Predicting Yelp Elite Squad Membership

Timothy Mak, tkmak@ucsd.edu, A12825008
Angela Zhang, anz020@ucsd.edu, A13063547
Les Wang, lnwang@ucsd.edu, A12985002

INTRODUCTION

Yelp is an online and mobile community sourced search engine for local businesses, restaurants, and entertainment. For this assignment we utilized datasets provided by Yelp about Yelp users and their activity from 10 different metropolitan cities[1].

The feature of this dataset that we focused on is Yelp's Elite Squad program. According to Yelp, the purpose of this program is to "recognize those in the Yelp community who are active users and role models both off and on the site"[2]. Members of the Elite Squad are designated by an Elite badge their profile[2]. Being an Elite Squad member also boasts the perks of various free invite-only events.

In terms of how Elite Squad members are chosen, a user must nominate themselves or have another user nominate them in order to be considered for the Elite Squad. However after the user is nominated, Yelp makes the final decision as to who becomes part of the Elite Squad. There are some criteria that are provided by Yelp as to how they choose users, but the exact method for how Yelp decides is only known to Yelp itself. Using the data provided by Yelp, our objective is to approximate how Yelp selects specific users to be part of the this Elite Squad. This could help Yelp users hoping to join the Elite Squad understand what aspects of their profile and activity they should focus on to better their chances. It could also help Yelp make more consistent decisions on choosing Elite members.

I. Dataset

A. Data Attributes

The dataset that is used in this report is comprised of three different data files from the Yelp Data Challenge[1]. The three sets of data used in this report are on reviews, users, and tips. The following contains the user data properties that are relevant to this report:

- 'user_id' - unique user identifier
- 'name'- first name of user
- 'review_count' - number of reviews written
- 'yelping_since' - date when user joined Yelp
- 'friends' - list of user IDs of friends
- 'useful' - number of useful votes given by user
- 'funny' - number of funny votes given by user
- 'compliment' - number of compliments received
- 'average_stars' - average rating of all reviews
- 'elite'- list of years they were an elite.

In total there are 1,518,169 users that make up this data set[1].

The following properties make up a Yelp review:

- 'review_id' - unique review identifier
- 'user_id' - maps to review to specific user
- 'business_id' - maps review to business
- 'stars' - rating out of five stars
- 'date' - date review was posted
- 'text' - review itself
- 'useful' - number of useful votes received
- 'funny' - number of funny votes received
- 'cool' - number of cool votes received

The review data given consists of 5,996,996 reviews made by Yelp users[1].

The last set of data are tips, which are similar to reviews but are shorter and do not contain a star rating. Below are the properties that make up a tip:

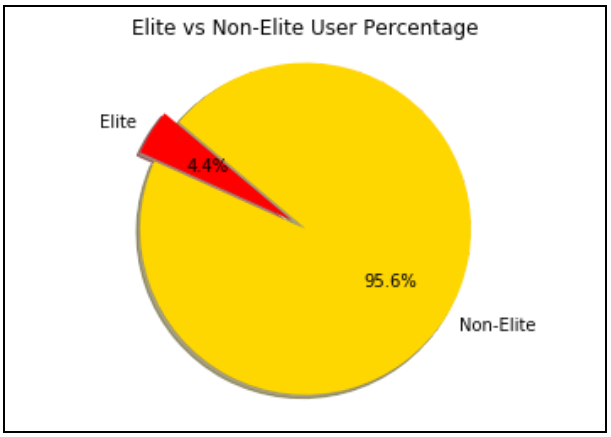
- ‘text’ - text of the tip
- ‘date’ - when tip was posted
- ‘likes’ - how many likes the tip received
- ‘business_id’ maps tip to business
- ‘user_id’ maps tip to user who wrote the tip

In total there are 1,185,348 tips in the data set written by Yelp users.

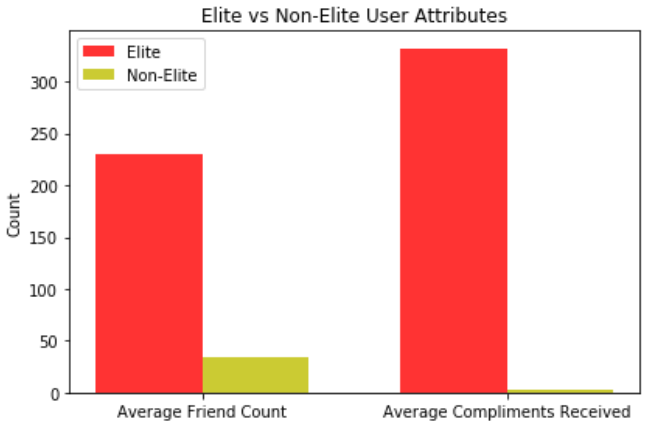
We were able to use all the user, review and tip data points in our logistic regression models without any issues of scalability or time constraints, but had to cut down for SVM.

