041	non-null	object
Music_dj	152041	non-null	object
Music_background_music	152041	non-null	object
Music_no_music	152041	non-null	object
Music_karaoke	152041	non-null	object
Music_live	152041	non-null	object
Music_video	152041	non-null	object
Music_jukebox	152041	non-null	object
Ambience_romantic	152041	non-null	object
Ambience_intimate	152041	non-null	object
Ambience_classy	152041	non-null	object

Figure 1. Data Summary

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. We will summarize two articles in this proposal. The first article is *Using Yelp Data to Predict BusinessClosure* by Michail Alifierakis, who is a PhD Candidate at Princeton University. The second article is from Stanford University. It is titled *Predicting New BusinessSuccess and Rating Using Yelp,* and is written by Aileen Wang, William Zeng, and Jessica Zhang. Additionally, there are several presentations on data mining techniques used specifically for this data, and we will summarize notable techniques found in the presentations. The first presentation is titled, "Data Crackers on Yelp Dataset". It was created by Prashanth Sandela, Vimal Gorijala, and Parineetha Gandhi. The second presentation is titled "Business Analysis Based on Location and Category". It was created by Keyur Mandani, Mikaelian Ovanes, and Hemanth Reddy.

Using Yelp Data to Predict BusinessClosure was written with the motivation to provide banks and investors with a model to predict whether loans and investments in a particular businessare a good idea. The author initially attempted to use the Yelp rating system to predict businesssuccess. 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 businesswas 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 businessstaying open, and weak prediction of a businessclosing. The author found the most important attribute leading to success was whether a businesswas a part of a chain, and the most important attribute leading to failure was businessdensity 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 BusinessSuccess 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.

The "Data Cracker on Yelp Dataset" presentation employed Naïve Bayes on user review data and classed the data into positive, neutral, or negative reviews. They also used decision tree induction, which they found to be useful for all types of attributes including, nominal, numerical, and ordinal. The decision tree was implemented in WEKA, which is not something we

considered using, but will require further investigation, and will probably be implemented in our project. The final strategy this presentation covered was KNN and was also implemented using WEKA. This presentation provided good directional support for our project, but it will not be helpful as a results confirmation tool, because the work analyzed user review data using NLP techniques to extract information, which is outside of the proposed scope of our project.

The second presentation titled, "Business Analysis Based on Location and Category" focused on geo data and provided insight into ways to categorize the data. The group who created this presentation worked with SQL data, so their analysis was query based, but they were able to categorize the type of business more easily than pushing the data through a python parsing package. The most interesting outcome of this presentation is a line graph visualization that shows average 'star rating' by business category. There are two distinct low-rating outliers—acupuncture, and bingo halls, and two distinct high-rating outliers—auto detailing, and bartenders. Sentiment analysis based on user- review would be an interesting application to apply to outliers like these. If nothing else, the potential hypotheses that may be generated are interesting, but may not be testable, and could lead to unfounded speculation.

It is interesting to see the differing methods the papers and presentations 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. T

PROPOSED WORK AND INSIGHTS

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-business tuples, and discarding businesses 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 businesses in the data set, but we are open to setting a minimum number of reviews further on in the process. As we have worked through the data we have concluded that the scope of the project needed to expand from just focusing on restaurants to focusing on all businesses. There is not sufficient data to only focus on restaurants. Additionally, we have found that the data is clustered around nine major cities in North America, which include: Phoenix, Los Angeles, Las Vegas, Charlotte, NC, Pittsburgh, Montreal, Urbana-Champaign, Toronto, and Madison. Since, the data is clustered around specific areas, we may be able to pull additional supplemental data if some categories need more support.

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.

We found the proposed data preprocessing to be unnecessary. All data appears to be consistent, and all attributes are usable. The most significant challenge regarding the data will be deciding how to use NaN values. Luckily, the NaN values are most abundant in binary attributes, so they have been replaced with zeros in preprocessing.

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 business success and failure.

We learned Pandas as a very handy merge function that mapped the zip code data to the business attributes data cleanly. The biggest challenge was in data type matching. Business attributes were objects while zip code data numerical, so we converted the zip code data to objects, then strings specifically. We then implemented an outer join, which merged all columns from both data sets on the zip code attribute, and created a new data set.

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 business 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 business 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".

We found that binning the data on attributes will help us decide which attributes should be considered candidates to split on. This can be seen further down in the results so far section.

We will also be implementing KNN on either check in frequency, income distribution by region, a candidate yet to be determined.

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 business 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 were considering using d3, which is a javascript language that makes interactive visualizations relatively easy to create, but decided against it. Instead, we will use Bokeh, which is a python library that allows for easy interactivity creation. Bokeh can output an html file that has all the isualization data encoded in it, so it lightweight, and can be hosted in a github.io dashboard. Outside of the typical pandas libraries we are going to use WEKA for KNN and decision tree analysis. No one in the group has used WEKA, so there will be learning curve blocker, but there are several tutorial resources available. IBM provides good introductory material, and can be found at https://www.ibm.com/developerworks/opensource/library/os-weka1/index.html?ca=drs-. The final tool we will be using is Carto, which is a mapping tool that works similarly to Tableau. It is a part of the github student developer pack, and will be very good for exploratory analysis, specifically looking at different attribute clusters based on geographic location. Examples of Carto visualizations can be seen in the "Results So Far" section of this paper.

