Exploring review quality and social network influence in the Yelp Academic Dataset

Taylor Dahlke (tjdahlke), Giovanni Campagna (gcampagn), Tamirlan Seidakhmetov (tamirlan)

Introduction

The primary interest of Yelp is to provide high quality reviews that are useful, insightful, and entertaining. Yelp users can vote on reviews by tagging them as *useful*, *funny* or *cool*. It's in the interest of Yelp to find out which users produce the best reviews based on these metrics, so that they can reward those users with "elite status" and other social perks. As a business, Yelp does best when users are creating reviews that are useful to other users, and wants to encourage that. Because of this, much of our work focused on one question: Can we build a model to help predict which users make the most useful reviews? We explore what variables we need to make such a prediction, with an emphasis on learning how social network information can help us better answer that question.

Dataset Findings

Every year, Yelp releases a dataset to be used in their academic challenge competition. This data set is extensive, and geographically covers eight cities in North America, and three in Europe. The dataset includes 2,225,213 reviews, for 77,445 businesses and 552,339 users. Most all of the information that is publically available on the Yelp website is included in the attributes for each of these three datasets. In particular, a subset of the friendship graph is included in the users table.

The dataset was selected primarily by geographical location. We find that many users are missing reviews (total count of user reviews is higher than the number of reviews given for that user in the dataset). We expect this is because users also review businesses that lie outside the selected cities (when traveling, etc). Fortunately, subsampling is not found within the social network established in the dataset. This points to the fact that for the cities selected, most of the users have Yelp friends within their immediate area. Nevertheless, aggregate attributes that reflect the real data are present for businesses and users.

Approach

We are most interested in predicting the normalized usefulness of a user (# of useful votes over all reviews / # of reviews), and are interested as to whether we can make use of derived attributes to better approximate that. We began by building a number of attributes from the business/ review data collected for each user such as:

- Relative usefulness of reviews when compared to other reviews for a business (mean and standard deviation).
- Relative earliness of review posting date when compared to other reviews for the same business (mean and standard deviation).
- Latitude / longitude of review locations for a user (mean and standard deviation).
- How traveled a user is (max of latitude and longitude standard deviation).

Next, we derived social network information for each user. To do this, we first selected from the original data all the users who have at least one friend (about 45% of all users). Then we searched through the friends of each user to build distributions of attributes. For example, a distribution of the number of fans that each friend of a user had. Then, we split these distributions into discrete bins that were globally defined for all users. Below are the categories of social network attributes that we built.

- Total size of user's second order network
- Second order network size for friends (binned distribution)
- Number of days since joining Yelp (binned distribution)
- Number of "good writer", "good photos", and "good list" compliments (binned distribution)
- Number of reviews for friends (binned distribution)
- Number of elite friends
- How far away friends are (distances from user to friend average review location), mean and standard deviation.
- How traveled friends are (binned distribution)

Also we looked at various statistics of friends' network using original dataset, particularly the mean and standard deviation, as well as the minimum and maximum of the number of friend's reviews, fans, useful votes and average stars for reviews.

Structural analysis of the user dataset

We begin with analysis of our dataset to better understand what attributes might allow us to best make predictions about a user's usefulness and elite status. Table 1 shows the top 5 most correlated attributes for each response variable.

Table 1: Correlation values of attributes with "normalized usefulness" and 'binary elite'

Given + social network derived attributes			
Correlation with "norm usefulness"		Correlation with "elite"	
Attribute		Attribute	
mean relative usefulness	**0.48**	**friends' max # of fans**	**0.63**
# of cool votes	0.36	**friends' max # of useful votes**	**0.62**
# of funny votes	0.36	# of reviews	0.58
Size of 2nd order network	**0.30**	**friends' max # of reviews**	**0.56**
# friends in 95 percentile for "good photo" compliments	**0.30**	**Size of 2nd order network**	**0.37**

By finding the correlation between the normalized usefulness and other attributes, we were able to get a rough idea of the most influential variables, most of which appeared to be attributes derived from social network data (in bold).

PCA

The next way that we chose to analyze our data set was by performing principal component analysis (PCA). We found by doing PCA that most of the variance (about 65%) was accounted for by the first four principal components (Figure 1).

