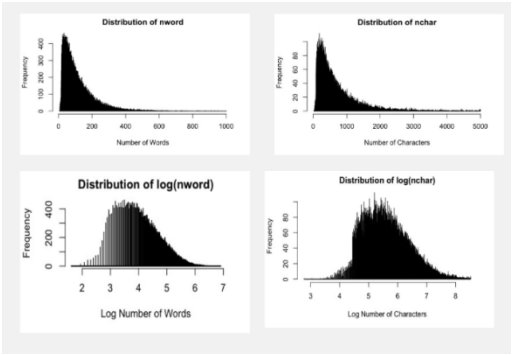
Yelp Predicting Project Write-Up
Group 21: Jing Duan and Yanxu Guo

Yelp is commonly used by users to see the numerous ratings and reviews that are left by the previous users for the local businesses and restaurants. The purpose of this project is to generate a predictive model to predict the Yelp rating based on the predictors and sentiment that are extracted from the review texts in the dataset.

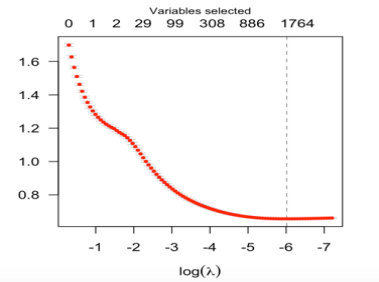
Our proposed model is a lasso regression model with 1597 predictors in total. All predictors were selected from the training dataset. According to the score in Kaggle, the root-mean-squared error of our prediction from our proposed MLR model is 0.797, which means the average error between our predicted star rating and the true star rating is around 0.8 star.

From the suggestion of the yelp html, we thought adding predictors is the main process of making the model more accurate. We used the tm package in R to help generate new words. It achieved our final predictors through 4 processes. It first changed all the words in the document to lowercase. Secondly, it removed all the punctuations and stop words, such as “i”, “me”, “a”, “and”, etc. Then it stemmed all the words, making “loved” and “love” both to “love”. Lastly, it removed all the sparse terms. We then used freq_terms function in the qdap package to choose the top 2100 high-frequency terms. After generating words predictors, we used tokens_select(wordnot, pattern = phrase('not_*')) to generate 250 phrases with the word “not”. We then manually chose 8 high-frequency phrases. With 148 original words that are not overlapping with our generated words, we had 2506 predictors in total.

Our first model was the multiple linear regression. We had 2 main changes to the dataset. 1.) Using log transformation to nchar and nword. 2) Using binary to indicate the appearance of the words for the 200-2506 terms and using its original frequencies for the original words.



We used log transformation because from the graph of nchar and nword, the graph was highly skewed to the right. By doing the log transformation, the distribution was becoming more normal, and the prediction could be more accurate. We used binary because it is more important to count the appearance of the words than how many times it appears. For instance, one of the words that is highly appeared is food. No matter how many times it appears, it conveys the same meaning and has nothing to do with the prediction.



Our model was improved by using lasso regression with cv of 5 folds. By choosing the variables that have slopes other than zero at the smallest lambda, the lasso model ended up having 1597 variables. Compared with our first model based on MLR, the variables that are chosen by lasso are more significant with smaller p values. Our MSE improved by 0.01 after using the lasso regression.

Our final model based on lasso regression has an R-squared value of 0.6447, which means our model can explain 64.47% of the variation in the data around the mean. The residual standard error is also low at

0.789, accounting for the accuracy of the model. The p-value of the model is relatively small, meaning most predictors are significant in the prediction process.

```
Residual standard error: 0.7885 on 53745 degrees of freedom
Multiple R-squared:  0.6447,    Adjusted R-squared:  0.6342
F-statistic: 61.11 on 1596 and 53745 DF,  p-value: < 2.2e-16
```

After we got our predictions of testing data based on our model, we refined the predictions by using the Law of Large Numbers. We calculated the distribution of our predictions by rounding the predictions to whole numbers, as well as rounding up any predictions less than 1 to 1 and rounding down any predictions greater than 5 to 5. As shown in the table below, you can observe that compared to the distribution of the training dataset, our model tends to overpredict 3 star and 4 star while underpredicting 1 star and 5 star.

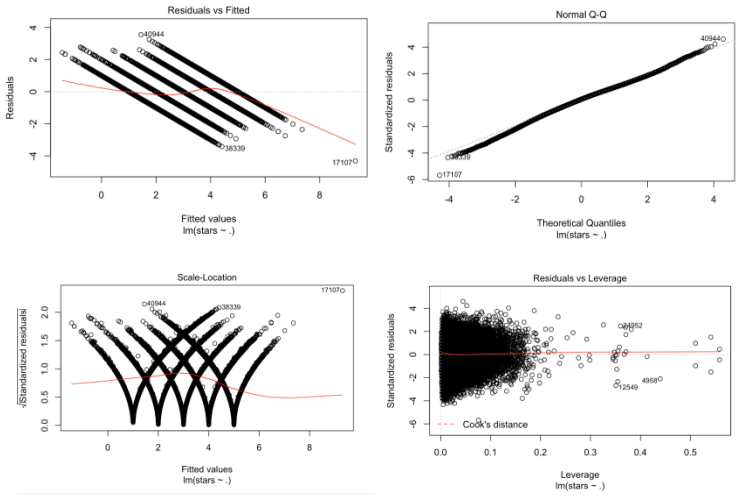
Dataset | Star	1	2	3	4	5
Yelp_Train	0.0968	0.1026	0.1458	0.3002	0.3546
Predictions	0.0346	0.1013	0.1991	0.4448	0.2202

