Ego Influence: Impact on Business ratings or not

Namesh R. Kher
Master's Information Management
College of Information Studies
University of Maryland, College Park
nkher@umd.edu

Sarika S. Hegde
Master's Information Management
College of Information Studies
University of Maryland, College Park
Sarika1@umd.edu

ABSTRACT

Yelp is a crowd sourced local business review and social networking site, where users post their reviews about a business using a star rating system on a scale of one to five. Our research is based on two things; firstly trying to understand the factors that affect the ratings that a business gets and secondly, since Yelp is a social network in itself, we try to understand how a user who has reviewed a business is influenced by his ego network and if that too could be one of the reasons why a person would prefer going to a restaurant. Often businesses like restaurants take paper based feedback from the customers that visit them. We use the online reviews from Yelp and train a model on it, through which we predict if a user would review or give feedback for a business based on the users friend's opinions.

This study is one of the first ones to explore the social network in a website like Yelp. In our research, we define a term *Ego Influence* as the combination of the following factors:

i. Social Influence
ii. Social Sentiment
iii. Closeness with friends in network

We derived some significant results from our study:

a. Our regression model suggests that *Closeness with your friends* (Personalized Page Rank) is a very strong factor that helps predicting the ratings that a business receives.
b. Besides closeness; Price, Average User review count and number of fans that a user has also form good predictors.
c. Our classifier gives 68% accuracy for predicting whether a user will review a restaurant or not depending on influence of his friends.

INTRODUCTION

Social networks have been around from a very long time now however studying social networks begun recently over a decade ago [16]. A social network comprises of a set of users, which are called social actors and these actors have relations in between them. Study of such data structures related to the field of social network analysis which help us know more about a network or community by revealing local patterns, global patterns and many network dynamics. Our focus in this paper involves social network analysis by measuring how does a users community influence the user.

With growing popularity of social networks where a users opinion has a huge influence on his peers and vice versa [1] (Muchnik, Aral, & Taylor, 2013), this research is focused on understanding how an ego network of user affects his opinion about a business. Unlike previous research papers which focused on Semantic Topic modeling approach, that concludes that the LDA model on text reviews can help predict average rating [3]. Our sentiment analysis on text reviews is based on a unigram model, which is discussed in detail in the Feature Collection Section (Part II). We applied basic text mining algorithm on the reviews to categorize it as a positive one, a negative one or no review. Alongside we also explore the closeness within the user's friend graph to understand how connected they are to each other. We define the closeness of friend to his network as the *"Personalized Page Rank Value"* that the user receives from that graph. The last feature that we collect for this analysis is the Social Influence factor for each user in our final dataset. This factor measures the users behavior towards a particular business based on his social network.

This paper is an approach to exploring the Yelp user graph and thereby measuring the extent to which users on Yelp get attracted towards a business. We choose to work with only the restaurant businesses and hence our research questions for our research problem are framed as follows:

1. **'*Given an ego node's sentiment score for a business, its personalized page rank (Closeness) and social influence measure can we predict if a user has given a review or not*'**

2. ***'What factors influence a persons visit to a restaurant'***

To explore the above, we start with explaining about the nature of the Yelp data by cleaning it to match it to our research problem and then perform some initial preliminary analysis on it. At the end of this stage we have our final data samples ready. We work on collecting our 3 main feature variables, which is described below, and then append these to our final samples. We also explain about the motivation for our research followed by some of our limitations. The features collection sections forms the methodology of how we collect the three variables we previously described. We also make use of other features provided by Yelp and finally combine all of them to develop statistical models to evaluate how well does our final data set help in answering our research questions. We also discuss a few potential applications that Yelp might consider using to enhance their business.

DATA

The Yelp dataset from the website [2], has five JSON objects with Business, User, Review, Tip and Check-in data of the user. To work with this big data problem we put Mongo DB to use to come up with an algorithm that would help put together an unbiased dataset combining the business, user and their review data. We excluded the tip and check-in data since they serve to purpose of the project.

