The Yelp Elite: Do they matter, and who are they?

# Problem and hypothesis



The Yelp Elite is a distinction made to reward members of the Yelp community who have developed a strong track record on the site. However, when it comes to understanding their actual power, we ask the following:

How different are the Yelp Elite from the general community and do their reviews have more impact on a business's success?

Our hypothesis is that the Yelp Elite are a more discriminating bunch of Yelp members and that their reviews will be a better “first stop” to look at when determining where to visit.

# The data

While there were several datasets available for use, our focus was on three particular datasets:

(1) Dataset of user profiles that included the user ids as well as a number of important variables on that user’s site behavior, as well as the years when the member was considered "Elite". Our target variable will be drawn from this dataset. The following variables are included in the dataset:

1) count of business categories visited (separate data set)

2) average\_stars

3) review\_counts

4) votes

5) compliments

6) fans

(2) Dataset of individual reviews that includes the userid and date of review so that it can be matched with the userid table

(3) Dataset of business information that includes the category of the business as well as the star rating of the business overall

Each of these datasets was publicly available through Yelp’s Data Challenge. http://www.yelp.com/dataset\_challenge

# Pre-processing steps

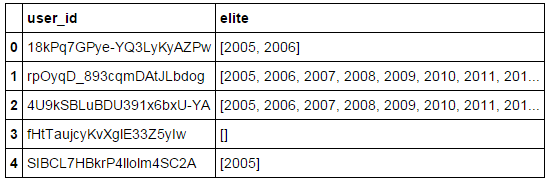
**Initial File Type Conversion**

Before any data manipulations could occur to run models, we had to convert JSON formatted Yelp! files to a csv format to be workable for our purposes. We had to convert the JSON strings that were extracted to a flat Python dictionary which could then be passed into Pandas.

**Pre-Processing the Users Table**

There were a number of steps that needed to be taken, in particular, to identify reviews that actually came from Elite members.

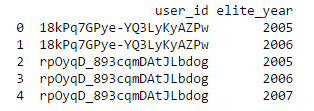
The first step in this process was sub-setting the user table to user id and the elite variable, which unfortunately, was stored as a string of a years that needed to be converted to an array that could be then further split.



*Figure 1: Elite column is stored as string*

In order to assure that in cases where the array was blank that there was indeed a blank, the length was then taken of the generated array to help gauge the number of years that a user was part of the elite.

Finally, we were able to see the unique user, year combinations following this processing as seen in Figure 2.

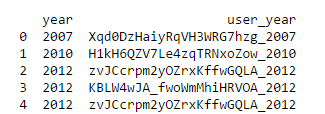


*Figure 2: Unique user/year combinations*

**Pre-Processing the Reviews Table**

Within the reviews table, the first order of business was to coerce the date field into a pandas version of dates that can be used for additional analyses.

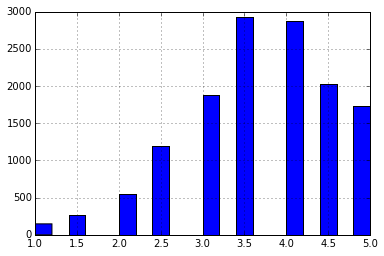
The next step was to concatenate the user id with the year to give us an indexing column that we can use to merge with the user table -- giving us the ability to flag particular reviews as coming from an elite source.



*Figure 3: User, Year combination in Reviews Data File*

# data exploration

Approximately 10% of businesses in the Las Vegas dataset we worked with had star ratings of 5 on Yelp's scale (that has spacing of 0.5 increments). There were a total of 13,601 unique businesses used in the analysis for Las Vegas. The data is right skewed with very few businesses having star ratings of 2 and below. The average rating is 3.66.

