Analysis of Yelp Rating of Reviews

Final Report
Yicheng Yuan, Lynette Gao, Emily Xu

Problem Statement and Background

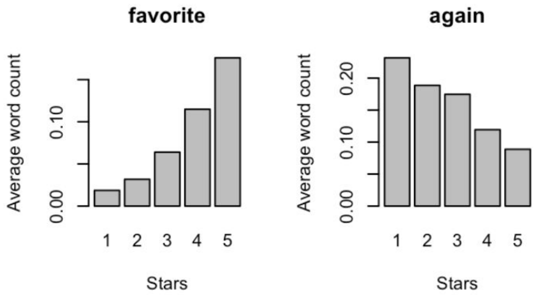
Yelp is a convenient and popular platform for users to submit reviews for diverse types of businesses according to their feeling of experiences. These reviews are considerably important to the businesses because they may influence decisions of potential visitors who use Yelp as well. Since 2016, there have been more than 121 millions reviews on Yelp and they are created into a data set called “Yelp Dataset Challenge”. For this project, we construct a sentiment analysis on reviews of restaurants, which is one of the categories of businesses. The data set we are given includes 81,434 reviews to 1,619 restaurants in Madison, which are separated into 48860 pieces of training data, 16287 pieces of testing data, and 16287 pieces of validation data. The purpose of this sentiment analysis project is to find out influencing factors to that make a review positive or negative, and construct a prediction model for the rating of reviews founded on texts and other provided variables in training data.

Selection of the Model

The original training data contains 54 variables, including stars, names of restaurants, text of reviews, date that reviews are posted, the number of people that find a review is useful, funny, or cool, city, number of characters and words, sentiments from -5 to 5, categories, and some words show up in text that might influence rating. The guideline of our model is to create useful new variables extracting from original variables because larger number of variables would lead to model’s increasing precision in fitting the outcome and decreasing prediction error. Generally, there are two ways we use to generate new variables.

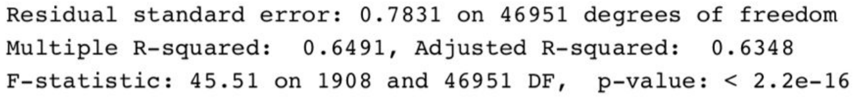
On one hand, we create dummy variables for new variables that are extracted from name, date, city and categories. For example, reviews of restaurants with the word “Burger” in their names or not are created into levels 1 and 0. The reason why we consider the word “Burger” in restaurants’ names is because users who like burgers are more likely to visit these restaurants. Likewise, reviews of restaurants with “Bar” in categories or not are also created into levels 1 and 0 because the word “Bar” represent the category of restaurants and has influence on rating of reviews to some degree. We create ten or more new variables for each of name and categories. In addition, we separate the city into Madison and not Madison. Moreover, we generate three dummy variables based on date, including weekdays and weekends, snowy months from November to April and non-snow months from May to October, and years before 2012 and years after 2012. Difference in weekdays and weekends, and months with snow or not might cause different moods of customers and then affect the ratings they give to the restaurants. The large range of years could result in more comprehensive and refined service from restaurants and greater demands from customers.

On the other hand, we calculate the number of each one word that shows up in each review text. The words we chosen are generally from our common sense, WordCloud Function, dictionary of text from 1 to 25000 with frequencies higher than 70, and Affin. Then we check histograms between potential variables and stars, twenty by twenty. Words that appear a clearly positive or negative relationship with stars in histogram are selected and added to the dataset, such as “favorite” and “again” in the histogram below. In addition to adding new variables, we also transform variables into log except categorical variables because the homoscedasticity is violated in residual plot of each variables. After applying log, the prediction error is smaller. In the end, we use multiple linear regression to fit variables, with stars as outcome and other variables as predictors.