From the actual data of about 1.6M reviews, 481K businesses and 366K users for a total of 2.9 M social edges, we started by extracting only the businesses that had restaurant as a category. At the end we were left with just over 21.5K restaurant businesses. Since we were trying to measure social influence we selected the users form the graph that at least have one friend in

their adjacency list. If a user does not have friends would signify that a user is a dangling node without outgoing or in coming edges and hence there cannot be any social influence for that user. Our final user database had 174K users out of the 366K users. To be precise, our final dataset consisted of 21,799 businesses that corresponded to 880287 restaurant business user reviews with a total of 174094 users.

INITIAL DATA SAMPLE FORMATION

Once we were ready with our cleaned dataset we begun with performing some initial preliminary analysis. We utilized the power of Mongo DB as a non-relational document oriented database where all our data resided. We used Mongo DB query language, which is actually Java Script based it self and hence fits the JSON format data provided by Yelp. In order to collect study our research problem we start making our samples, from our cleaned data set explained above. The ***base approach*** of our sample formation lies in collecting ***(uid, bid)*** combinations where a user might or might not have reviewed for a particular business. To these samples we add a column called ***Reviewed***, which is either a 1 or 0 signifying that the user has reviewed for that business, or not respectively. For collecting such samples we first study our dataset by running a few mongo queries and have the following observations:

1. Top 500 restaurant businesses had at least about 284 reviews at minimum and top 300 businesses had at least 381 reviews.

2. The most popular business had about 4137 reviews and least popular ones had single or no reviews.

3. Most of the users in the dataset had very few friends in their network and very few had large number friends.

Users	Friend List Size
150,000	Between [1, 25]
About 10,000	Between [26, 500]
Rest of them	Between [501, 3830]
Total users: 174,094	

Figure 1. The table shows distribution of the friend list sizes for the users in the cleaned yelp dataset. Only 25K of the users have more than 25 friends in their social network.

Figure 1 shows the distribution of the friend list sizes for the users in our cleaned yelp graph. We observe that only 25K of the users had more than 25 friends in their networks, which is about only 14.2% of all the users. After these quick observations we began sampling our data and built three different approaches for it which were based on our base approach as explained above. We finally went ahead with the last one.

Approach 1

1) *Build a set of top N reviewed businesses -> B {bids}*
2) *Get mix of different sized ego networks*
   a) *Small (Less than 50 friends)*
   b) *Large (> 200 && < 500)*
   c) *Medium (> 500 && < 1000)*
   d) *Very Large (> 1000 && < 3830)*
3) *Build a set of above users -> U {uids}*
4) *Cross multiply B and S*
5) *Add a column called Reviewed 'R'. Place a 1 if a review exists for the (uid, bid) combination else 0.*

The above approach was inspired by the biased yelp graph that we had. Above 85% of the users had few friends in the network and hence we wanted to get mix user samples. By this we mean that we tried obtaining users who had a large friend list, a small one and medium sized one. Eventually we discarded this approach because maximum of our collected samples i.e., ***(uid, bid)*** combinations had more negative samples (0 in the Reviewed column) than positive ones. We wanted to have a near to equal number of these and hence went ahead with another approach.

Approach 2

1) *Get mix of different sized ego networks (Small, Medium, Large) and form a set of users -> U {uids}*
2) *Get the businesses for which the above users have reviewed and add them to the set of businesses -> B {bids}*
3) *Get the users of the above users and also add them to the set of users -> U {uids}*
4) *Cross multiply B and S*
5) *Add a column called Reviewed 'R'. Place a 1 if a review exists for the (uid, bid) combination else 0.*

The objective of the above approach was to get a fair number of positive and negative samples. This was on the assumption that if a user has reviewed for a business then the users friends might also have which related to what we are trying to find out. However when analyzing the collected samples we again observe more number of negative samples (Many users did not review for a business). Our final approach is as follows.

