Yelp DataSet – Calculating Popularity Zones
CSE 190 Assignment 2

Tyler Daniel
University of California, San Diego
tdaniel@ucsd.edu

INTRODUCTION

This assignment will cover the Yelp Dataset Challenge which provides the data that lists the popularity of a business. Popularity will be determined by the involvement or activity from consumers. These "Popularity Zones" will show how large the activity is made from the consumers. The Yelp Dataset Challenge, along with the information from the lectures of the CSE 190 Data Mining course taught by Professor Julian McAuley, will become operated on to produce the expected output. When considering the popularity zones, there will be several categories to be examined. First of all, the sheer number of reviews in the area will provide a better example of the popularity. Then the amount of check-ins at the businesses will determine the reputation. Finally, the reviews and rating for the restaurants, bars, or any other business will determine the consumer trends for all business in each area.

The specifications required the dataset to be quite large; in fact, it must have more than 50,000 samples. I through several possible datasets, such as social media sites like Facebook and Reddit. Facebook would describe how popular a post, page or personality in the amount of "likes" and comments. This option would be similar to the Yelp popularity zones, but being only restricted to certain pages created by any person skews perception. Another huge deterrent was the fact that Facebook does not have a "dislike" option to show negativity for a post, page or personality. This misrepresentation would not be that effective of an analysis so I looked for a different option.

The concept of having "likes" and "dislikes" led me to consider examining the dataset for Reddit. The large community and various posts about nearly anything could provide an interesting dataset to go through. However, further research into the Reddit history behind the posts shows that there is well over a billion aspects to consider which involves positive and negative karma, number of comments, users, and topic categories. This vast amount of information would provide an interesting challenge but a little too much for the consideration of this assignment.

Then I came across the Yelp Dataset Challenge and the contest that was set up. In this contest, Yelp is looking for several topics for the programmers to delve into, such as cultural trends, location mining, urban planning, seasonal trends, infer categories, and change points. The location mining appeals to this project since it means to determine if a business truly thrives because of the area or something more. The popular zones that includes several businesses like restaurants, bars, and stores is an interesting topic in itself. The amount of reviews and quality of ratings more certain businesses stand out for consumers and others to be shunned. Therefore, the dataset provided from the challenge can help determine what makes a popular zone.

The data provided and all the lessons learned from class and in other assignments will help produce the output needed. Looking though the previous works, the topics of regression, classification, clustering and prediction will help create a more sophisticated model of the data. The goal of the assignment is to find the most popular areas with the hopes that these popular zones can lead to more business.

ANAYLSIS OF DATASET

The dataset provided from the Yelp Dataset Challenge described several useful topics such as check-in, user, review, business, and tip. Going

through each category will show which zones will be the most popular. The check-in data will give information on the number of times people have checked in at a business and the time of day. The user category will define the profile of the user/reviewer which will include the age of the account and connections. The review data provides several points as it is the most descriptive. The review covers the rating a person gave for a business, the business name itself, the popularity or usefulness of the review, the day the person gave the review and obviously the text of the review itself. The business category will provide the location of the business, the name, the number of reviews, the average ratings, and hours of operation. Finally, the tip data will give small information from the reviewer that tells the reader what to do at the business.

The interesting finding from the dataset is the cities provided in the data. The cities in no particular order are: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas and Madison in the U.S., Edinburgh in the U.K, Karlsruhe in Germany, and Montreal and Waterloo in Canada. This will provide a vast spread of data as trends from each city and country will differ. Using the dataset for all this cities will mean that our averages when calculated later on might not be as representative of some areas of business.

Using the information provided in the dataset, I should be able to conclude the stronger areas of popularity from all the categories listed above. Several combinations of ratings and reviews will be tested upon to show the popular zones needed to be shown. To calibrate the results, the dataset will be separated into two categories as shown through lecture: training and test set. The most accurate depiction of the data using the test set will determine the area of popularity.

PREDICTIVE TASK

Based on the dataset provide by Yelp, this project's predictive task is measuring the popularity of a location (location mining as referred to in the challenge) by examining business ratings in the locations, the relative costs for businesses, and the information or text in the reviews. By simple inference, the predictions that can be made are: the larger number of reviews and higher ratings provide a more popular area and the lower the number of reviews and low ratings signifies a more desolate area.

