ASSIGNMENT 2 – Evaluating The Yelp Elite Squad

Gian Costa
Data Mining and Predictive Analytics
gcosta1992@gmail.com

Arturo Aguilar
Data Mining and Predictive Analytics
ara003@ucsd.edu

Eric Jiang
Data Mining and Predictive Analytics
erjiang@ucsd.edu

University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093

ABSTRACT

In this paper, we attempt to create a model that accurately predicts if a user, given their features, has ever been Elite. We do this by considering the features of users, such as their review count, number of votes their reviews have received, number of friends, etc. By using supervised learning, and more specifically, Linear Regression, we plan to create a model that considers weights of different features and see if those weights, applied to a given user's corresponding features can be added to together and compared to a tuned threshold to produce a binary classifier/predictor.

1. INTRODUCTION

Yelp is well known as the most popular tool used by people to discover new restaurants through reviews. Along with the general helpfulness of the reviews themselves, the application has integrated a social aspect that has drawn more people in and encouraged them to contribute and establish a presence on Yelp. This social aspect can be seen through the inclusion of a "Friends" feature, a "Fans" feature, votes and likes on reviews and much more. A unique social feature that we were most interested in exploring however, was the Elite user feature. Yelp has created a hand-picked exclusive community of users who produce exceptional reviews and are most active. According to their website, "Members of this exclusive, in-the-know crew reveal hot spots for fellow locals and are the true heart of the Yelp community, both on and offline" (Yelp). By creating a more competitive environment for users who want to become part of the club and creating this sense of importance in users who are Elite, Yelp has increased the quality and quantity of its reviews.

We wanted to try and discover what features of different users make them "Elite". According to the Yelp site, users qualify based on authenticity (that the user is a real person), contribution (creating reviews that are actually useful to other people) and connection (how active users are, both with their own reviews and others). However, Yelp specifies that each addition to the Elite Squad is picked very subjectively and there is no guarantee of admission. An accurate predictive model that predicts if users have ever been Elite could provide valuable insight to users looking to give themselves the best chance to becoming Elite because it would expose the user features that are most significant in helping them achieve this.

2. DATASET DESCRIPTION

Because our team consists of three amateur foodies, we naturally took interest in the superfluous, scrumptious data from Yelp. Conveniently, an academic dataset consisting of "1.6M reviews and 500k tips by 366k users for 61k businesses", provided in JSON format, is easily accessible online for the Yelp Dataset Challenge. Although this sample data is (probably) considerably smaller than all that lies within the corporate database, it was still too much data for us to work with given the scope and time allowed for our assignment. After exploring each category of data and deciding on our task and goal of interest, we ended up using only using the user data (for our predictive task), which is formatted as follows:

```
{
   'type': 'user',
   'user_id': (encrypted user id),
   'name': (first name),
   'review_count': (review count),
   'average_stars': (floating point average, like 4.31),
   'votes': {(vote type): (count)},
   'friends': [(friend user_ids)],
   'elite': [(years_elite)],
   'yelping_since': (date, formatted like '2012-03'),
   'compliments': {
      (compliment_type):(num_compliments_of_this_type),
...
},
'fans': (num_fans), }
```

The user json file provided to us by Yelp contained 366,715 entries. We divided these entries into a training set, a validation set and a test set:

Training Set: 240,000 entries

Validation Set: 60,000 entries

Test Set: 66,715 entries

Of the 366,715 user data entries, here are their following statistics:

- Average 'review_count': 32.214809866
- Average 'average_stars': 3.7187821878
- Average number of 'votes': 122.21769494
- Average number of 'friends': 7.02501670234
- Average number of years 'elite': 0.224340427853
- Number of years 'elite': {0: 341414, 1: 3612, 2: 6922, 3: 5166, 4: 4152, 5: 2419, 6: 1584, 7: 802, 8: 387, 9: 186, 10: 57, 11: 14}
- Number of users 'yelped_since' (increasing order): [(51, '2004'), (691, '2005'), (1148, '2015'), (3974, '2006'), (10676, '2007'), (19390, '2008'), (32968, '2009'), (50505, '2014'), (50722, '2010'), (63483, '2013'), (63897, '2012'), (69210, '2011')]

- Average number of 'compliments': 1.44551763631
- Average number of 'fans': 1.57533779638