Approach 3 – Final

1) *Get N businesses, which have, review counts within a particular range and add them to the business set -> B {bids}*
2) *Store the corresponding users who have reviewed for the above bids is the users set -> U {uids}*
3) *Start writing 'X' positive (uid, bid) samples using the above created sets.*
4) *Start writing 'Y' negative (uid, bid) samples using the above created sets where $Y \approx X$*

For collecting equal positive and negative samples we follow a reverse approach where first collect samples where a user has reviewed for a business and store 'X' samples of the same. As we collect samples we store the uids and bids in their respective sets. After collecting the positive samples we collect 'Y' negative samples for those *(uid, bid)* combinations. We repeat the above steps for a different range of businesses in terms of their review counts. For instance first we do this for businesses having review counts between 3500 min and 4000 max, and then for businesses having review counts between 3000 and 3500 and so on.

Reason for doing this: Initial analysis on the business and corresponding reviews for the same showed that very few businesses had reviews above 2000. Most of them had less than 1000 reviews. Hence to keep the dataset unbiased and unbiased only businesses having many reviews; we divided the businesses

in to different ranges based on number of reviews they had received and considered certain number of businesses from every range. We selected the users who had given the reviews for business as well as those who did not. At the end of our sample collection we had about 155K samples if [*(uid, bid) combinations with Reviewed column]* with an al most equal number of positive and negative samples. In the feature collection phase we work on creating the three variables we explained above and add it to the final samples. We also make use of other variables provided by the yelp dataset.

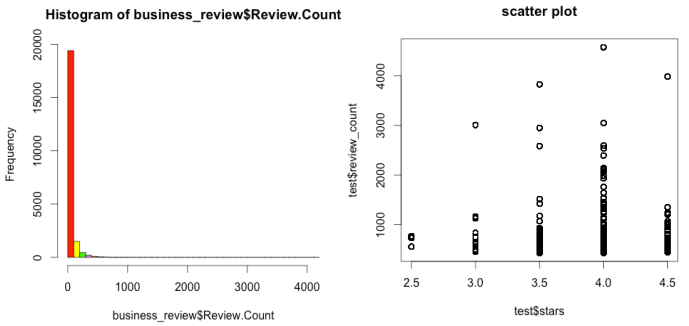

Figure 2 a): Businesses with many reviews have got a rating of 4-5 Figure 2 b): Majority businesses have very few reviews

Figure 3. Google Map plots using R to show the restaurant locations considered for the research. Insight: Arizona showed the highest number of restaurant businesses, restaurants with highest reviews were in Las Vegas. Thus location could also be used as a factor to understand the impact on business for further studies.

MOTIVATION

A lot of research has been done on understanding what affects a users rating. For instance, researchers have proved that the customer's response to a restaurants average rating is affected by the total number of reviews that it has received and is unaffected by the size of the reviewers friend network [14]. Other work by (G. Ganu, N. Elhadad and A. Marian) shows that textual information (user reviews) give better results than numerical star ratings when measuring the quality of a business.

By contrast we work towards measuring an individual's closeness with his friends along with their opinions about a business and how these factors prompt him to visit the place. Social networks represent relationships between a user and his friends (people known to him). This study makes an effort towards understanding a non- traditional user graph like yelp, which has evolved over time with the primary intention of providing opinions.

We took a sample of the graph and analyzed the same social network that persists in Yelp. We found that instead of connected communities we found there were different clusters in the graph, which can be interpreted as people being close to their friends. We used Node XL to make the cluster representation. Exploring such a graph to understand closeness and social influence was a big challenge.

LIMITATIONS

As online communities are rapidly increasing in size, a lot of insights about a users behavior can be derived from the behavior and actions of the users ego network. Yelp was launched as a platform for people with the primary intention of giving opinions about businesses and rating them. Over the years it has managed to grow immensely as a social network allowing users to have friends and fans (similar to Twitter followers). On constructing the user graph from the provided dataset we observed that unlike popular social networks, the Yelp graph consisted of many users that had no friends and also many users with very small social communities (small ego network sizes) which is not very likely in other online networks like Facebook or Twitter. Every dataset has limitations and ours is no such exception. Since our research is restricted to users who have at least one friend, our dataset is biased, as it does not represent the entire population of Yelp dataset.