The baseline of the project, which will be the first method of solving this problem, will use a linear regressor on the training set of the ratings from the dataset. The idea is to separate the cities with a more fundamental difference in cultural trends. Easily separated will be the individual cities in Europe, which are Edinburgh, UK and Karlsruhe, Germany. When looking over the data in the U.S., the city of Phoenix, AZ would be the most culturally different when compared to the other U.S. cities. Multiple issues can arise with this distinction because the vast difference of cultures in differing countries will differ on several categories of interest. This will also include the fact that these cities are so large in population that the minority of an interest could still be a large body of people. However, this project has the goal of finding the opinion or interest in the majority of the population, rendering these issues insignificant. The number of check-ins for a business will be translated into a predictor using the ratings. Then, the test set will be used to determine the number of check-ins for a business. The test set will be the cities that were not listed in the above.

The mean-squared error (MSE) will be used to check the accuracy of the predictor. The mean-squared error of an estimator "measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated (2)." Thus, the smaller the MSE, the closer the fit is to the data. The higher the MSE means that the predictor needs to be changed and perhaps add another category to create a more accurate model. If the possibility of the predictor based on ratings produces a high MSE, a new predictor will be made based on the total number of reviews for the businesses. If that predictor fails as well, then a new predictor will be constructed using both ratings and reviews to hopefully make a more accurate model. If the first two predictors fail, the last one should work and provide the most accurate representation of the popularity zones. These popularity zones will predict the most consumer dense locations for businesses.

LITERATURE

The Yelp Dataset Challenge provides a list of topics for programmers to examine in the data given. The broad use of the data provides an opportunity to find several trends such as cultural preference, location popularity and the effects of time. With the specific cities chosen in this contest, we can see that the goal for cultural trends and location mining reflect the project well. The people in different countries, and even different cities inside the country, will have different likes and dislikes. The countries with multiple cities will certainly have different preferences which need to be taken account for. Obviously, the differing countries will have different preferences because of the distance between them will change their likes.

Working with the Yelp Dataset after several rounds of the challenge provides several reports for reference. This is useful because the challenge is still fairly new meaning that there is not as much literature written for it compared to other projects. The data for each round varied slightly, but was mostly similar to the dataset that is being tested on. Most notably, Professor Julian McAuley was a part of a pair that was one of the winners for round one of the challenge (1). His paper, which was called "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text", predicted "how a user will respond to a product." To do this, he obtained highly interpretable textual labels for latent rating dimensions and more accurately predicted product ratings by harnessing the information present in review text. By combining the accuracy of rating prediction with mean squared error and the likelihood of the review corpus, we can see how certain users rate businesses.

MODEL

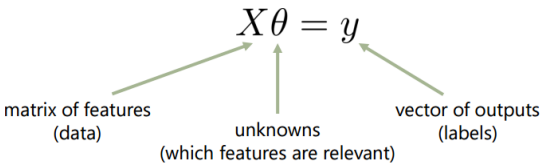
In order to fully optimize the model, the dataset must be fully understood. Check-ins for the businesses were broken up in accordance to hour of the day. The data will show multiple numbers for a business which need to be combined to give the accurate representation of check-ins. The large list of check-in numbers did not provide a clear depiction of total check-ins, so some inference on the association of check-ins for businesses needed to be made. To solve this issue, lists and dictionaries were used to organize the data in an efficient way. This allowed the number of check-ins to be tabulated easily.

Linear regression is "an approach for modeling the relationship between a scalar dependent variable *y* and one or more explanatory variables (or independent variables) denoted *X* (4)." This will follow the function provided below:

$$f(\text{data}) \xrightarrow{?} \text{labels} \tag{2}$$

Simple enough, we put the data into a function and the function will output information on the labels.

After this, the linear regressor will follow the predictor of:



(2)

Following the labels from above, X is the matrix of features (data), Θ is the unknowns (which features are relevant), and y is the vector of outputs (labels) (4). When placed on a graph, the predictor is created by the representation of a line and the correlation between the data.

As stated in the predictive task portion, we need to use the mean squared error to create an accurate representation of the data.

Mean-squared error (MSE)

$$\frac{1}{N}\|y - X\theta\|_2^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i \cdot \theta)^2 \tag{2}$$

In words, the difference of the predictions and the observed values corresponding to the inputs to the function which generated predictions are squared then multiplied by the mean (3). Thus, the terminology of mean squared error is created.

Assessing the validity of the mean-squared error requires the mean and variance for the

predictors. This will ensure an accurate representation on the average of all businesses.