B. Data Analysis

In terms of interesting findings from the data set, there were a couple of notable statistics derived from the data set. As shown in the graph below we found that only 67,109 or 4.4% of all created yelp accounts of the dataset are currently or have ever been part of the elite squad.

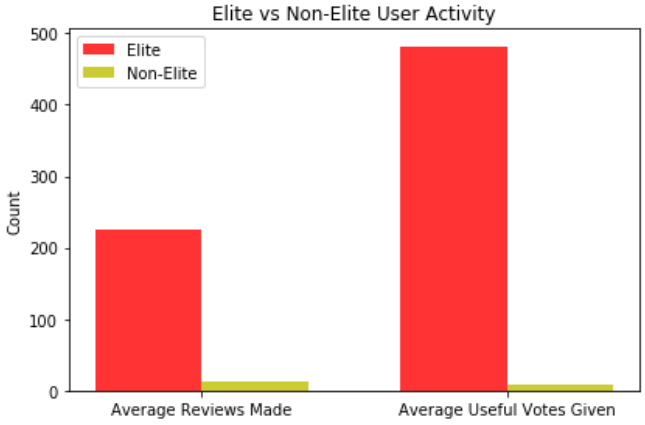


Additionally we found in our initial assessment of the dataset that there are significant differences between an Elite Squad user and a Non-Elite Squad user in terms user attributes. Specifically we found that there are vast differences with the number of friends and compliments received by other users. The following graph contains the findings of what we found on these differences:



Elite Squad members have significantly more friends on Yelp and receive on average a larger number of compliments from other users. It is important to note however, that when users connect their Yelp and Facebook accounts, Yelp considers all “friends” of the user who have Yelp accounts as “friends” of the user on Yelp as well.

Additionally, we discovered that the activity of each group of users varied vastly as well. The amount of activity was measured by how many reviews a user made and how many votes the user gave to other reviews besides their own. In the graph below, the average number of reviews and votes made by each group is shown:



Similar to user attributes, Elite Squad members tend to have more reviews made and useful votes given than Non-Elite Squad members.

II. Predictive Task

Our task - given information about a user, is to predict whether they are/have been a part of the Yelp Elite Squad. The Yelp website has an article about what the Yelp's Elite Squad is and explains some of the criteria that they look for. Some of the qualifications that they mention are *"well-written reviews, high quality tips, a detailed personal profile, an active voting and complimenting record, and a history of playing well with others"*[3]. We thought that it would be best to find user properties in the dataset that are highly correlated with these qualifications and base our model off of these properties.

For *"well-written reviews"*, we decided to consider two factors: the length of the review and the number of useful, funny, and cool votes that the review received. We noticed that a lot of Yelp Elites write reviews consisting of multiple paragraphs going over various aspects of the businesses such as the service, ambience, overall experience, etc. This is in contrast to non-elites who often write short blurbs. The way that we would extract the review length is similar to how we have been doing it in class. Also, if a review is to be considered *"well-written"*, then it would most likely have numerous votes. For elite-written reviews, many of them have numerous votes under all of the three categories (useful, funny, and cool) while non-elites usually have either a couple or no votes at all. Since each type of vote means different things and could have potentially different weights when considering whether it is *"well-written"*, we go through the review data and count these votes separately instead of summing up the total number of votes.

In order to create a feature for *"high quality tips"*, we approached it similarly to how we created the feature to determine *"well-written reviews"* in that we simply examined the number of likes the tip had. Since tips are meant to convey short suggestions, we chose not to factor in the length of the text.

When we were looking through the dataset properties to come up with a feature that would be correlated with *"a detailed personal profile"*, we came to the conclusion that there were no good indicators in the dataset that would explain this. Normally, a Yelp user profile includes an about page containing fields such as hometown, hobbies, favorite movies, etc. which would help classify whether a user has *"a detailed personal profile"*. However, since Yelp wanted to anonymize the data, they excluded these fields from the dataset.