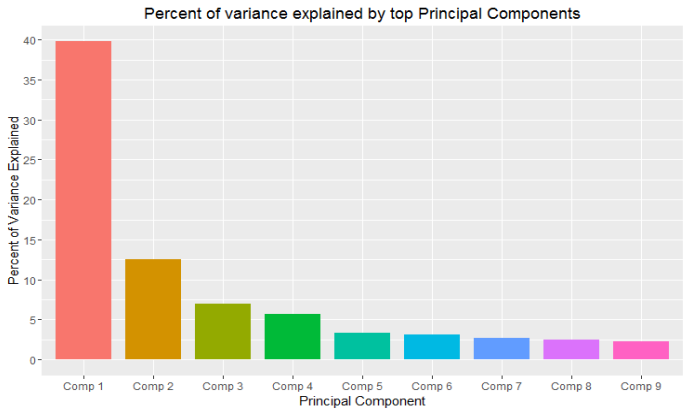


Figure 1. Percent variance explained by each Principal Component

Clustering

One hypothesis that we made after heuristically reviewing different users in our dataset and on the Yelp website itself, was that elite users tend to be friends with other elite users. Many of them seem to meet a special events hosted by Yelp. Further, many elite users seem to get compliments from these same elite friends. We also find that elite users tend to be more helpful (average normalized usefulness of 2.12, while non-elite have a mean of 1.20). In order to confirm our hypothesis that users are split in two rough clusters of non-useful, non-elite users, and useful elite users, we run K-Means on the data, with different cluster numbers. We split the data in a training set, on which we trained the K-Means algorithm, and a test test, which we assigned to the nearest cluster center, and we measured the test error on an average of 10 splits, with 10 initialization of K-Means each time. Data was preprocessed by standardization and PCA with various numbers of components, as well as log-scaling for attributes that would grow with the number of friends and the number of reviews, because they would naturally vary significantly. Not surprisingly, the error was observed to decrease as the number of cluster increased, but we found a local minimum with 8 clusters.

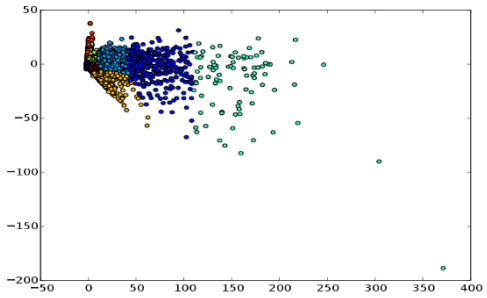


Figure 2. Clusters with K = 8, using 4 PC

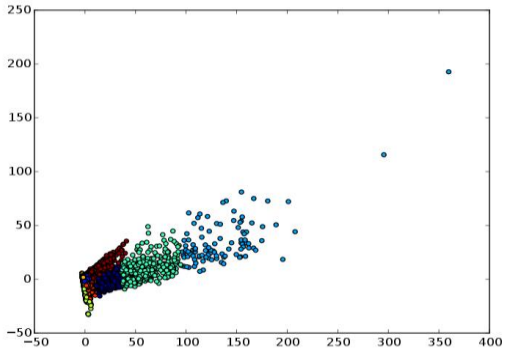


Figure 3. Clusters with K = 8, after PCA 4 preprocessing and log-scaling of large attributes

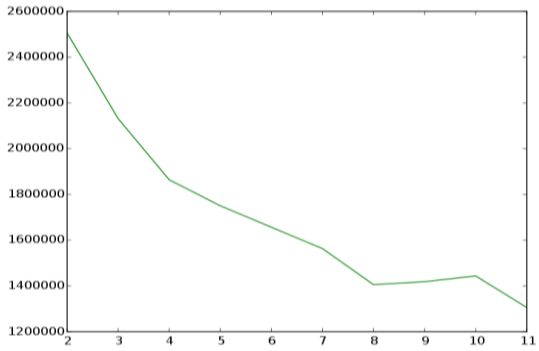


Figure 4. Test error (total euclidean distance to the centers) of K-Means with different K's

Empirical analysis of the cluster centers with K = 8 and PCA 4 preprocessing suggests that the 4 more marked clusters tend to be affected the most by the compliment quantiles attributes in the first order network (that is, the number of friends with a given, binned, number of compliments), suggesting that friends with the same number of compliments are close together. This seemed to agree with our hypothesis regarding the clustering of elite users / "patting each other's back" with compliments.

Separability of elite binary

Following our initial hypothesis that elite status provides some structural separability to the users, we trained a supervised model to predict the elite status of the user, given all other attributes.

We chose SVM with a radial basis function (RBF) kernel, because that would limit our bias and extract the most from our features. We used a training set of 53,877, because the full dataset was too computationally intensive, and because we rebalanced to make sure we had 50% elite and 50% not elite (despite having only about 5% elite in the original data). We obtained the following results (where precision and recall should be intended for a elite = +1):

Table 2. Summary for prediction of elite users

	Accuracy	Precision	Recall
Train	93.5%	91.9%	94.6%
Test	93.2%	92.8%	94.6%

This very high accuracy suggests there is a clear separation for elite and not elite, which we can also see by plotting the users in PCA 2 space. The graph also shows the decision boundary of the SVM, although a large number of points appear to be misclassified because the decision boundary is plotted in PCA 2 space, but the SVM is trained on the full dimensionality of the data.