MILESTONES

We are currently in the 13th week of the semester, and we will present the project in the sixteenth week, so that gives us four weeks to complete the project. Cleaning, preprocessing, and integration are complete. We expected preprocessing to take more time than

cleaning and integration, so it received 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. We are left with three weeks 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 21st along with, because the evaluation and visualization produce common data, which can be applied to both modes of communication. The project should be polished and ready to present by Monday April 29th.

RESULTS SO FAR

Our results are based on preliminary data exploration. We have several visualizations, and summary data that are relevant to the scope of the project.

The first visualization was derived from the integrated dataset yelp_money_merge.csv. It shows the mean income distribution for the Phoenix area.

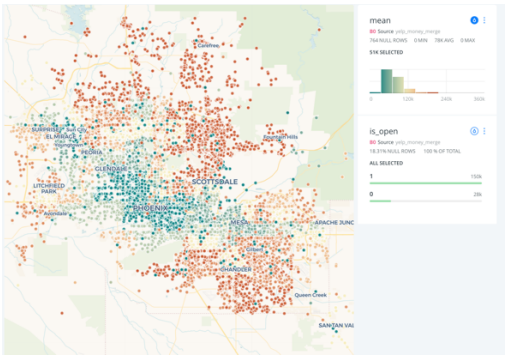


Figure 2. Phoenix area mean income distribution.

This map shows distinct clustering of lower to middle incomes around the Phoenix city center, and then increased income outside the city center. This map is interesting in the context of the second visualization, which shows open and closed businesses for the same geographical region.

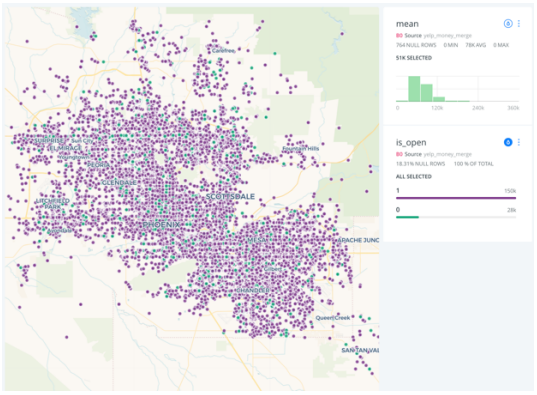


Figure 3 Business open or closed map

Although the clustering is not as dramatic as in the previous visualization there are apparent 'clusters' around the Phoenix city center, which is where lower incomes are clustered as well. There is also a distinct line of business closures around the 101 highway which forms the norther boundary around the Phoenix Metro area.

The third geo visualization is a chloropleth map of the distribution and clustering of the number businesses represented in the in North America

Figure 4 Clustering of represented businesses

This visualization shows the nine metro areas where the Yelp data was gathered from. There is a good mix of geographic regions, which will help to normalize income distribution. Each cluster except for Los Angeles contains at least one area near the maximum sample size for an area at 33k samples.

The following two visualizations are query counts, and are not based in data mining techniques, but are summary visualizations that motivate the attributes that need further investigation.

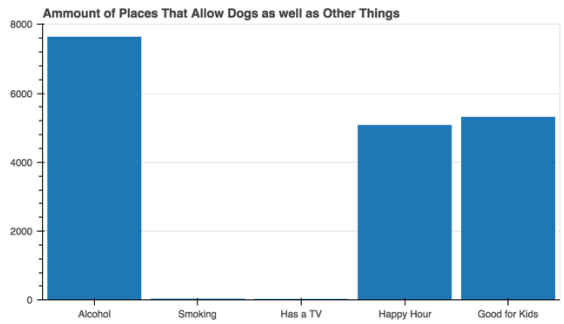


Figure 5 Attributes present in data

This visualization shows attributes that are present in the data to a significant degree. Knowing this information will allow us to use decision tree induction for classification. For example, if the known label is open or closed (maps to business success or failure), then possible candidates for the first tree split can be Alcohol, Happy Hour, Good for Kids. Similarly, we can use frequency of check in data to look for businesses outside of the trend. The visualization below shows check in frequency for all businesses across all seven days.

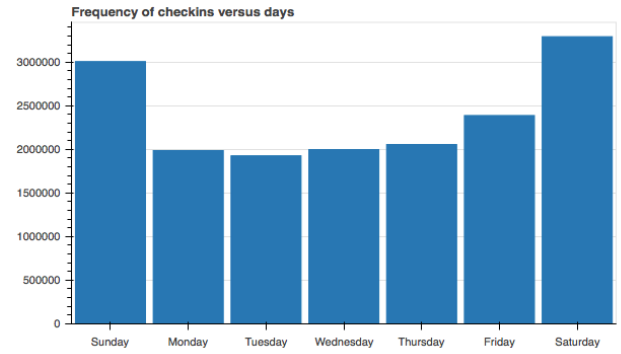


Figure 6 Checkin Frequency

This information becomes significant when we drill down to specific businesses, regions, or income grouping and see flat trend, where check in is high for all seven days or check in is low for all seven days. Another interesting insight check in data can give us outside of frequency is distribution. The data contains 3911217 check ins total, and 146350 unique check ins, so if the check ins are evenly distributed across all business, then there should be approximately 26 check ins per business, but that is probably not the case.

Now that our data has been integrated, and clustered or binned we can begin to drill down, and analyze the classes of the data by employing KNN classification, and decision tree induction classification which will produce further insights as to what attributes contribute to business success, or failure.