In spite of having business location data we did not have the user location data. Hence our research does not include user locations as a feature in understanding whether it affects a business. With the user location data (if provided) we could answer questions like whether a particular user visits businesses (restaurants) only close to his location or does he also travel away from home to try a new restaurant based on his friends' reviews. Also, the current implementation of Sentiment Analysis of text reviews is based on a simple unigram model that classifies a review as good, bad or great, unlike other deeper work that has happened in the space of sentiment analysis which focuses on Topic-based prediction of text reviews [9]. If such a model is used for opinion mining we believe that the accuracy of our model would increase.

FEATURE COLLECTION

I) Closeness within an Ego Network:

The PageRank algorithm developed by Google [Page Rank Paper Citation] computes the global importance of a Web Page by assigning a numeric score to it which influences the ranking of search results. As a step forward the notion of Personalized Page Rank aims at ranking results with respect to a given source. The same paradigm can be applied to a social network graph. Just like how twitter uses Twitter rank (extension of Page Rank algorithm)

which measures the topic similarity between users and the link structure [4], we work towards creating similar variables for the purpose of our research.

We calculate the Personalized Page Rank for a user in the Yelp Graph considering it as a measure of closeness to its own Ego network. The user data has *friend list* associated with it. A graph G (V, E) data structure consists of nodes or vertices 'V' and connections between these as links or edges 'E'. The user is a Vertex 'V' in the graph and it has an adjacency list which are its Edges 'E'. We make use of this to reconstruct the yelp graph that would help us calculate the closeness between users and his friends and their influence on him. We calculate closeness between user networks using the *Personalized Page Rank algorithm* [5].

$$P(n) = \alpha \left(\frac{1}{|G|}\right) + (1-\alpha) \sum_{m \in L(n)} \frac{P(m)}{C(m)}$$

Figure 3. Formulae for Personalized Page Rank [Wikipedia]

||G| is the total number of nodes or the size of the graph and alpha is the random jump factor. With other available programming models like MPI we would have to take care of details like synchronization and fault tolerance [6]. Hence we choose the Map Reduce framework, which is a widely used programming model for solving graph problems. We make use of the open source Hadoop implementation and calculate Personalized PageRank values for 32,000 users of the Yelp Graph. We study and learn the algorithm and the Map-Reduce Page Rank implementation explained by Jimmy Lin and Michael Schatz [6] [Prof. Lin's MR Book] and extends the same to calculate Personalized Page Ranks. In a typical implementation for Personalized PageRank the Page Rank mass lost in the dangling nodes (nodes without out links) are put back into source node. We exclude redistributing the floating mass contributions, as we have pruned our graph to exclude the dangling nodes meaning, we did not include the users who did not have a friend list at all.
So we extract a feature that we term 'Closeness' to understand how strength between the nodes can impact decision-making. One could say that 2% variance in closeness had a positive impact on business ratings. So closer a user is to his friend circle more are chances of him reviewing a business.

II) Social Influence:

Since we are working on understanding the influence of an ego on a user we had to come up with a feature that would help better describe influence.

Social influence feature was created to represent whether a users friends reviewed a business before him has impacted his decision to visit the restaurant or not. To calculate this feature, we compared the timestamp of the users review and reviews of his friends and for every friend that had given a review assigned a score 1 and if not then 0.

This way we averaged the total score to come up with social influence factor that shows a fair negative correlation with business rating and closeness and a positive correlation with Social sentiment feature i.e. people who have less friends are most likely to give rating to a restaurant and their social influence is going to affect their opinion about it. There is a direct relation between social influence and social sentiment. Increase in social influence causes significant increase in sentiment score and vice versa.