Understanding of the Model

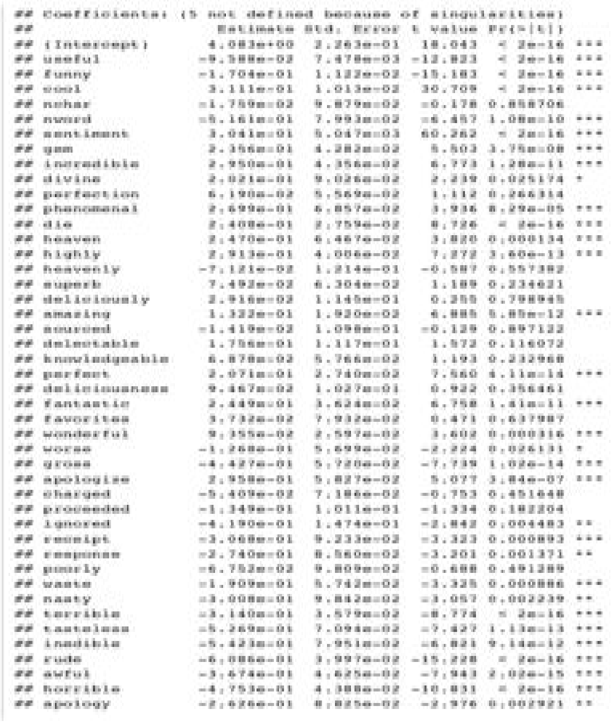
Multiple linear regression is used to fit the 48860 pieces of training data. In total. The best random forest model incorporated 1908 variables on four categories of features: the category of the restaurant, review text (words indicating emotion), feature of the date, and the sentiment of the review. According to the model summary, the residual standard error of the model is 0.7831. In our data, it implies, the actual star of each review can deviate from the true regression line by approximately 0.7831 on average. In other words, given that the mean star of all restaurants is 4.083 and that the residual standard error is 0.7831, we can say that the percentage error is 19.18%. In addition, the R-squared statistics provides us a measure of how well the model is fitting the actual data. The R-squared we get is 0.6491. In other words, roughly 65% of the variances found in the response variable(star) can be explained by the predictor variables we have used in the model. However, it should be noticed that R-squared will always increase as more variables are included in the model. Therefore, the adjusted R-squared seem to be a more important indicator here, which is 0.6348. Finally, the F-statistics is provided for the inference of whether there is a relationship between our predictor and the response variables. The further the F-statistic from 1 the better it is. In our example, the F-statistic is 45.51 which is relatively larger than 1 given the size of our data. Therefore, it is sufficient for us to reject the null hypothesis that there is no relationship between the star of the restaurant and the predictors. In conclusion, with a model that is fitting nicely, we could start to run predictive analytics to try to estimate the star rating for a random review given the information/predictors in the model. Now, it is time for us to take a closer look at the coefficients of the model. The coefficients of each predictor implies the effect of each variable on the actual star rating of the review. There are three different kinds of coefficients which may have different meaning. Firstly, considering the coefficient for categorical variables, such as the category of the restaurant. For example, if the restaurant is categorized as fast food, holding other variables constant, it is likely that the star rating will decrease by 0.2352. The second type of the variables is continuous variable, such the number of time one specific word showing up in the review text. For instance, the coefficient for the word cool is 0.311, which is saying that every extra time the word cool appears in the text, the star rating goes up by 0.311. In this case, the standard error can be used to compute an estimate of the expected difference in case we ran the model again and again. In other word, we can say the effect of the predictor cool is 0.01. Also, it should be noticed that the small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between the response variable and each predictor. Following this general idea, it is possible for us to come up with more general ideas on other variables.



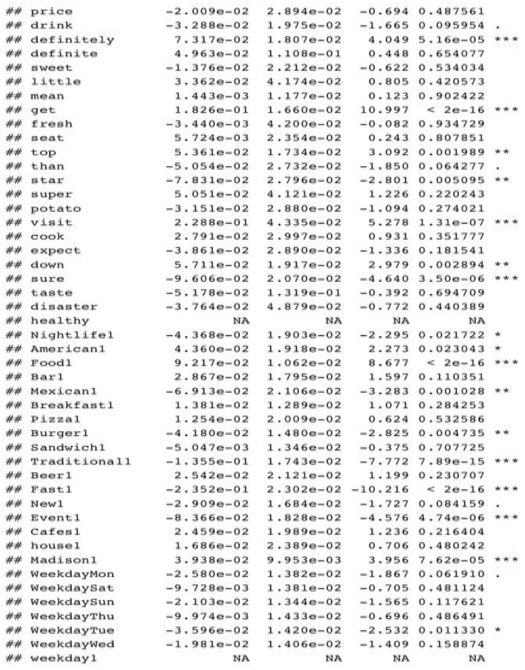
```
Residual standard error: 0.7831 on 46951 degrees of freedom
Multiple R-squared:  0.6491, Adjusted R-squared:  0.6348
F-statistic: 45.51 on 1908 and 46951 DF,  p-value: < 2.2e-16
```