According to the Law of Large Numbers, the average results obtained from a large number of trials should be close to the population distribution. Both of our training and testing datasets contain over 30,000 samples, hence the sample sizes are large enough for the datasets to have similar star rating distributions. Since the predictions are biased toward 3 star and 4 star, we decided to adjust the thresholds for each star level so that the distribution of our predictions can be similar to the distribution of the training dataset. To adjust for such a situation, we first created different thresholds for different star levels based on the training data quantile for each star, then did some trials to test different thresholds' distribution. The final thresholds we came up with were as follows: for predictions within the interval of 1.5 and 2.25, we adjusted them to be 1.49; for predictions within the interval of 2.5 and 2.89, we adjusted them to be 2.49; for predictions within the interval of 4.0 and 4.29, we adjusted them to be 4.51; and for all other predictions, we kept them as they were. Finally, we were able to adjust the distribution of our predictions to be similar to the training dataset as shown below. And this adjustment reduced our RMSE on the Kaggle by about 0.008. However, due to the adjustments we used, we have a stepwise discontinuous predicted value rather than a continuous predicted value.

Dataset | Star	1	2	3	4	5
Yelp_Train	0.0968	0.1026	0.1458	0.3002	0.3546
Refined Predictions	0.1026	0.0966	0.1356	0.2981	0.3669

The plots for checking the assumptions of our lasso regression model are shown on the right. Based on the residual plot, it seems like that the linearity assumption is violated, but there are reasons for such behavior. The five parallel lines result from the fact that there are only five possible values for our response variable: stars, and the pattern of the slopes are due to the conditional probability. According to an article about the linear pattern of the residual plot (Searle 1987, 1), the parallel lines are due to the plot is a plot of conditional variables, which means it is a plot of [(y-y_hat)|y=c] against (y_hat|y=c), or the same as c - (y_hat|y=c) against (y_hat|y=c). Hence it is a straight line, one for each c (each possible outcome of response

variable) with a slope of -1. Thus it should not be surprising to see five parallel lines with the slope of -1 in the residual plot. Based on the other three plots, our assumptions for the model are met. From the QQ plot we can see all the points are roughly a straight line which means the normality assumption is not violated, and our sample distribution is close to Normal distribution. From the Scale-Location plot, we can see there is no clear trend, hence the homoscedasticity assumption is met. Based on the last plot, there is no outstanding high Cook's distance point in our model.



There are three main strengths of our proposed model: 1) The model we produced is based on high-frequency predictors that were effectively selected. 2) We have 1597 predictors in the final model, which gives us an R-squared value of 0.6447. This is a relatively high R-squared value for a model to predict a dataset of over 50,000 samples. 3) The model is hardly influenced by a single character or word.

There are three main limitations to our model and process: 1) It is hard to find the most frequently used phrases through data based on our process. 2) We chose to use the stem of the words during the predictor selection process, but words with the same stem may not mean the same, like both correct and corrupt have "corr" as the stem. 3) We did not cancel the collinearity problem in our model and we did not add interaction terms as predictors.

Here are some suggestions for future studies: 1) Use correlation as the first step to narrow down the predictor selection rather than solely based on frequency. A high correlation with the response variable is an effective way to choose predictors. 2) The appearance of the words is more crucial than the frequencies of the words. A method we have tried to solve the issue is the TF-IDF, but because of the time limit and we did not figure out the idea behind it, the method was not used. Also, change all the frequency of predictors to binary. Even if we have changed the frequency to binary, it is awkward that we only changed the 200-2506. if having more time on the last day of the project, we would change all the words to binary. 3) Maybe try to group words with similar meaning together and form interaction terms based on them to reduce collinearity. The interaction has been thought of as a factor that influences the model. However, we did not come up with a solution to reduce the interaction effects in time. 4) Stepwise regression can be applied to select predictors, but have to do a small amount like 100 predictors at a time, otherwise, R Studio will freeze. We have tried to use stepwise regression to reduce the number of predictors while we had over 1500 predictors selected, but the R ran for the whole night and became frozen with no result. Hence maybe running a small group at a time would work.

In conclusion, the Top 5 methods of improving the MSE scores are 1.)Adding word predictors 2.)Adding predictors with not. 3.)Binary transformation 4.)Log transformation 5.)Changing the score above or below the stars boundary to the boundary scores("1", "5").

We believe our model is a robust model for predicting stars in Yelp based on the review texts using 1597 predictors included in our model. However, there are some limitations to our

model. Overall, our model is able to predict about 65% of variability of yelp star rating in the given dataset.

Jing Duan and Yanxu Guo are the members of team 21. The work has been evenly done by each group member. For the code part, we write together by both thinking about the projects and adding on our former code if we have new ideas. Yanxu did the most part of data collection and model refining, while jing did most part on analysing the model through graphs and finding the characteristics of each model's predictor. Based on this, the presentation is also assigned equally such that Yanxu talks about predictor selection and model refining and Jing discusses inference and diagnosis. In the end, we think together of the advantages and disadvantages of our project and the future improvement.

References:

Searle, Shayle R. "Parallel Lines in Residual Plots". *Ecommons.cornell.edu.* October, 1987. https://ecommons.cornell.edu/bitstream/handle/1813/33060/BU-945-M.pdf;sequence=1

Machine Learning with test data using r. *Pluralsight*. 2019. https://www.pluralsight.com/guides/machine-learning-text-data-using-r/

Jing Duan:5
Yanxu Guo:5