Algorithm: Social Influence calculation

1. Consider a user and a business
2. Check if the user has reviewed the business or not
    i. If reviewed
    ii. Check if the review date of friend
    iii. Else = 0

    b. If user- friend review date is before user review date
        i. If yes, then social influence = 1
        ii. Else = 0

    c. Calculate the average social influence
3. Add the column to the dataset

III) What Your Friends think:

Online review systems help users in decision-making. Businesses are often dependent on positive reviews from their customers to gain popularity attract more customers eventually increasing their customer base. The more positive reviews a business receives, higher are the chances of more people coming in. We use this analogy to measure the influence of friends within a network. A user would definitely visit a place and give positive reviews about it if the user has received positive reviews about the same restaurant from his/her friend circle. This eventually would matter to a business and give them some feedback on ways to improve their service.

In order to calculate this parameter we make use of a Unigram Sentiment Analysis model that we developed in R using text-mining package, where we try and understand whether a given review is positive of negative. Our approach was to take the user reviews, clean the text to remove the stop words, and then create a corpus of words. We considered a list of positive and negative words against which compared our text words.
On the basis of occurrence of words from the dictionary we create a numeric score for our review.

Algorithm: Sentiment Analysis on Yelp Reviews

1. Store positive words -> P {positive set}
2. Store negative words -> N {negative set}
3. **for** all reviews **do**
    a. Sentiment Score = 0
    b. Tokenize the text into words
    c. Clean words (remove punctuations, lower casing)
        i. **for** all Word 'w' **do**
        ii. **If** 'w' in P then SS += 1
        iii. **If** 'w' in N then SS += -1
        iv. Else Continue
    d. Save Sentiment Score

Using the above algorithm we calculated the sentiment behind the

text review of the user for a particular business. Majority of the businesses had received an average positive review and few had negative reviews. We did not consider the topic similarity in the reviews [4].

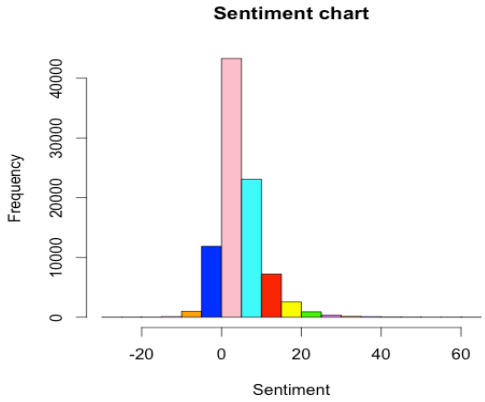

Figure 4. Analysis on sentiments derived on the text reviews show that very few businesses received negative reviews. Majority users had a very neutral opinion about a restaurant.

To test the accuracy of the sentiment score we calculated from our model, we gave a list 200 of reviews to some users and asked them to classify it in to 'positive', 'negative' or 'neutral' subsequently taking the majority opinion.

	Predicted Negative	Predicted Positive
Negative cases	TN 40	FP 44
Positive Cases	FN 33	TP 83

So the accuracy of our model is fairly good of around 61.5%. We used our sentiment model to come up with a new feature of Sentiment score. For this we considered a user, and created a list of his friends. We aggregated the sentiment score of each of his friends review for the business under consideration and added a new feature to out dataset 'Social Sentiment'.

Algorithm: Sentiment Analysis on Yelp Reviews

1. Consider a uid for a bid
2. sentiment_score = 0
3. For all *(uid, bid)* do
   i. FRIEND_SET <- set of friends
   ii. **for** all FRIEND_SET **do**
   iii. sentiment_score = sentiment_score+1
4. Add Sentiment score in a new column for the *(uid, bid)* pair

STATISTICAL MODELS

I) *'Given an ego node's sentiment score for a business, its personalized page rank (Closeness) and social influence measure can we predict if a user has given a review or not '*

The first question we try to answer in the research was whether a person reviewing a business depends on his friend's opinion and action. We worked on training our dataset to build a classification model, which when given a set of users, their closeness with their friends, and their friends opinions about a business, would predict whether a user has reviewed that business or not. This will help support our idea of a referral system, which possibly could be based on sentiments of users friends who are most likely to influence him.

We used decision trees to build the classifier. Decision trees can continue to grow infinitely by choosing splitting features, dividing it into smaller and smaller partitions for perfect classification or until algorithm runs out of features to classify further [15].