Understanding of the estimation

Our model is a multiple linear regression model. The estimated value for the test and validate dataset are derived by plugging the according values for each variable into the fitted model. The values for each variable is transformed by taking log of them instead of being used directly from the columns.



```
## Coefficients: (5 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.083e+00  2.263e-01  18.043  < 2e-16 ***
## useful          -9.588e-02  7.478e-03 -12.823  < 2e-16 ***
## funny           -1.704e-01  1.122e-02 -15.183  < 2e-16 ***
## cool             3.111e-01  1.013e-02  30.709  < 2e-16 ***
## nchar           -1.759e-02  9.879e-02  -0.178 0.858706
## nword           -5.161e-01  7.993e-02  -6.457 1.08e-10 ***
## sentiment        3.041e-01  5.047e-03  60.262  < 2e-16 ***
## gem              2.356e-01  4.282e-02   5.503 3.75e-08 ***
## incredible       2.950e-01  4.356e-02   6.773 1.28e-11 ***
## divine           2.021e-01  9.026e-02   2.239 0.025174 *
## perfection       6.190e-02  5.569e-02   1.112 0.266314
## phenomenal       2.699e-01  6.857e-02   3.936 8.29e-05 ***
## die              2.408e-01  2.759e-02   8.726  < 2e-16 ***
## heaven           2.470e-01  6.467e-02   3.820 0.000134 ***
## highly           2.913e-01  4.006e-02   7.272 3.60e-13 ***
## heavenly        -7.121e-02  1.214e-01  -0.587 0.557382
## superb           7.492e-02  6.304e-02   1.189 0.234621
## deliciously      2.916e-02  1.145e-01   0.255 0.798945
## amazing          1.322e-01  1.920e-02   6.885 5.85e-12 ***
## sourced         -1.419e-02  1.098e-01  -0.129 0.897122
## delectable       1.756e-01  1.117e-01   1.572 0.116072
## knowledgeable    6.878e-02  5.766e-02   1.193 0.232968
## perfect          2.071e-01  2.740e-02   7.560 4.11e-14 ***
## deliciousness    9.467e-02  1.027e-01   0.922 0.356461
## fantastic        2.449e-01  3.624e-02   6.758 1.41e-11 ***
## favorites        3.732e-02  7.932e-02   0.471 0.637987
## wonderful        9.355e-02  2.597e-02   3.602 0.000316 ***
## worse           -1.268e-01  5.699e-02  -2.224 0.026131 *
## gross           -4.427e-01  5.720e-02  -7.739 1.02e-14 ***
## apologize        2.958e-01  5.827e-02   5.077 3.84e-07 ***
## charged         -5.409e-02  7.186e-02  -0.753 0.451648
## proceeded       -1.349e-01  1.011e-01  -1.334 0.182204
## ignored         -4.190e-01  1.474e-01  -2.842 0.004483 **
## receipt         -3.068e-01  9.233e-02  -3.323 0.000893 ***
## response        -2.740e-01  8.560e-02  -3.201 0.001371 **
## poorly          -6.752e-02  9.809e-02  -0.688 0.491289
## waste           -1.909e-01  5.742e-02  -3.325 0.000886 ***
## nasty           -3.008e-01  9.842e-02  -3.057 0.002239 **
## terrible        -3.140e-01  3.579e-02  -8.774  < 2e-16 ***
## tasteless       -5.269e-01  7.094e-02  -7.427 1.13e-13 ***
## inedible        -5.423e-01  7.951e-02  -6.821 9.14e-12 ***
## rude            -6.086e-01  3.997e-02 -15.228  < 2e-16 ***
## awful           -3.674e-01  4.625e-02  -7.943 2.02e-15 ***
## horrible        -4.753e-01  4.388e-02 -10.831  < 2e-16 ***
## apology         -2.626e-01  8.825e-02  -2.976 0.002921 **
```



```
## price           -2.009e-02  2.894e-02  -0.694 0.487561
## drink           -3.288e-02  1.975e-02  -1.665 0.095954 .
## definitely       7.317e-02  1.807e-02   4.049 5.16e-05 ***
## definite         4.963e-02  1.108e-01   0.448 0.654077
## sweet           -1.376e-02  2.212e-02  -0.622 0.534034
## little           3.362e-02  4.174e-02   0.805 0.420573
## mean             1.443e-03  1.177e-02   0.123 0.902422
## get              1.826e-01  1.660e-02  10.997  < 2e-16 ***
## fresh           -3.440e-03  4.200e-02  -0.082 0.934729
## seat             5.724e-03  2.354e-02   0.243 0.807851
## top              5.361e-02  1.734e-02   3.092 0.001989 **
## than            -5.054e-02  2.732e-02  -1.850 0.064277 .
## star            -7.831e-02  2.796e-02  -2.801 0.005095 **
## super            5.051e-02  4.121e-02   1.226 0.220243
## potato          -3.151e-02  2.880e-02  -1.094 0.274021
## visit            2.288e-01  4.335e-02   5.278 1.31e-07 ***
## cook             2.791e-02  2.997e-02   0.931 0.351777
## expect          -3.861e-02  2.890e-02  -1.336 0.181541
## down             5.711e-02  1.917e-02   2.979 0.002894 **
## sure            -9.606e-02  2.070e-02  -4.640 3.50e-06 ***
## taste           -5.178e-02  1.319e-01  -0.392 0.694709
## disaster        -3.764e-02  4.879e-02  -0.772 0.440389
## healthy                 NA         NA      NA       NA
## Nightlife1      -4.368e-02  1.903e-02  -2.295 0.021722 *
## American1        4.360e-02  1.918e-02   2.273 0.023043 *
## Food1            9.217e-02  1.062e-02   8.677  < 2e-16 ***
## Bar1             2.867e-02  1.795e-02   1.597 0.110351
## Mexican1        -6.913e-02  2.106e-02  -3.283 0.001028 **
## Breakfast1       1.381e-02  1.289e-02   1.071 0.284253
## Pizza1           1.254e-02  2.009e-02   0.624 0.532586
## Burger1         -4.180e-02  1.480e-02  -2.825 0.004735 **
## Sandwich1       -5.047e-03  1.346e-02  -0.375 0.707725
## Traditional1    -1.355e-01  1.743e-02  -7.772 7.89e-15 ***
## Beer1            2.542e-02  2.121e-02   1.199 0.230707
## Fast1           -2.352e-01  2.302e-02 -10.216  < 2e-16 ***
## New1            -2.909e-02  1.684e-02  -1.727 0.084159 .
## Event1          -8.366e-02  1.828e-02  -4.576 4.74e-06 ***
## Cafes1           2.459e-02  1.989e-02   1.236 0.216404
## house1           1.686e-02  2.389e-02   0.706 0.480242
## Madison1         3.938e-02  9.953e-03   3.956 7.62e-05 ***
## WeekdayMon      -2.580e-02  1.382e-02  -1.867 0.061910 .
## WeekdaySat      -9.728e-03  1.381e-02  -0.705 0.481124
## WeekdaySun      -2.103e-02  1.344e-02  -1.565 0.117621
## WeekdayThu      -9.974e-03  1.433e-02  -0.696 0.486491
## WeekdayTue      -3.596e-02  1.420e-02  -2.532 0.011330 *
## WeekdayWed      -1.981e-02  1.406e-02  -1.409 0.158874
## weekday1                NA         NA      NA       NA
```