For evaluating whether the user has *"an active voting and complimenting record"*, we decided to look at the number of votes that were sent by the user as well as the number of compliments that they received. Unlike when we were evaluating *"well-written reviews"*, we decided to consider the summation of votes sent compliments received. Since we are looking to see if a user is *"active"*, looking at the summation should be sufficient since the type of vote sent or compliment received has no effect on the activeness of the user.

Lastly, when evaluating users on their *"history of playing well with others"*, we decided to look at the number of friends that they had. This criteria is correlated with being connected to numerous Yelp friends since an elite would most likely be involved in their local community and have numerous interactions with people. We also hypothesized that an elite user would have a greater amount of Elite friends compared to a non-elite user, since one must be nominated to be considered for eliteness and Elites seem to be more likely to nominate others than non-elites.

III. Model

Since we are building a binary classifier, we decided to use two of the models that were mentioned in class[4]: Logistic Regressions and Support Vector Machines. We decided not to build the Naive Bayes model since computing probabilities based on our non-binary features would be difficult. Also, as mentioned in class, the Naive Bayes model runs into the problem of double-counting. After we built the two predictors, we would choose the one with the better prediction and optimize that predictor further.

A. Baseline Models

In order to have something to compare these models to, we created two baseline models. One of them always predicted that the user is not a Yelp Elite. As seen in our data analysis, the ratio of non-elite to Elite members is so large that the model actually performs quite well with an accuracy of 0.95579. The other baseline model we used is one in which we calculate the feature averages of elites and non-elites and base the prediction on whether a user's features are closer to the elite averages compared to the non-elite averages. For this baseline, we used 4 of the primary features from the user dataset: number of reviews, number of friends, number of compliments received, and number of votes given. This gave us an accuracy of 0.95996, which was slightly better than our other baseline.

B. Initial Model Issues

When we first tested our models, we noticed that they were both performing worse than our baseline models. After further investigation, we found that the underperformance was being caused by the fact that the weights that were associated with our features were insignificantly small. This was caused by our training set having so little elite user data points. In terms of our training set, we initially just took the first 50,000 user data points. However, we forgot to take into account the fact that the ratio of elites to non-elites was so low in our training set which caused our models to underestimate the effect that the features had on our predictions. So, we decided take a different approach and use k-fold cross-validation with 10 folds to split the training and testing sets multiple times for each model. This way, we were able to employ every data point in the user, tip, and review datasets as both training and testing data at some point to produce the best estimate of mean squared error, accuracy, precision, and recall. After changing our validation pipeline, both of our models outperformed the baseline models.

C. Multiple Logistic Regression

For the logistic regression model, we started with 6 basic features from the user data: number of reviews, number of friends, number of fans, number of compliments received, number of votes given, and average rating given. We first performed a logistic regression with each of these features separately to determine the "correlation" each feature had with eliteness ("correlation" in quotes since we aren't determining a linear relationship but more so an association in general). We then used a forward stepwise selection process to determine the best combination of features that produced the most accurate model using multiple logistic regression.

The next features we decided to add were more metrics regarding friends, specifically the number of Elite friends a user had, and the ratio of Elite to non-elite friends. Then we brought in the tip dataset to create features from the number of tips a user has written, the number of likes a user's tips has received, and the ratio of likes to tips. Finally, we used the review dataset to build 8 more features: average review length, number of Useful votes received, number of Cool votes received, number of Funny votes received, number of reviews with paragraphs, number of reviews with bullet points or numbering, number of reviews with pros/cons, and number of reviews with colons.

D. Support Vector Machines

Using the feature set that had the best results with logistic regression, we used the same function for running a Support Vector Machine model to see if we could improve upon our result even more. For this model, we had to cut down the dataset to 25000 Elite users and 25000 non-Elite reviews in order for the model to run in a timely fashion.

E. Principal Component Analysis

To further understand the importance of each feature and see if we could cut down without losing accuracy, we also ran PCA on all 19 features. We then fit a series of logistic models on the top k=1,2,...,19 features using 10-fold cross validation to determine the one with the lowest mean squared error.

IV. Literature

The dataset that we are using came from the Yelp Open Dataset and is used to "teach students about

databases, to learn NLP, or for sample production data while you learn how to make mobile apps"[1]. There are no similar datasets since it is specific only to Yelp. The state-of-the-art methods in this research area that are currently being used are logistic regression and support vector machines.