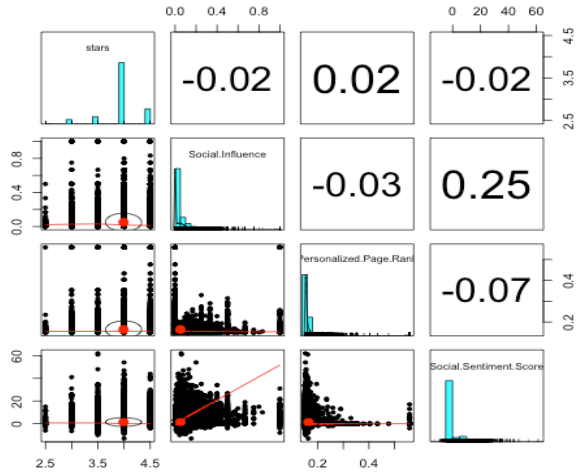

Figure 5. Above figure shows positive correlation between Personalized Page Rank (represents the closeness of a user within his friend list) and Negative Correlation between Social influence and Social Sentiment.

We got fairly good results with decision trees and Rules, after failed Naïve Bayes and SVM tests. The results derived were vague since the classifier gave no false positives or negatives. We believe this may be because our hypothesis of a user reviewing a business is based on his friends reviewing the same business did not match the machine learning models.

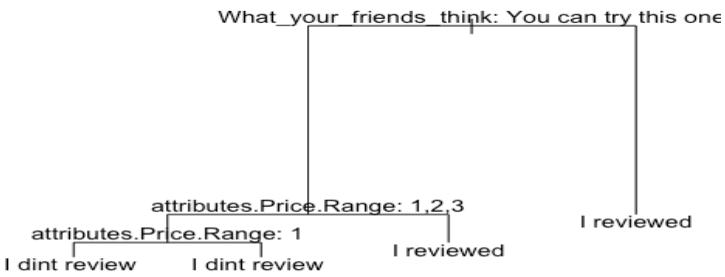

Figure 6 a) Classification Trees using Decision trees. To classify who has reviewed for a business depending on the social factors, we start at the root of the tree. We see that the users whose friends have given a decent review for a restaurant and depending on the price range tend to review for the similar business.

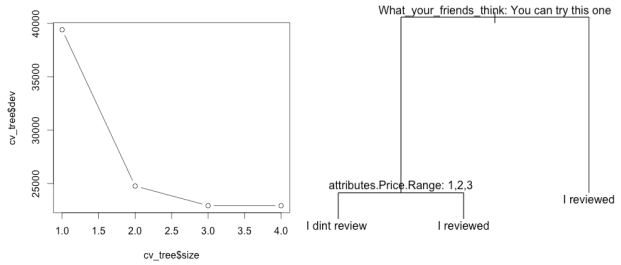

Figure 6 b) Tree model After Pruning

To support our observation with the decision tree model, we used the classification rules to represent knowledge in the form of logical if-else statements that assign class to unlabeled samples.

OneR: The One Rule algorithm is a simple classification technique that divides the data in to groups based on similar values of the feature [15]. Our results showed that ***69%*** of instances of our data were correctly classified based on single feature ***"What your friends think"***. Around 38% of instances were wrongly classified as businesses reviewed by the user.

Ripper: Since OneR uses only one feature for rule building we use an improved algorithm to consider other rules as well that might affect the classification process. The *JRip* algorithm on our dataset produced four rules.

Results showed, if your friends think that "You can try a restaurant" and it's a pricey business, users have not reviewed for it. Similarly, if it's a an average priced restaurant and friends have good things to say about it, then users have reviewed for those business as well. *This supports our idea that social influence in the sentiment form is a factor for a person to visit a restaurant*.

II) *'What factors influence a persons visit to a restaurant'*

The second part of our research was to analyze what factors decide whether a business gets a review or not and to understand if social influence is one of those factors.

