**Predicting User Hot Spots Using the Yelp Dataset**

**CSE 190 Assignment 2**

**Pedro Villaroman**

UC San Diego

San Diego, CA

pvillaro@ucsd.edu

**Phong Tran**

UC San Diego

San Diego, CA

pdt004@ucsd.edu

**Abel Jara**

UC San Diego

San Diego, CA

abjara@ucsd.edu

**ABSTRACT**

*In this project, the patterns related to the popularity of a certain location with consumer businesses based off of customer activity are observed and evaluated. We term these locations of popularity as “hot spots” where we see a large amount of customer activity. In order to do this, we utilize the Yelp Dataset Challenge dataset for Challenge Five, which provides the necessary information to accomplish our goal. We manipulate and mine the dataset using the various tools we have learned from our CSE 190 Data Mining and Analytics course taught by Professor Julian McAuley. In the end, we hope to achieve useful results that will help us understand consumer trends based off of popularity of nearby restaurants quantified by user activity such as reviews on the restaurants, ratings on the restaurants and check ins at the restaurant.*

**INTRODUCTION**

Upon receiving the assignment, our group, consisting of Abel Jara, Phong Tran and Pedro Villaroman have looked through several different data sets ranging from simple reddit posts and reddit user comments to even Pornhub datasets consisting of the most popular types of pornography in certain regions of the United States. We looked at these datasets in order to find interesting data that we may be able to manipulate for useful and beneficial results that could potentially help others understand the business of the data set and user trends. However, many problems with the data sets arose while observing the datasets we have come across. For example, Reddit limited what kind of posts we could see by only allowing us to see 25 posts per page (a problem that we could have overcome with code, but decided not to) and only give us limited reviews for a certain post. Overall, the dataset could not give us the necessary information we need in an efficient manner that would help us understand user trends. In regards to the released Pornhub dataset we found online, the dataset only contained 53 data points documenting the average time a user spends watching a video of a certain category.

After much thought and consideration with the observed datasets, our group has decided to use the Yelp Dataset Challenge dataset given for Challenge Five because of ease of access and relevant information that we could use to explain user trends. As keen lovers of food and avid explorers of the “next best thing”, we wanted to think about what makes a restaurant or barber shop so popular, and how it becomes even more popular based off simple user trends. The information that is needed to solve this problem, like how often a user visits a restaurant and whether or not we can find out when he or she visited the restaurant, can easily be found from what Yelp explains it has provided.

With this data, we aim to utilize the tools we have learned from Professor Julian McAuley and his Data Mining class. We believe that tools like logistic regression from Homework 1 and latent-factor models from Homework 3 will be the most beneficial to our project. Later on, we will use graphing and modeling tools from pbpython.com that will help us visualize the data in a more efficient and easily understandable manner. In the end, we hope to achieve our goal in understanding the customer trends and hopefully be of use to future business owners looking to find the best place to start a business.

**EXPLORATORY ANALYSIS**

We chose to explore the Yelp dataset from the fifth Yelp Dataset Challenge because of the amount of the vast amount of data points it provides along with the following fields of data: business, review, user, check-in, and tip. Business holds the information related to a certain business associated with Yelp such as the name of the business, the location, the rating it has, the amount of reviews it has, and even the operational hours. The review field holds information about the review like which restaurant the review is for, the rating a user gave a business, the number of votes a review received, the date of the review and the text of the review. The user field holds information for a certain user like friends, and how long the user has used Yelp. The check-in field allows us to see when and how many times a restaurant was checked in at. The tip field gives us the information for a business tip like if a user should buy a certain item at a business or not.

Based off what was given in the dataset, we concluded that our project could be accomplished by using certain pieces of data given and also by manipulating certain pieces to find useful data that we need. We also concluded that we could use several different types of features given to reach what we want and even use a combination of these features together. Depending on what gives us the best and most accurate results compared to the test set (we plan on dividing the dataset into a training set and a test set) we will be able to find out what is the best feature representation for our problem.

A limiting factor that we noticed from the Yelp dataset was that we are only given specific cities that range from many different backgrounds. We were given data from Edinburgh in the U.K., Karlsruhe in Germany, Montreal and Waterloo in Canada, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, and Madison in the United States. This could pose as a potential problem because we could be getting different types of trends for different cities. However, we will take this into account later on. But based off these findings, many other types challenges could arise.