Among all of the different features, 'elite' caught our attention as the average number of years of elite status appeared to be considerably low – about 2.7 months or 82 days – along with the clear trend of significantly less users holding elite status for more years. From the Yelp website, becoming a part of the Yelp Elite Squad is a real treat, where "members of this exclusive, in-the-know crew reveal hot spots for fellow locals and are the true heart of the Yelp community, both on and offline." But wait, there's more; becoming a selected Elite entails…

> Only a shimmering smorgasbord of stuff that'll change your life: Nifty new friends, über-local gatherings, invites to fun (and free!) parties at least once a month, and a shiny profile badge. Most importantly, you'll join the ranks of some of the most influential tastemakers on the site and in your city. Desperately seeking schwag? You'll have first dibs on everything from Yelp sunglasses and lip balm to sweatbands and temporary tattoos. Represent!!

And while becoming Elite is free of charge, what's the catch? Hence, our goal is to divulge (and exploit) the truth behind the Yelp Elite Squad system of selection. Yelp claims that their process of figuring out who deserves the quirky title every year (terms last until/nominations happen at the end of every calendar year) considers not only the "frequency and quality of reviews", but also being "model Yelpers that engage on the site by sending compliments, voting Useful, Funny, and Cool (UFC) on reviews, participating respectfully on Talk, and consistently posting quality content." To see for ourselves whether elite status users indeed reflect better numbers, we've accumulated the following average statistics:

Table 1. This table displays the average statistics of Elite users compared to Non-Elite users

Average Statistics	Elite Users	Non-Elite Users
# Reviews	245	16
# Average Stars	3.78013675349	3.71423541507
# Votes	1336	32
# Friends	55	3
# Compliments	8	0
# Fans	16	0

Evidently, Elite status users are superior in the number of reviews that they've written; the average number of votes that their reviews have received; their number of friends; the number of compliments they've received; and their number of fans. Our work aims to determine which of these features best allow us to predict whether a user deserves to be a Yelp Elite.

3. PREDICTIVE TASK

3.1 Model Selection

Our goal was to predict, given a set of user features, whether or not that user had ever been Elite. Essentially we needed a reliable supervised learning technique to model the relationship between these input and output variables, so that we could predict output based on input. We also needed a binary classifier that would predict 0 to indicate that the user had never been on the elite team and 1 to indicate that the use had been on the elite team for at least one year. We attempted using the following models:

Model 1: A Linear Regression model whose result could be compared to a threshold to make a binary classification

Model 2: A Logistic Regression model for that would result in a 1 for true having been elite, or 0 for having never been elite.

We decided to use the first option because we found consistently higher prediction accuracy in the test data. We describe why we chose model 1 further in the results.

3.2 Model 1 Description

In this model we used linear regression with various user features to fit theta values and determine which features were significant for this predictive task. Our linear regression model:

$$X\theta = y$$

Here our vector of outputs (y) contains the number of years each user has been elite. The feature matrix (X) contains any combination of 'reviewCount', 'votes', 'friends'(count of friends), 'votes'(count of votes). Based on the description of Elite users qualifications on Yelp's site, we decided that one or more of these features would most likely be the important features used by Yelp to decide who is eligible to be Elite. From Table 1 we can see that on average, Elite users have substantially more friends, votes, reviews, and fans. This was also a solid indication that these features would have some (or most) importance in determining if a user had ever been Elite. We tried a combination of different features in our feature matrix to see which ones yielded the thetas that produced the highest prediction accuracy. We will explore those different combinations in the Results Section.

To use our Linear regression model to achieve this binary classification, we needed a threshold that the linear regression model result could be compared to predict a binary value. A prediction of 0 would mean the user had never been Elite. A prediction of 1 would mean the user had been Elite at some point. Essentially our model would look like the following:

$$\textbf{prediction} = \begin{cases} \mathbf{1} \;\; \boldsymbol{if}\ (\theta_0 + \theta_1(\boldsymbol{feature1}) + \theta_2(\boldsymbol{feature2}) + \boldsymbol{etc.}) > \boldsymbol{threshold} \\ \mathbf{0} \;\; \boldsymbol{otherwise} \end{cases}$$

where 'feature 1', 'feature 2', etc. would be the user in questions' features, the theta values would be the weights, or importance values fitted by the linear regression model that we trained. We decided that the best way to chose the optimal threshold value would be to tune it based on the validation data set. Then we would use this final optimal threshold value when making predictions on our test data.