Our correlation results showed that the better the closeness between a user and his friends in an ego network, it affects the ratings in a positive way. On the other hand the social sentiment and social influence factors show a negative correlation with the business ratings. Also 17% of variation in the price is reflected on the stars that the business gets. They show a negative correlation with each other.

To understand what factors allow businesses to have a good rating, we performed multiple linear regressions with business star rating as the dependent variable and combination of user and business variables as the independent variables. We use the *stepwise regression*, which includes the models choice of predictive variables, which is an automatic process. We use the backward elimination technique that involves starting with the candidate variables and checks for the improvement in the model performance by deleting the variables at successive iterations and comparing the model at each step. This continues till no further improvement is possible.

The formula returned at the end of the analysis shows the following features as the best predictor:

Business star rating <-
Closeness within and ego network + Number of fans a user has + Total number of review counts that the business has received already + the average reviews that the uses has given + Average rating that a user usually gives + How pricey the restaurant is

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.61718 -0.08598  0.02318  0.13309  0.75242

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.108e+00  1.281e-02 320.706  < 2e-16 ***
Social.Influence     -9.766e-04  1.164e-02  -0.084 0.933155
Social.Sentiment.Score -1.921e-04  3.241e-04  -0.593 0.553492
Personalized.Page.Rank  1.363e-01  2.891e-02   4.713 2.44e-06 ***
review_count         -5.053e-05  9.287e-07 -54.407  < 2e-16 ***
fans                  1.238e-04  3.736e-05   3.313 0.000923 ***
user_average_stars    3.160e-02  2.915e-03  10.843  < 2e-16 ***
attributes.Price.Range -7.045e-02  1.662e-03 -42.394  < 2e-16 ***
user_review_count     -8.637e-05  5.492e-06 -15.727  < 2e-16 ***
```

Figure 7a) Backward Elimination in Stepwise regression.

The maximum residual value of 0.752, shows that the model under predicted the ratings by 0.75 for at least one observation. The stars (***) indicate predictive power of the independent variables in the model. Three stars indicate significance level of 0, which signifies that the feature is extremely unlikely to be unrelated to dependent variable.

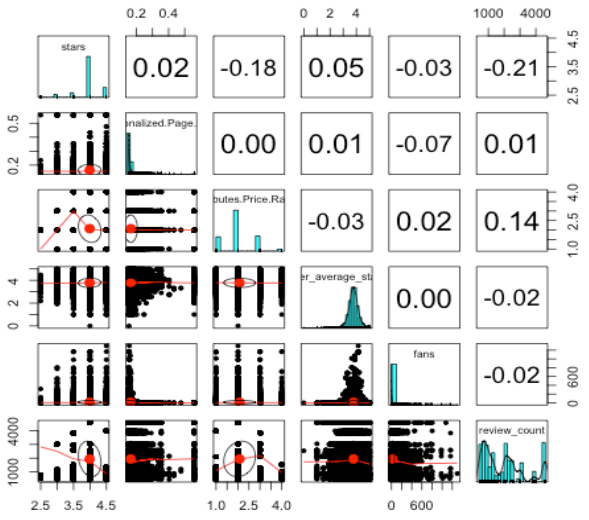

Figure 7b) Correlation between predicting variables.

We observed that there was a positive correlation between predicted and actual values that confirms our research question II.

APPLICATIONS

One of the potential applications of our study could be business recommender system for users. Based on a person's closeness with his online network the system could make

suggestions on what places (restaurants or other businesses) the person could visit. This would be more apt for location-based services where one looks for good nearby places to visit and the suggestions could be the ones recently visited by their friends or even places that their friends have good things to say about.

This concept can also be extended in the field of health care by providing recommendations to a user about hospitals, health centers and doctors. This would surely help users in case of health issues or emergency situations. The recommendations could be based on the price or even the sentiment of reviews given by an individuals network.

Extending our work further we also kept Yelp's business model in mind when working on our research problem. Yelp is majorly based on crowd-sourced reviews. They may probably make use of this idea to build a software application through which they could machine learn and predict which users are likely to visit a particular business. Using this application they could share coupons for different businesses with their users, who have a high certainty to visit a restaurant just by analyzing what their network has to say about it. The coupons might not pertain to restaurant businesses and can be of any other types too.