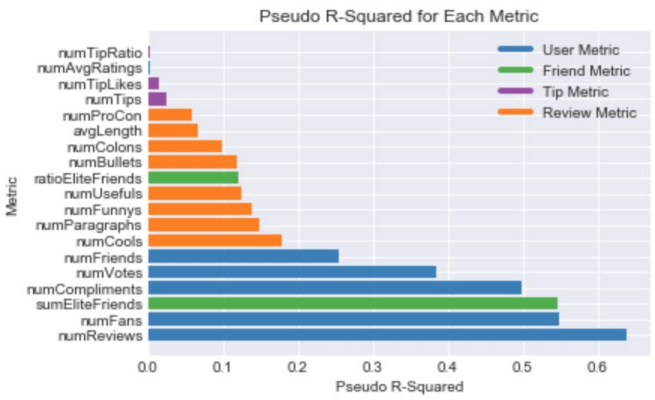
In terms of other research papers that explore this topic, we actually found a paper done by students from UC San Diego[5]. They also used similar predictors and data properties. One notable difference was that they factored in locality since elites are more likely to have written reviews for local businesses which is absent from our dataset. They also took an interesting approach for one of their features in which they looked at the problem as a community detection problem. One of the perks of being a Yelp Elite is that they are invited to exclusive Yelp events where they get to meet each other. Therefore, it's more likely that elites will have a lot of elite friends.

Another piece of literature that did similar research was an article by Diana Lam who based her features off of a user's activity, popularity, and reviews[6]. When looking at the reviews, she took a similar approach to what we did in class which was that she found the top 50 unique words used by elites compared to non-elites. She concluded that the most important factor to consider when determining if a user is an elite is there popularity. Overall, her approach was similar to ours but she placed more of an emphasis on analyzing the review content and structure. She also used a Random Forest Classifier as her final model, which we did not try as it seemed out of our scope.

V. Results and Conclusion

A. Feature Importance

We interpreted the Pseudo R-Squared statistic of each logistic regression as a "correlation coefficient" of sorts to determine the importance of certain features and groups of features we used. For all features except Average Rating, Elite Friend Ratio and Tip Ratio we normalized the data with a log normalization.



In general, user metrics had the highest correlation followed by friend, review then tip metrics.

User Metrics	**Pseudo R-Squared**
Number of Reviews	0.6377
Number of Friends	0.2536
Number of Compliments	0.4974
Number of Votes	0.3832
Number of Fans	0.5476
Average Rating	0.002275

The user metric with the highest correlation was number of reviews, followed by number of fans and then number of compliments. This matches our intuition, since writing reviews is the primary way of contributing on Yelp it makes sense that it would be factored in with higher importance when determining Eliteness. Fans and compliments are both features that do not seem to be widely used on Yelp, so only very active (and more likely elite users) would be the ones to have any activity in those respects. The user metrics with the least impact were average rating and number of friends. These both make sense as well, since average rating is on a closed scale and number of friends varies for Facebook connected users.

Additional Friend Metrics	**Pseudo R-Squared**
Number of Elite Friends	0.5460
Ratio of Elite to Non-elite Friends	0.1193

From the additional friend metrics however, we see that the number of Elite friends does in fact have a higher correlation. We can attribute this to the fact that Eliteness is based on nominations, and Elite users are more likely to give nominations to friends.

Tip Metrics	**Pseudo R-Squared**
Number of Tips	0.02428
Number of Tip Likes	0.01398
Ratio of Tip Likes to Tips	0.001400

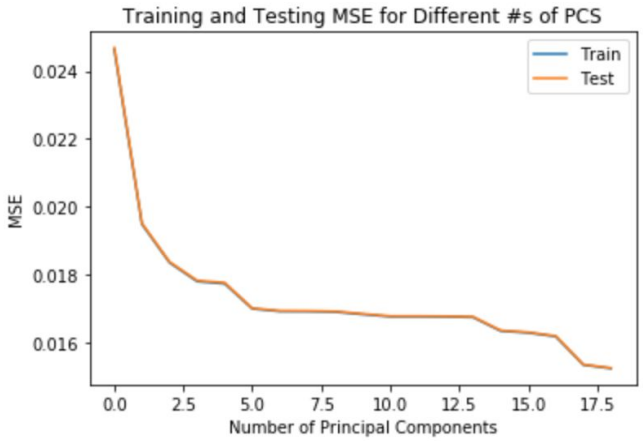
The tip metrics did not seem to affect Eliteness as much, possibly because they are short and easy to write for non-elite members as well.