In making our predictions we competed with a baseline prediction accuracy of 0.94314, which can achieved by just predicting 0 (or "user has never been Elite") for all users. This questionably high baseline is due to the fact that in the test data there were only 3793 Elite users out of 66,715 total users.

3.3 Model 2 Description

In this model we used logistic regression with the average length of all reviews by a user to fit theta values. Our logistic regression model:

Our input feature vector was the average length of a user's reviews first sentence. Fitting it to the testing dataset, we used scipy built-in function fmin_bfgs to yield the best parameters theta by which to form the decision boundary that our test set data would be compared against. This returned a theta of [-2.69053, 0.0015697] that when used on the test set, would determine if the given data resulted in a probability above or below .5 (above being 1 and 'elite'; below being 0 and so not 'elite').

This resulted in a 0.94314 % accuracy, which while at first seems favorable is really no better than having given all users a value of zero for the same reasons as noted at the end of 'model 1 description'

We conclude that our initial intuition that the length of the first sentence of the user's reviews is negligible indeed does not give us much insight on what determines a user having at some point been 'elite'.

4. RELATED LITERATURE/STUDIES

4.1 Dataset Information

For our assignment, we've decided to use the dataset provided for the Yelp Dataset Challenge (www.yelp.com/dataset_challenge). Although Yelp hasn't specified exactly how the thousands and millions of entries from different data categories have been chosen, it is stated that the dataset "includes information about local businesses, reviews and users in 10 cities across 4 countries" (U.K.: Edinburgh; Germany: Karisruhe; Canada: Montreal and Waterloo; U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison). Also, we aren't told how Yelp has used the data provided in this academic dataset for their business. However, many of the challenge participants have certainly used the dataset for a wide variety of unique data mining tasks (the seven grand prize winning submissions from the first three rounds can be found from the link provided earlier).

4.2 Similar Datasets Studied In the Past

Since we have only really worked with the user data from the dataset (for our predictive task), we will list how the winning submissions from the challenge have studied all of the various categories of data (business, check-in, review, tip, user).

Inferring Future Business Attention by Bryan Hood, Victor Hwang and Jennifer King

- **Data:** business, review, user
- **Task:** 1) "predict the number of reviews that a particular business should have at the current time" to compare current business attention to what the Yelp market expects, and 2) predict "how many reviews [a particular business] should have in the next six months" to describe the trend in business attention – how its attention will change in the future
- **Methods:** K-means clustering (on users); sentiment analysis (of reviews); PCA, Univariate Feature Analysis, Greedy Feature Removal, K-Fold cross-validation (for feature selection); Support Vector Regression (for prediction)

Improving Restaurants by Extracting Subtopics from Yelp Reviews by James Huang, Stephanie Rogers and Eunkwang Joo

- **Data:** business, check-in, review, user
- **Task:** "describe latent subtopics discovered from Yelp restaurant reviews" to point out customer demands and discover what customers care about in rating restaurants
- **Methods:** Latent Dirichlet Allocation

Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text by Julian McAuley and Jure Leskovec

- **Data:** review, user
- **Task:** "Hidden Factors as Topics" (HFT) model that uses (user) ratings and review text for product recommendations
- **Methods:** HFT model

Clustered Layout Word Cloud for User Generated Review by Ji Wang, Jian Zhao, Sheng Guo and Chris North

- **Data:** business, review, user
- **Task:** "present the clustered layout word cloud; a text visualization that would assist in making decision quicker based on user generated reviews"
- **Methods:** Grammatical dependency parsing; Clustered layout word cloud

Valence Constrains the Information Density of Messages by David W. Vinson and Rick Dale

- **Data:** review, user
- **Task:** "extend and explore the potential role of context in the observed information density of messages"
- **Methods:** Review-internal entropy; Average unigram information; Average conditional information; Conditional information variability

Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach by Jack Linshi

- **Data:** review
- **Task:** improve personalization of user star ratings using "an approximation of a modified LDA which conditions topics' term distributions not only on the Dirichlet parameter, but also on star ratings"
- **Methods:** Latent Dirichlet Allocation (modified, approximation of)

On the Efficiency of Social Recommender Networks by Felix W.

- **Data:** business, review, user
- **Task:** 1) "proposal of stochastic model for recommendation diffusion", and 2) "an algorithm for social recommender network optimization"
- **Methods:** Recommender systems models

(Note: The fifth round of the Yelp Dataset Challenge suggests using its dataset to research cultural trends; location mining & urban planning; seasonal trends; inferring categories; natural

language processing; and predicting attributes. Hence, these are all ways that Yelp data has been/is currently being studied!)