**LITERATURE**

For the dataset that we are using, Yelp released a series of challenges and problems that they suggest we try solving or predicting. However, we did not get the chance to see this challenge before we started with our own predictive task (which we will take more in depth about later). Since the dataset consisted of data from different cities in different countries, it is geared towards observing various cultural and urban trends within different populations. For example, would a group of people Karlsruhe, Germany prefer a certain type of food or business to another type of food whereas a group of people in Phoenix thinks otherwise? Yelp released a lot of meaningful and challenging questions that could definitely be solved with the provided dataset.

Since the Yelp dataset was recently posted, there are no known pieces of literature that are geared specifically towards the dataset we are using. However, there are reports of other projects for previous Yelp Dataset Challenges that contain data similar to the data in the dataset we are using. In the first Yelp Dataset Challenge, the data featured reviews and businesses from the greater Phoenix metropolitan area, which is different than the data that we have in the current dataset. For example, Professor McAuley had a winning submission to the first dataset challenge, predicting how a user will respond to a product using latent factor models and building predictors on user trends by first exploring the “hidden dimensions” of customer reviews using a wide set of different datasets [1].

In order to understand how users review certain things, the paper aims to first understand rating dimensions by “using an objective that combines the accuracy of rating prediction (in terms of the mean squared error) with the likelihood of the review corpus (using a topic model)” [1]. The models also serve as tools to find more informative ratings that could potentially predict the category a review is in or the genre a product is in.

**PREDICTIVE TASK**

We decided on what our exact problem was going to be after looking at the Yelp dataset. From the Yelp dataset, we want to predict the amount of check-ins and activity in an area populated with businesses by measuring the content of the reviews, the ratings for the businesses in the area the categories of the businesses, and the average prices of the businesses in the area. We predict, before we start fitting predictors on the dating, that the higher the ratings and number of reviews there are for a set of restaurants in a city, the higher the amount of check-ins there will be.

There are, of course, several different approaches we could take to solving the problem. The first approach will serve as our baseline. The first method we will employ is training a linear regressor on a training set of data’s ratings, where we will isolate the cities we believed would have the most unique data: Edinburgh in the U.K., Karlsruhe in Germany, Montreal in Canada, and Phoenix in the United States. However, since these states have culturally different backgrounds, it was brought to attention that we could run into the problem that one set of people in a country rates differently than another set of people on the same items in a different country. That will not matter though since we are trying to predict the majority of the Yelp user base trends on the ratings. Since we are given the amount of check-ins a certain restaurant has, we can then build a predictor based off the ratings that will help us predict how many check-ins a restaurant will receive in the test set, which are the other cities we haven’t used in the training set.

After building the predictor, we plan on using the mean-squared error to assess the validity of the model. If we have a high mean-squared error, then we need to either change the way we approach the problem, or add another feature that could predict the amount of check-ins better.

If the predictor with just ratings yield poor results, we will build a predictor based off the number of reviews a business has. Once again, we will use the mean-squared error to evaluate our results and see the accuracy of the predictor.

If the predictor that we have built does not accurately predict the check-ins (which we expect from just having a predictor on ratings) we will build a feature set with both the rating and the number of reviews given to a business. Both of these features together should allow us to fit the model much more precisely, thus yielding accurate predictions. Once again, we are going to use the mean-squared error to assess the validity of our approach.

By building this predictor, we should be able to accurately see where the user “hot-spots” occur and maybe even show future business owners where the best places to start a business would be.

**ALGORITHMS**

The first thing we had to do in order to solve our problem was to understand the structure of the dataset. We observed that the check-ins data field was built based off the number of check-ins at a certain hour on a certain day. For example, at 1pm to 2pm on a Saturday, there were four check-ins and at 5pm to 6pm on Monday, there were eight check-ins instead of giving us the total amount of check-ins for the restaurant. We also noticed that the information that was needed was not directly given and we had to associate which check-ins is the correct check-ins for a certain business. For example, the amount of check-ins we would have would be associated to a random, nonsensical sequence of letters and numbers that was associated to the actual business.

We did overcome this problem through the utilization of dictionaries and lists, so we were enabled to move onto building predictors for our features.

Linear regression is a type of supervised learning where we will attempt to infer the “underlying function that produced the labels associated with the data”, or our output, from the labeled data, our input [2]. Basically, we will have a function similar to the format below:

f(data)  labels

Where we take the data, put it into a function, and hopefully get useful information regarding the labels that are used.