RESULTS

Test 1

City, Country	*Mean Check-in*	*Variance Check-in*
Las Vegas, United States	251	936015.03
Waterloo, Canada	26	494.50

Base of Predictor	*Predictor Value*
Values of Ratings	1006707109643.5606
Number of Reviews	1.307584287e+12
Ratings and Reviews	2.321568837e+12

City, Country	*Mean-Squared Error*
Las Vegas, United States	924684.33
Waterloo, Canada	494.69

Test 2

City, Country	*Mean Check-in*	*Variance Check-in*
Las Vegas, United States	251	936015.03
Waterloo, Canada	26	494.50

Base of Predictor	*Predictor Value*
Values of Ratings	-4.93266149e+12
Number of Reviews	4.72945378e+12
Ratings and Reviews	-218949259887

City, Country	*Mean-Squared Error*
Las Vegas, United States	389769.52
Waterloo, Canada	282.46

CONCLUSION

The results of the project do not output what was to be expected. The large variance and high predictor values for the two example cities, Las Vegas and Waterloo are the reason for the undesired output. As speculated before, there are several reasons as to why the variance and predictors were so large. Some businesses do much better than several around them. The large number of check-ins for casinos and the McCarran International Airport create a significant discrepancy for model. A method to fix this issue would be to remove these businesses of high success, but doing so would not predict such an accurate or truthful model. The inconsistency of the check-ins along with the close grouping of the ratings also made the outcome irregular. Looking back at the experiment, a possible change to produce a better result would be to categorize the locations even further. Certain areas fair differently than others. The city of Waterloo is considerably small than that of Las Vegas. Las Vegas has more business in areas surrounding the strip but suffer as you travel further away from the center of the city.

An example following this logic is shown below. This is a heatmap of the number of reviews in Phoenix, Arizona. By focusing on these areas of more activity, the predictor could have done a more accurate representation of popularity.

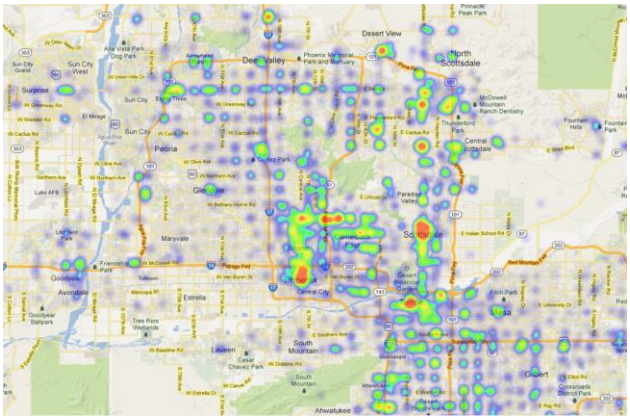


Figure 1

REFERENCES

(1) McAuley, Julian, and Jure Leskovec. "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text." (n.d.): n. pag. 2013. Web. 01 Dec. 2015.

(2) McAuley, Julian. *CSE 190 – Lecture 1.5* (n.d.): n. pag. Web. 01 Dec. 2015.

(3) "Mean Squared Error." *Wikipedia*. Wikimedia Foundation, n.d. Web. 01 Dec. 2015.

(4) "Linear Regression." *Wikipedia*. Wikimedia Foundation, n.d. Web. 01 Dec. 2015.

Most of the code

```
def parseData(fname):
  for l in urllib.urlopen(fname):
    yield eval(l)

print "Reading data..."
review = list(parseData("yelp_academic_dataset_review.json"))
checkin = list(parseData("yelp_academic_dataset_checkin.json"))
user = list(parseData("yelp_academic_dataset_user.json"))
business = list(parseData("yelp_academic_dataset_business.json"))
print "done"

businessList = []
for i in business:
        businessList.append([i['business_id'], i['city'], i['areaPopular']])

checkin_Bus = []
for i in checkin:
        checkin_Bus.append(i['business_id'])

checkin_List = []
for i in checkin:
        checkin_List.append(i['checkin_info'])

checkinTotal = []
for a in range(0, len(checkin_List)):
        checkinTotal.append(sum(checkin_List[a].values()))

businessPopular = []
areaPopular = []

totalCount = dict(zip(checkin_Bus, checkinTotal))
checkin_Sorted = sorted(totalCount.items(), key = operator.itemgetter(1))
checkin_Sorted.reverse()
for x in checkin_Sorted[:5000]:
        businessPopular.append(x[0])

for i in businessPopular:
        if i == businessList[i][0]:
                areaPopular.append(businessList[i])
```