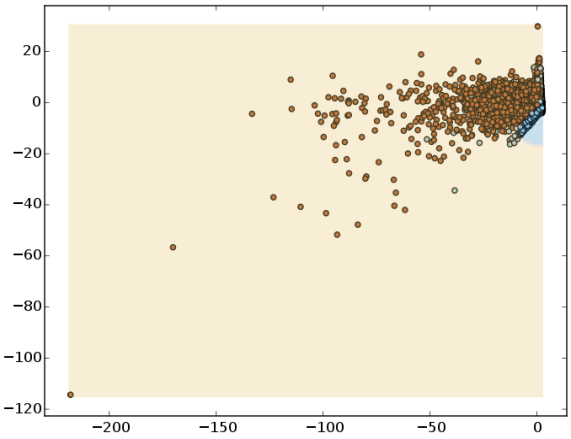


Figure 5. Elite binary, with SVM decision boundary (brown elite, blue not elite; PCA with 2 components for illustration purposes)

Graph analysis of the user social network

We built a graph from the social network data, linking each user to each other. Using this graph, we were able to run an assortativity metric to understand the preference of users to befriend others based on certain attributes. In this case, positive values indicate a preference toward clustering based on that attribute, while a negative number indicates the opposite. While none of the attributes had particularly strong values, the relative ranking of attributes

still tells us something about how users are related to their friends.

We found that users tend to connect to each other when they have similar spatial extent of where they give reviews; i.e, users who seem to travel a lot tend to be friends, and those who mostly give reviews locally befriend users who do the same. We also found that users tend to be friends with users who share a similar latitude spatially. Also, the relative earliness of reviewing a business compared to other users who reviewed a business was an attribute that clustering occurred on (i.e; hipsters tend to group with other hipsters, and trend followers tend to group with other trend followers). We also found smaller grouping strengths based on users being elite or not, as well as based on how useful a user typically is. This suggests to us a possible link between the social network derived attributes, and the ability to predict normalized usefulness and elite status, since these attributes tend to cluster in friend groups. Therefore, information about the friend group could be useful to predicting these attributes. We later find this to be the case.

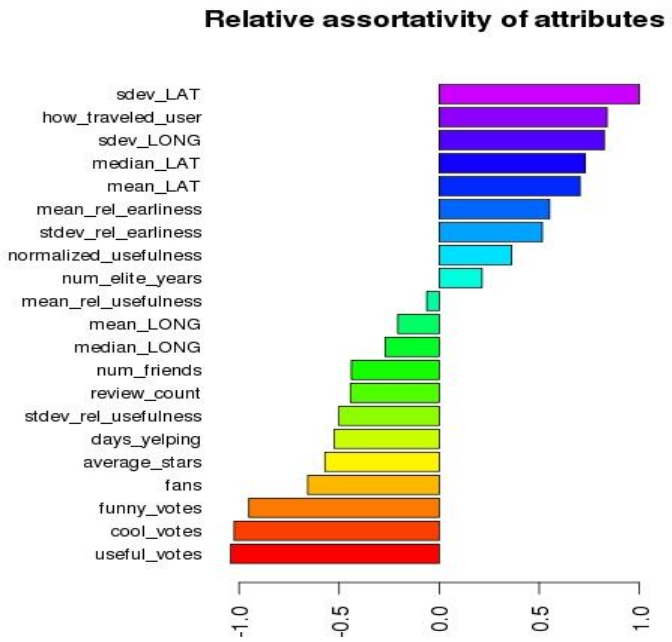


Figure 6. Relative assortativity of attributes

Predicting usefulness of a user

Linear and Quadratic Regression: Methodology

We test supervised learning techniques to predict the normalized review usefulness (#useful review votes for user / # user's reviews) for each user. We started by making two datasets, one with the top 5 given attributes, and one with the top five given+derived attributes (see Table 1). We split our datasets again into 75% training and 25% test, and then did 10 fold cross-validation doing both linear regression and quadratic regression to generate mean-squared error results (see Table 3). We found that we are able to get a much better fit to our data using the dataset that included the derived attributes, with better R^2 values, as well as lower mean-squared error (MSE) on our test set.

Linear and Quadratic Regression: Results

Table 3: Regression results

Data type		R^2	MSE
Linear Reg	Given attributes	0.14	2.75
	Given + derived attributes	0.282	2.32
Quadratic Reg.	Given attributes	0.166	2.41
	Given + derived attributes	0.304	2.02