Analysis of the model

There are strengths and weaknesses of our model. The first strength is the variety of the variables. The dataset we are analyzing has already been given variables extracted from the review text part of the data. For example, “superb”, “amazing”, “sourced”. These variables indicate the number of appearances of these words in each review text. There are two variables in the original dataset that can be directly used, which are “nword” and “nchar”. These two variables can be used directly or be put under statistics transformation directly.

In addition to the variables derived from the review text, we extracted more variables from the text session. This kind of variables takes a great percentage of the variables. Furthermore, we investigated other variables which the original dataset is equipped with. We managed to derive variables from other columns. For example, “date”, “location”. From the date, we were considering if the timing going to the restaurant would influence the review. The possible difference could come from weekdays and weekends, whether it snowed or not on that day.

The second strength is how we processed the data.Except for the dummy variables, we did not use the statistics directly. We tried data transformation, which were taking log and taking square of the numbers. To obtain the best transformation, we tried the combinations of these methods. Taking log only, taking square only, taking log first then taking the square, taking square first then taking the log. As it turns out, taking the square of the statistics would result in large variances. Taking log only is the best way to transform the data.

For the weakness, the first one is that there are many similar variables, which are derived from the review text session, are not combined in our model. For example, the variables called “delicious”, “deliciousness” and “deliciously” are separate ones. After our discussion, we agreed that these words usually have the same meaning and might be considered as similar expression of satisfaction.

The second weakness of our model is that the outliers are not eliminated. Not all the data provided in the dataset are suitable for fitting models. When we have decided which words can be extracted from the review text by looking at the histogram of number of appearances of the words vs star rating, we did not eliminate the outliers when fitting the model. This could have negative effect on the final model.

The third weakness of the model is that not all of the variables are as useful as we expected. For example, the variable “Madison”, which is a dummy variable indicating if the restaurant’s location is in Madison or not, has a small p-value. However, its coefficient is also very small. If a restaurant is in Madison, the star rating of the review would increase by 3.938e-2, which is negligible compared to other ones. However, when we tried to remove these variables that seem useless, the root mean square error for the reduced model increased, which indicate that these variables still have an effect of the model. In order to receive the best result, we did not remove these variables.

Conclusion

To sum up, each word show up in text could influence the sentiment of review and essentially determine whether a review is positive or negative. Through expanding data set by generating effective variables and using log transformation, the multiple linear regression, with stars as outcome and other variables as predictors, can fit the data set more precisely and the prediction error of fitted model is much smaller. Therefore, the multiple linear regression model we get from training data could fit in testing data and validation data and possibly produce more precise ratings of reviews.

Team Contributions

All team members contributed equally to the project (33.33% each). Each member is responsible for 1/3 of the write-up and 1/3 of the presentation. Key accomplishments are listed here:

- Lynette Gao: Examined using various models.
- Emily Xu: Data Cleaning to manage dataset size. Removed general stop words.
- Yicheng Yuan: Extracted Semantic Features. Identified yelp specific stop words