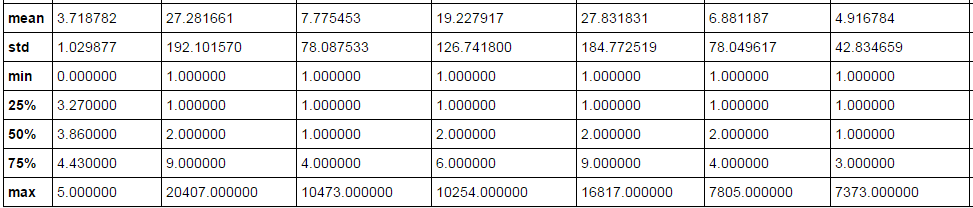


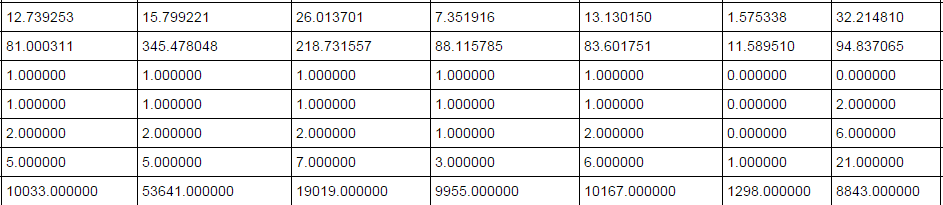
*Figure 4: # of businesses with star ratings of...*

In terms of the users that are being modeled in the second analysis, their average stars are relatively in line with the overall business ratings, being skewed to the right with a median of 3.86, as seen in the descriptive stats from Figure 5.

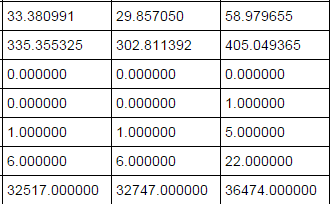
A couple of other noteworthy points from the descriptive statistics that will be important considerations when looking at the modeling algorithms:

* Regardless of elite status, the two most common compliments received by users are simple “Thank You” messages with no connotation included and cool and hot. The compliment handed out the least is for the reviewer to “write more”
* Not surprisingly, the compliment with the most variance is on photos – likely driven by the fact that not all reviewers include photos to be commented on in the first place. Of interest, however, is that although the generic “Thank You” would appear to be handed out most often, it also has the most variance, indicating a strong skew in the data (and probably an outlier).



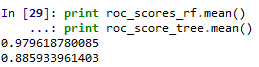


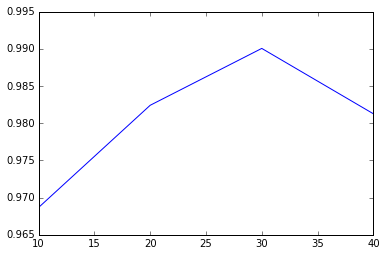


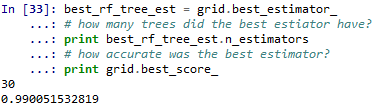


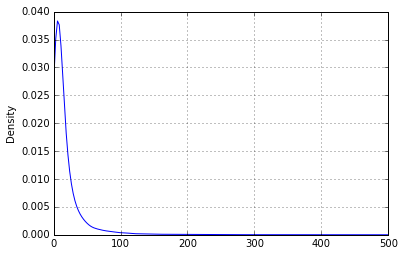
# how you chose which features to use in analysis

# details of modeling process

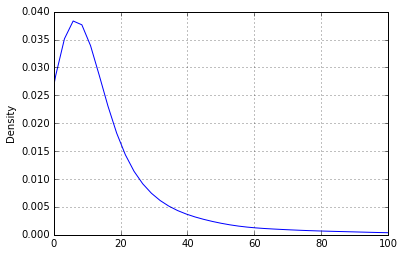








Within elites, a shocking number only have fans



# Challenges, successes, and conclusion

To take advantage of this template’s design, use the Styles gallery on the Home tab. You can format your headings by using heading styles, or highlight important text using other styles, like Emphasis and Intense Quote. These styles come in formatted to look great and work together to help communicate your ideas.

Go ahead and get started.