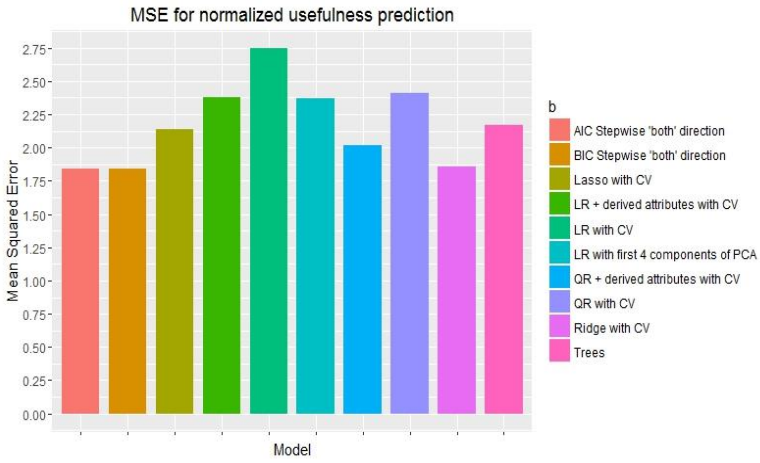


Figure 7. Summary of regression analysis

Feature Selection

In order to choose only important features and avoid overfitting, stepwise regression in both directions, which includes both backward and forward regression was done. As a metric to choose the best model Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used. AIC chose 50 out and BIC chose 42 of total 64 attributes. Then Linear Regression was run. In fact, both models showed similar performance and outperformed all other models (Figure 7).

Lasso and Ridge Regression

Another regularization methods that were used to avoid overfitting by penalizing having many attributes in the model. Despite the fact that Lasso model had fewer non-zero attributes, Ridge regression showed almost the same performance as stepwise regression models and outperformed Lasso regression (Figure 7).

Regression Trees

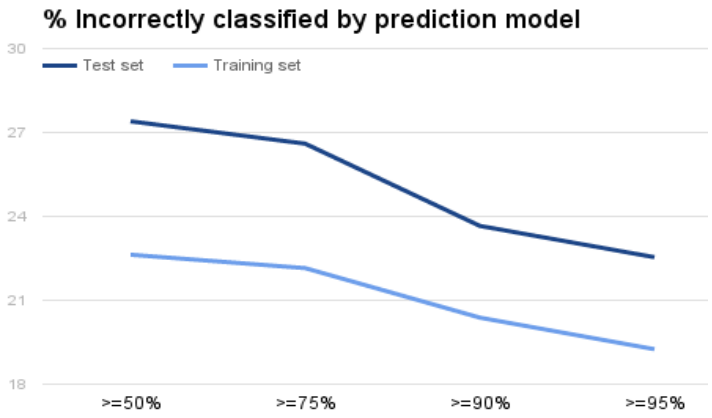
Regression trees was used, however it is a very computationally expensive model. So parameters were chosen heuristically, by making some optimization. Optimal parameters are bucket number = 42, cp = 0.005806, split = 43. However performance of regression trees was not satisfactory (Figure 7).

Support Vector Machines: Methodology

Because we had limited success predicting the normalized usefulness of a user when we treated the response as a continuous variable, we changed approach. Since it is most interesting to identify users who have high normalized usefulness, we decided to split along a quantile threshold for normalized usefulness. For example, we assign a response of 1 to users who had a usefulness above the the 75th percentile, and 0 to the rest. Then we randomly sampled from this data set to make test and training sets that were composed of equal parts most useful (1) and least useful (0). We made our training and test sets the same size (1000 users each). We then fit a SVM model to these datasets (using a linear kernel, which we found worked the best). This was done with 25 fold cross validation to get the error values shown. We repeated this process for varying splits of the dataset along percentiles of usefulness, which we plot below.

Support Vector Machines: Results

We found that we were able to get increasingly better predictions of a user's usefulness when we were trying to predict for users that had higher normalized usefulness. This suggests to us that the more useful users are better separated from the rest. The figure below shows this decrease of error in both the test and training sets when we attempt to predict the more useful users.



Conclusions

We began our project with the hypothesis that we might be able to predict user characteristics, particularly usefulness, by using derived attributes from review and social graph information. Cursory analytics like simple correlations between variables showed us that these derived attributes had higher correlations with usefulness than many of the given attributes. More advanced unsupervised learning methods like PCA supported our theory of gaining added value from these derived attributes by showing us that these attributes explained much of the variance in our data. In practice, the application of using these attributes was modestly successful. Our linear and quadratic regression showed that the attributes do indeed help us better predict usefulness as a continuous variable. Further, our SVM application showed us that some levels of usefulness are easier to predict than others. Our clustering and elite classification experiments showed us that our data had higher separability on compliments and elite status, both of which are loosely associated with normalized usefulness. We later confirmed by doing associativity on our social network graph that the attributes that tend to cluster users include normalized usefulness and number of years being elite, suggesting a link between the separability we find among elite / non-elite users and the influence of social network data in helping to predict normalized usefulness. While being able to predict the usefulness of a user is not a simple thing to accomplish, we suggest through our statistical analysis, unsupervised learning methods, and regression / classification experiments that our network derived attributes might be helpful in more accurately reaching this goal.