Review Metrics	**Pseudo R-Squared**
Average Review Length	0.06551
Number of Useful Votes	0.1235
Number of Cool Votes	0.1786
Number of Funny Votes	0.1374
Number of Reviews with Paragraphs	0.1477
Number of Reviews with Bullet Points/Numbering	0.1183
Number of Reviews with Pros/Cons	0.05768
Number of Reviews with Colons	0.09875

Out of the review metrics, the most important ones were number of Cool votes, Funny votes and Useful votes received followed by number of reviews with paragraphs and bullet points/numbering. The amount of recognition a user receives for their reviews through votes can be attributed to how "well-written" it was, and paragraphs/bullet points are both signs of well-thought out and formally written reviews. The latter two can also be used the weed out the long reviews that are just rants or complaints, versus the ones that are actually well-written.

Metrics Used	**Pseudo R-Squared**
User Metrics Only	0.6958
User and Friend Metrics	0.7152
User, Friend and Tip Metrics	0.7160
User, Friend, Tip and Review Metrics	0.7210

Adding on the various types of metrics was able to improve our R-Squared minutely, but the most important group of features by far was still the User Metrics we began with.



Using PCA, we were able to determine how big of a difference each principal component really made on our predictive accuracy in the end. We can see that although the MSE continues to decline as the number of principal components we use increases, it does seem to level out around between 0.018 and 0.016 where the slope becomes less steep. This tells us that that once we're using 5 or 6 principal components, adding on more does not improve our accuracy by much. So if we wanted to reduce the dimensionality of our dataset, we could restrict to using the first 5 or 6 PCs while still maintaining a high accuracy.

B. Evaluating Performance

Using multiple logistic regression with just the User Metrics, we were able to achieve 0.976403 testing accuracy, which accounts for around 47% of the missed accuracy from our initial baseline model. Each of the additional types of metrics improved upon the accuracy marginally, giving us 0.977963 including friends and

tip metrics, and 0.978341 including review metrics. This final accuracy accounts for over 51% of the missed accuracy from baseline.

SVM ended up being a lot less accurate than we originally thought, worse than the baseline in fact. This may be due to us having to cut down the dataset size in order to accommodate the model, since SVM is a much more complex and time consuming model.

	Logistic Regression with User Metrics	*Logistic Regression with All Metrics*	*SVM*
Accuracy	0.976403	0.978341	0.632
Precision	0.773815	0.749589	0.637
Recall	0.552789	0.765922	0.626
MSE	0.021246	0.015287	0.367

While support vector machines try to find the biggest margin between positive and negative data points, the regression model places weights on each feature and bases its predictions on whether those features are present. Using logistic regression would result in a higher accuracy if elite badges are awarded based off of some checklist of user characteristics that one must have in order to qualify as being a Yelp elite. Since there are cases where some users have lots of friends but do not compliment as much or instances where users write high quality reviews but in various locations, this would cause the positive and negative data points to be randomized and close to each other when using support vector machines to find the maximized margin between elites and non-elites.

C. Final Model

Our final and most accurate model used logistic regression. The feature combination with highest accuracy was: number of reviews, number of friends, number of compliments, number of votes given, number of fans, average rating, number of elite friends, number of tip likes, average length of reviews, number of useful votes, number of funny votes, number of cool votes, number of reviews with paragraphs, and number of reviews with bullets. Although some of these features did not have very high R-Squared values to begin with, they still gave us a slight increase in accuracy so we decided to include them. Since logistic is a relatively fast model, we did not have to sacrifice much time in order to include extra, marginally-useful features so we decided to include as many as possible without decreasing our accuracy.

D. Future Work and Possible Improvements

An important aspect of Yelp that we were not able to capture in our model is the influence of photos. While there is a photos dataset, it does not include data about the user that posted it and whether it was a "good" or "bad" photo. Taking pictures of a business has a huge impact on its rating since it gives potential customers the chance to see what their experience there would be like. However, since that information is not a part of the dataset, we cannot create those features. Also as mentioned before in Section II, we do not have the user data to determine whether their profile is detailed enough. If we were to continue doing research in this area, we would factor in the features that were mentioned in the literature that we did not include.

References

[1] https://www.yelp.com/dataset
[2] https://www.yelp.com/elite
[3] https://www.yelp-support.com/article/What-is-Yelps-Elite-Squad?l=en_US
[4] http://cseweb.ucsd.edu/classes/fa18/cse158-a/slides/lecture3.pdf
[5] https://pdfs.semanticscholar.org/bfd2/59f54861a9551858041bb17e249df3cb07d8.pdf
[6] https://www.dianalam.com/2016/02/28/yelp-classification.html