All Stars and Useful Reviews: A Clustering Exercise

Data Mining Info 290T: Homework 2

Teammates: Lucia, Shannon, Carina, Michelle

1) RESEARCH QUESTION

Our **research question:** Do "all star" reviewers give “useful” reviews?

By the feature’s definition, to be an “all star” reviewer, a reviewer must have left at least one review in each ‘star’ category (1 - 5). We want to understand if whether a reviewer is an all star or not is correlated with the usefulness of a reviewer’s reviews.

2) DATASET

The dataset we are using (yelp_reviewers) is a transformation of an original dataset (yelp_reviews). The original dataset was one scraped from Yelp that detailed businesses, reviewers, type of review (cool, funny, useful), and stars. The transformed dataset had a number of extra features developed by students in the class mostly revolving around different ways to interpret cool/funny/useful and number of stars.

This transformation happened through multiple steps, the first and most important of which was to group the dataset by “user_id” so that we would have one row of data for each user, telling us how many useful/funny/cool reviews they wrote on average. We calculated other variables, such as the log or the percentage of useful/funny/cool reviews, the year each user wrote the most reviews, and the average length of the reviews’ text per reviewer.

We selected q16l and q6 as our features. q16l, “all_star”, is defined as a binary feature that takes a value of 1 if the reviewer has left at least one review for each category of stars (i.e., at least 1x “1-star” review, …, 1x “5-star” review). q6 was “usefulness.”

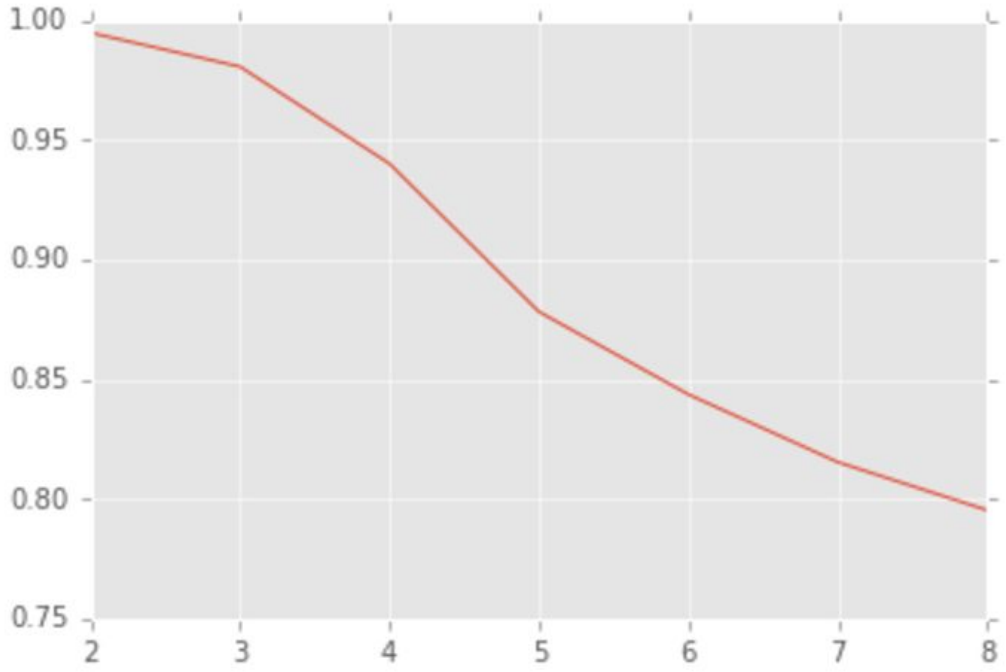
3) METHODS

To determine whether there is a correlation between all-star and useful reviews, we clustered the dataset using k-means.

We scored our clusters by their silhouette score, “a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).” While we also considered the inertia method, which measures inner-cluster variance, we felt the silhouette score would be a more relevant metric for our research question. A higher silhouette score relates to a model with better defined clusters - taking into account difference between and within clusters (as opposed to just within).

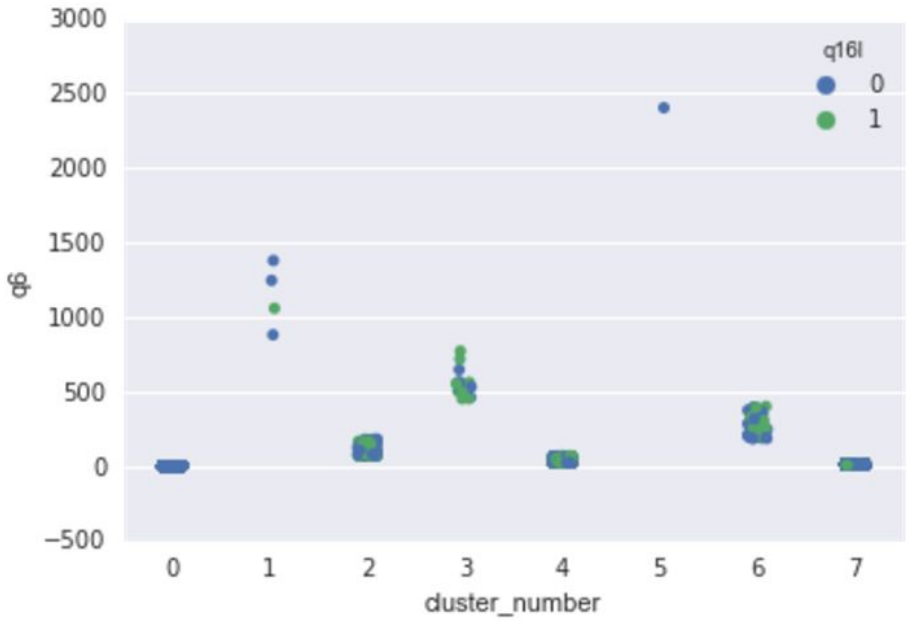
4) RESULTS

When we clustered and scored our sliced dataset, we noticed that our silhouette scores were very close to one (k = 2, 0.99 to k=8, 0.79).



When we investigated the nature of the clusters when k=2 and when k=8, we noticed that there was always one cluster that contained the majority of the reviews. When k=8, the largest cluster had 149,511 entries out of 171,639. We also looked at how the ‘all-star’ data was grouped within the clusters, and we saw that they mostly fell within 2 or 3 groups

depending on the k value. This leads us to the conclusion that there is **no significant correlation** between the fact that a user provides reviews in all 5 'star' categories (= an all-star reviewer) and being considered useful by her peers on the platform.



In the chart above, blue dots represent non all-star reviewers, and green dots represent all star reviewers. As can be inferred, the green dots, all star reviewers, are spread across all clusters, further indicating the lack of correlation between the two selected features.