4.3 State-of-the-art Methods/Methods Borrowed

It appears that several of the most popular methods used to study this type of data are variations of recommender systems; latent factor models; Latent Dirichlet Allocation; and natural language processing. This makes sense based on the social recommendation nature of Yelp and the extremely large corpus of review text. Since our model was considerably simpler than the ones mentioned above, we did not use any of the relatively complex methods to make our predictions. However, like one of the winning submissions that placed heavy emphasis on feature selection using different regression models, we also used regression models to select our features. Additionally, our main feature – user "voteCount" – was also considered by others who used the user data.

5. RESULTS

We were able to tune our model's (Model 1) accuracy by changing the features we examined and the threshold that we compared the Linear Regression result to (See Section 3.2). To find the best features combination, we trained our Linear Regression Model and the threshold with different combinations until the accuracy of the predictions on the test data converged. The following table displays the prediction accuracy that resulted from different combinations of features in the feature matrix (X):

Table 2. This table displays the accuracy of the test predictions based on which features were included in the feature matrix when fitting the theta values for the Linear Regression Model.

Features Considered	Accuracy
reviewCount	0.96432
reviewCount & votes	0.96600
reviewCount & votes & friends	0.96687
reviewCount & votes & fans	0.96628
fans & votes	0.96207
reviewCount & votes & friends & fans	0.82838

As seen in the table, the combination of features that yielding the best prediction accuracy was 'reviewCount, 'votes', and 'friends'. This accuracy of 0.96687 translates into 2210 total missed calculations out of 66715 test entries. While this seems astoundingly accurate, it is important to remember that a prediction accuracy of 0.94314 can be achieved by just predicting 0 (or "user has never been Elite") for all users since only 3793 of the 66715 are Elite. Therefore our model (Model 1) was only 2.37% more accurate than the baseline.

The fitted theta values for this model were the following:

θ_0 = -0.00590570

θ_1 ('reviewCount') = 0.00719299

θ_2 ('votes') = -0.00007644

θ_3 ('friends') = 0.002220511

These values indicate that if reviewCount, number of votes, and number of friends are 0 then the user has been Elite for

-0.00590570 years and theses weights can be multiplied be each user's respective features to find the number of years they have been Elite. Obviously this "number of years a user has been Elite value" is useless to use if we want to predict a binary value (whether they have been Elite or not) and these thetas, or weights, applied to the users features may not always yield an accurate prediction. This is why we tuned our threshold on the validation data set so that we could have a more accurate threshold to compare this "number of years Elite value" to in the test data set to produce a binary prediction.

For the logistic regression model (Model 2), it was necessary to pre-compute the data. From the user data we extracted only the 'elite' values and the 'user_id', which was necessary to identify which reviews were written by a user from the review data. From the review data, we extracted 'text' and the 'user_id'--again, which was used to identify its author.

The initial values of theta for which fmin_bgfs was also fed were randomly selected to be a value, x, for -1<=x<=1, which then outputted the optimized theta values for which we used on the test data to predict whether or not a user had at some point been elite.

Often, some values for the initial theta used in fmin_bgfs yielded low results (as poor as .06 percent accuracy), but a re-initialization of the theta values was all that was needed to yield the .94 percent accuracy, which was the best this model could do; however, it could not outperform the accuracy of having simply predicted all users as having never been 'elite', and for this reason we concluded that first sentence length was not a feature that in any way affected whether or not a user had at some point been 'elite' for the given data.

6. CONCLUSION

It is frustrating to know that we only improved on the baseline by 2.73% but there are numerous reasons to consider. Perhaps we did not find the optimal combination of features to consider. In our Model 2 we attempted a different approach on features and instead looked at the average length of the first sentence of each user's reviews. However this proved to be an insufficient model because considering this feature by itself did not improve on the baseline.

While our approaches did not yield any astonishing improvement on the baseline predictions, they did provide insight on the importance of review count, vote count, and friend count when considering if a user has been elite. Admission to this exclusive club will remain at the discretion of Yelp and their subjective choices may be sometimes impossible to predict, but if one is to better their chances, they might start by boosting these stats.

7. REFERENCES

[1] Hood, Bryan, Victor Hwang, and Jennifer King. Inferring Future Business Attention. N.p., n.d. Web. 29 May 2015. <http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_InferringFuture.pdf>.