INTRODUCTION

In 2005, Yelp introduced an "Elite" users feature, an exclusive perk for active and engaging users on the site. The feature includes access to Elite-only food events, networking, and special recognition around the site. For these reasons (but especially for the free food!), users have long looked for ways to become an Elite member. Unfortunately, Yelp is extremely secretive about their "Elite Squad;" the council that selects these members remains anonymous, and Yelp doesn't even release the number of users currently in the Elite Squad. Not only that, but there are no specific guidelines for how to obtain this coveted status, nor have any Elite users found a common set of hard and fast rules to maintain their statuses on a yearly basis. For this reason, the present report aims to assess the effect of different forms of engagement on Yelp on the likelihood of a member having Elite status using two supervised learning methods: logistic regression and *k*-NN, on the Yelp Open Dataset.

METHODS

Data

Data were collected from the 10GB, open-source Yelp Open Dataset, an all-purpose dataset for machine learning and Yelp-run data challenges, updated yearly. The present study utilizes the review and users subsets, contain information from over 6.6 million reviews from 1.6 million Yelp users.

Natural Language Processing

First, using MapReduce, we ran an NLP (Natural Language Processing) model to find the average readability score across each user's reviews in the reviews dataset. We employed the TextStat package, specifically the Flesch reading ease function, which takes in a chunk of text and returns a metric describing the readability and complexity of that text. The Flesch reading ease score is calculated with two ratios that are subtracted from a constant (206.835): total words/total sentences and total syllables/total words. Based on these ratios, complex and/or compound complex sentences along with long words depress the score whereas simple sentences with simpler, monosyllabic words would generate higher scores. Review scores max out at

121.22 (most easily read) while the lower bound is not stated. For reference, a typical article from the Harvard Law Review scores in the low 30's, while the Harry Potter series averages 72.83. This score was used as a proxy for the complexity and utility of a specific user's reviews. We predicted that a review from an Elite user should be complex, due to its descriptive and informative language, but also clear and comprehensible enough to still be useful to other users. Average readability scores for each user (across all of their reviews) were then joined with the users dataset.

Logistic Regression

We trained a logistic regression model in R 3.5.3 with a random sample of 150,000 users, about 10% of the overall dataset. Our logistic regression model used the scores that were obtained from the Flesch Reading Ease results, the number of reviews that the user posted, number of fans of the user, number of average stars that the user gives, how many years the user has been a Yelp member, whether the user has ever been elite, and lastly, total number of compliments the user has given and received. The outcome of interest was a binary variable indicating if a given user was ever Elite (Elite status changes on a yearly basis). We used variance inflation factor tests, as well as deviance tests (logistic regression equivalents of F-tests) to assess the importance and weight of different variables on the model.

Primarily, we aimed to assess the accuracy of this model; that is, calculating the percentage of elite or non-elite users accurately predicted as such by the training model, using MapReduce. For each user, the mapper calculated the predicted value from the logistic regression model and assigned them either a zero (if their predicted probability of ever being elite was less than 0.5), or a one (if their predicted probability was greater than or equal to 0.5). Keys were two integer strings; the first integer was the actual ever elite status of the user, the second was the predicted value. After combining and reducing, we generated four tuples of 0,1 pairs, with the frequency of each tuple across all 1.6 million users in the dataset.

***k*-NN Analysis**

As a secondary analysis, we hand-coded a *k*-nearest neighbors (*k*-NN, for short) algorithm to predict whether or not a user was elite. The *k*-NN method of accomplishing this entailed searching through every other user in the dataset to find the *k* "closest" or most similar other users. After finding these *k* "nearest neighbors" to our single user, we then determined if the single user was elite or not by majority vote among its nearest neighbors. We used MapReduce on Google Cloud clusters to accomplish this. We started by creating a formula for calculating the similarity between two users:

$$\textbf{Similarity Score} = 0.35\left(1 - \frac{\sqrt{\textbf{review\_count}}}{116}\right) + 0.125\left(1 - \frac{\sqrt[3]{\textbf{friends}}}{53.65}\right) + 0.125\left(1 - \frac{\sqrt[4]{\textbf{fans}}}{24.5}\right) + 0.125\left(1 - \frac{\textbf{avg stars}}{4}\right) + 0.125\left(1 - \frac{\textbf{score}}{120.78}\right) + 0.07\left(1 - \frac{\sqrt[3]{\textbf{total compliments given}}}{65.4}\right) + 0.05\left(1 - \frac{\textbf{years yelping}}{14.1}\right) + 0.03\left(1 - \frac{\sqrt[3]{\textbf{total compliments received}}}{65.2}\right)$$