CONCLUSIONS

Our research results do support the idea of the previous research work of (Michael Luca) [14] which shows that the number of reviews that a business receives impacts the rating that it receives. On the similar lines, we can say that a user is most likely to review a business if his friendship bond is stronger within his network. Mostly restaurants, which are very pricey, tend to receive fewer reviews. This may help us conclude that a user is most likely to visit a restaurant if his friends have good things to say about the place and its not very costly. Also we observed that, if a user's network has less social influence and social sentiment score then the user rating is comparatively higher. One may say that influence does impact what rating a user would give to a business. Also, the classifier that we have implemented could form the basis of a referral system. Based on what a user's network has endorsed him for, the recommendation system could send job alerts to the user.

ACKNOWLEDGEMENTS

We would like to thank Prof. Vanessa Friaz Martinez (Assistant Prof. at iSchool, Univ. of Maryland, College Park) for encouraging us to research in this space and constantly mentoring us throughout the project. We are also thankful to Prof. Jimmy Lin (Prof. and Associate Dean for Research at iSchool, Univ. of Maryland, College Park) for his useful guidance.

References

`

[1] Muchnik, L., Aral, S., & Taylor, S. J. (2013, August 9). Social Influence Bias: A Randomized Experiment. Retrieved from http://web.natur.cuni.cz/~houdek3/papers/Muchnik%20et%20al%202013.pdf

[2] Yelp Dataset Challenge. (n.d.). Retrieved from http://www.yelp.com/dataset_challenge

[3] Linshi, J. (n.d.). Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach. Retrieved from http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf

[4] Weng, J., Lim, E. P., Jiang, J., & Hi, Q. (2010). Twitterrank: Finding Topic-Sensitive Influential Twitterers. Retrieved from http://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1503&context=sis_research

[5] Lin, J., & Dyer, C. (2009). Data Intensive Text Processing with MapReduce. Synthesis Lectures on Human Language Technologies. doi:10.2200/S00274ED1V01Y201006HLT007

[6] Lin, J., & Schatz, M. (n.d.). Design Patterns for Efficient Graph Algorithms in MapReduce. Retrieved from http://lintool.github.io/bigdata-infrastructure2015s/content/Lin_Schatz_MLG2010.pdf

[7] Jimmy Lin and Chris Dyer. Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers, 2010.

[8] (Everett et al., 2005) Everett, M., & Borgatti, S. P. Ego network betweenness. Social Networks. Doi: 10.1016/j.socnet.2004.11.007

[9] (Ganu et al., 2009) Ganu, G., Elhadad, N., & Marian, A. Beyond the Stars: Improving Rating Predictions using Review Text Content

[10] (Longke et al.,n.d) Longke Hu., Aixin Sun., & Yong Liu. Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction. School of Computer Engineering, Nanyang Technological University, Singapore.

[11] (Lu et al., 2011) Lu, B., Ott, M., Cardie, C., & Tsou, B. K.. Multi-aspect Sentiment Analysis with Topic Models. Doi: 10.1109/ICDMW.2011.125

[12] (Moontae et al., 2010) Moontae Lee., & Patrick Grafe. Multiclass sentiment analysis with restaurant reviews.

[13] (Ying et al., 2012) Ying, J. J., Lu, E. H., Kuo, W., & Tseng, V. S. Urban point-of-interest recommendation by mining user check-in behaviors. Doi: 10.1145/2346496.2346507

[14] Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. Retrieved from http://www.hbs.edu/faculty/Publication%20Files/12-016_0464f20e-35b2-492e-a328-fb14a325f718.pdf

[15] Lantz, B. (2013). Machine Learning with R. Retrieved from October 2013

[16] Despalatovic, L., Vojkovic, T., & Vukicevic, D. (n.d.). Community structure in networks: Girvan-Newman algorithm improvement. Retrieved from http://docs.mipro-proceedings.com/cts/CTS_09_2540.pdf