The equation for the linear regressor assumes the form:

XΘ = y

Where X is the matrix of features (data), Θ represents the unknowns (or the features that are relevant), and where y is the vector of outputs (the labels). This equation represents a line through a graph of points that will show the best correlation between the data, and act as our predictor.

In order to verify that our predictor is valid, we want to use the Mean-Squared error as described below:

 [3]

Which takes the sum of squares of the difference between the prediction value we attained from multiplying the predictor by the test set value and the actual value and then dividing it by the number of values in the test set.

In order to verify that our MSE is valid, we have to calculate the mean and the variance for each of our predictor features, so we could attain an accurate representation. Basically, the MSE is proportional to the variance of the data we are testing, so we should be able to judge the accuracy of our predictor based off the values of the MSE and variance.

**RESULTS**

Experiment 1

Mean and Variance for Check-ins in Waterloo

Mean for check-ins in test set = 20

Variance for check-ins in test set = 486.5

Mean and Variance for Check-ins in Vegas

Mean for check-ins in test set = 236

Variance for check-ins in test set = 922585.03

|  |  |
| --- | --- |
| **Feature** | **Predictor** |
| **(1) Ratings** | 1005908162257.2361 |
| **(2) # Of Reviews** | 1.30777524e+12 |
| **(3) Both** | 2.3136834e+12 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Prediction** | **Actual** | **MSE** |
| **(1)** |  |  |  |
| **(2)** |  |  |  |
| **(3)** |  |  |  |

Experiment 2

Mean and Variance for Check-ins in Waterloo

Mean for check-ins in test set = 20

Variance for check-ins in test set = 486.5

Mean and Variance for Check-ins in Vegas

Mean for check-ins in test set = 236

Variance for check-ins in test set = 922585.03

|  |  |
| --- | --- |
| **Feature** | **Predictor** |
| **(1) Ratings** | -4.95766150e+12 |
| **(2) # Of Reviews** | 4.73864834e+12 |
| **(3) Both** | -219013160000 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Prediction** | **Actual** | **MSE** |
| **(1)** |  |  |  |
| **(2)** |  |  |  |
| **(3)** |  |  |  |

**CONCLUSION**

Based off our results, we did not end up with what we wanted. It started when we calculated the variance for the two cities we were going to test on and received very high values for each city, Waterloo and Las Vegas. With this high variance in mind, we decided to calculate the predictors and again, got absurdly high numbers, as seen in the tables. There are many reasons why the predictors had such high values for each of our features. Firstly, when predicting the check-ins using ratings, we saw that the check-ins varied greatly while the ratings were very close together. This was giving us very skewed results. We then noticed that some businesses in each city had a much higher number of check-ins than other businesses in the area. For example, the Las Vegas airport had over 60,000 check-ins, giving a stark contrast to the mean of the set, where we got 236. To fix this problem of the high variance, we attempted to eliminate the places with a high number of check-ins; however, we concluded that this was contradictory to what we wanted to predict. We still need a way to predict the number of check-ins for a business.

**Things We Wanted to Do**

After having the opportunity to play around with the Yelp dataset, we realized that there was so much more to the problem that we could have accomplished. We found out that each business data field has a ‘neighborhood’ attribute associated with it that could give us much more meaningful results regarding our proposed problem. With the neighborhoods, we could magnify the specific spots in a city that people frequent, which we could then share with business owners who want to find the exact places to potentially start a business. However, we found out that not all businesses were associated with a neighborhood and some even had more than one neighborhood. This would have skewed our predictors, leaving out useful data that may or may not be helpful.

We do have a solution to combat this problem though. Each business is also associated with latitude and longitude coordinates, which would have allowed us to use the cluster algorithm and force a business into a neighborhood by distance. With that, we would have much more useful information regarding potential locations beneficial or harmful to business growth.

**REFERENCES**

[1] J.McAuley,J.Leskovec. Hidden Factors and Hidden Topics:Understanding Rating Dimensions with Review Text. In ACM RecSys. 2013

[2] J.McAuley. Supervised Learning – Regression.CSE190 Lecture 1.5. April 7.2015

[3] Mean Squared Error. Wikipedia. Wikimedia Foundation, n.d.Web.June 2.2015