**All variables (bolded) are the difference between the single user and the prospective neighbor/other user, among that particular variable.*

The weights for each component of the formula were determined by the statistical significance each variable had in our logistic regression: variables that were more statistically significant were more powerful predictors of eliteness, thus we figured they would operate similarly in the *k*-NN algorithm. We obtained the similarity score between a single user and all other users who were its prospective neighbors using a mapper function. For a single user, once we obtained the similarity scores for its comparisons with all other users, we yielded key-value pairs of ***key:*** user, ***value:*** (other_user, similarity score). The reducer took in a single user and a generator object containing each of the key-value pairs corresponding to that particular user from the mapper. Using a PriorityQueue of max length = *k*, we obtained the *k* other users with the largest similarity scores for the single user. Using these nearest neighbors, we then predicted if the single user was elite or not by majority vote of the nearest neighbors. The reducer then yielded one of the four strings: "00", "01", "10", or "11". The first integer corresponded to the actual elite status of the single user; the second integer corresponded to the predicted elite status of the single user based off of the *k*-NN. These strings were piped to a CSV, and by counting the occurence of each string in a second, more trivial mapreduce scheme, we assessed the accuracy

of our algorithm. The algorithm could not be run without Google Cloud; runtimes for a <1% subset of the data took nearly 40 minutes.

RESULTS

The logistic regression model is summarized in Table 1. Though we initially assessed the validity of splitting up each type of compliment received/given into its own individual term, this model had poor predictive power, so we ultimately decided to combine these terms into total counts for each. All terms are significantly important to the model, save for the total number of compliments a user received. However, after assessing the term using a deviance test, we asserted this term was still important to include in the model.

term	estimate	std.error	statistic	p.value
(Intercept)	-4.1184859	0.1688817	-24.386805	0.0000000
score	-0.0185642	0.0010639	-17.449798	0.0000000
review_count	0.0236431	0.0004504	52.497742	0.0000000
fans	0.0000776	0.0000044	17.719083	0.0000000
average_stars	0.4126102	0.0264123	15.621892	0.0000000
years_yelping	-0.0171222	0.0076540	-2.237032	0.0252842
total_compliments_given	-0.0019019	0.0003567	-5.332629	0.0000001
total_compliments_received	0.0005216	0.0003997	1.304855	0.1919421
friends	0.0056643	0.0004384	12.919983	0.0000000

Table 1: Results of Logistic Regression model

Each estimate represents the change in the log odds of a user ever being Elite for a one unit increase in that term. For example, for every one unit increase in the Flesch reading ease score, the model predicted a -0.0186 unit decrease in the log odds of a user ever being Elite, holding all other terms constant ($p < 0.01$). For score, this directionality is expected, as higher scores indicate lower reading complexity. The average number of stars across the user's reviews was also highly predictive of having been elite; a 1-star increase in the average number of stars given is expected to increase the log odds of ever being Elite by 0.4126 on average, holding all else constant ($p < 0.05$). Besides these indicators, we also included information regarding the total number of reviews a user had written, the number of fans and friends they had, how long they had been on Yelp, as well as the total number of compliments they had received on their profile, and the total number of compliments they had given on other users' reviews.

After running MapReduce on the testing dataset only, we found the following four pair frequencies, summarized in Table 2.

Actual Status	Predicted Status	Frequency
1	1	27006
0	0	1430572
1	0	26819
0	1	6470
TOTAL:		1490867

Table 2: Results of Logistic Regression predictivity

Overall, the model was accurate in predicting a user's status; our model was able to predict statuses correctly 97.8% of the time. Overall, this leads to a false positive rate (users were not Elite but predicted as such) of just 0.43%, and a false negative rate (users are Elite but predicted as not Elite) of just 1.8%. Despite these promising findings, our true predictive power is masked by the overwhelming number of non-elite users in the dataset. Despite a 97.8% accuracy rate overall, the accuracy for accurately predicting Elite status is essentially null- the model correctly predicted just 50.2% of ever Elite users as such.

For k-NN, we noticed that there was not almost no occurrence of false negatives. After looking at the distribution of the dataset, we understood these results to be reflective of the fact that the data points of elite users tend to be very distinct from those of non-elite users. Therefore, the lack of false negatives confirms our hypothesis that elite clusters would tightly cluster together, which would be reflected in a model utilizing k-NN/a priority queue.

Actual Status	Predicted Status	Size of K	Percentage of Observations
1	1	**5**/10/15	**3.5%**/3.5%/3.14%
0	1	**5**/10/15	.**28%**/1.6%/.36%
0	0	**5**/10/15	**96.22%**/94.9%96.5%

After running multiple tests and computing averages, we found that taking k = 5 yielded the best results, in which best results is defined as minimal percentage of false positives.

DISCUSSION

We decided to use Flesch Reading Ease Test for "scoring" the complexity of the reviews because it is standardized for all reviews. The Flesch Reading Ease Test also has distinct cutoffs for each score interval in terms of the complexity of the sentences. These distinctions between complexity levels helped to ground our idea of what makes a review "complex." Further, we had to account for some cases in which a review was not in English. In these cases, the Flesch Reading Ease Test analysis gave us a negative number, which we decided to clear out from our analysis. Therefore, our analysis only included users who wrote reviews in English.

Despite the overall predictive power of the logistic regression model, it was surprisingly not at all useful in accurately predicting ever Elite users as such. We tested our model for robustness in R, and each term was found to be extremely statistically significant in predicting the log odds of Eliteness. Nonetheless, there still remain a number of reasons the model may not have accurately predicted Elite users. For one, Elite users made up just 3.7% of the overall sample in this dataset, with only 6574 Elite users making up the training dataset. The fact that these users made up such a small proportion of the overall dataset may have made it difficult to extrapolate predictions about other users, especially given the fact that many of the covariates in our model could have been broadly overlapping between Elite and non-Elite users.

There may have also been omitted variable bias present here- it is entirely possible that there are some factors Yelp uses under the hood in their Elite decisions. Or, while these factors are important in determining Elite status, Yelp may use an unknown weighting scheme of their own to aid in their decision. In this case, our model would not have accounted for this, as there is no weighting associated with each variable (besides the magnitude of each coefficient, though this is not entirely one-to-one with a weighting scheme). We attempted to account for this in the KNN analysis, by weighting variables in our distance algorithm, with much success.

Through *k*-NN analysis and *k* selection, we found that *k*-NN had more predictive power for specifically elite users in comparison to logistic regression.

It is also important to note that the users need to be nominated, either by themselves or by someone else, in order to be considered to become elite. It might be the case that the "elite" status is not well-known enough to ensure all potentially elite users were actually elite. We may have had users that were good candidates to become Elite, but their lack of knowledge about the program would have affected their ability to actually become Elite. This can be a strong explanatory factor as to why there are so few Elite users in our dataset, as well as why it was difficult to predict them via quantitative methods.

CONCLUSION

Yelp users have long hypothesized a number of ways to increase their odds of being an Elite user, including leaving well-thought out reviews, engaging in other users' content, and amassing large quantities of fans and friends. Despite their best efforts, however, our logistic regression model found very little predictive power in these factors on predicting Elite status; our *k*-NN algorithm, having no false negatives, seemed to be better at predicting the "eliteness" of a user, given that the user was elite. Though there was still a noticeable amount of false positives, which could possibly indicate that some non-elite users have elite characteristics but simply do not receive the distinction, this is less problematic considering that the focus of our study is elite rather than non-elite users. Yelp has always been secretive about the selection process for attaining this coveted status, and it seems our analysis only aids that there are a number of secret, unpredictable, and unobserved factors that affect a user's chances of ever being Elite. There is certainly more analysis to be done, but perhaps Yelp is also just very good at concealing the true path to becoming Elite.

References

Harris, Jenn. “For Some Yelp Reviewers, It Pays to Be Elite.” *Los Angeles Times*, Los Angeles Times, 1 May 2015, www.latimes.com/food/la-fo-0502-yelp-20150502-story.html.

“How to Write Plain English.” *Guide to Academic Writing Article - Management - University of Canterbury - New Zealand*, web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml.

Olivia. “How to Become Yelp Elite: The Perks and Benefits.” *Birds of a FIRE*, 5 Dec. 2018, www.birdsofafire.com/yelp-elite-perks-fancy-free-food-cool-experiences/.

“What Is Yelp's Elite Squad?” *Yelp*, www.yelp-support.com/article/What-is-Yelps-Elite-Squad?l=en_US.