ISE 599: Applied Predictive Analytics, Fall 2019

Disaggregation of Yelp Review Ratings

GitHub Link:
https://github.com/zelongchen3/Disaggregation-of-Yelp-Review-Ratings

Professor: Mayank Kejriwal

Group:

Peiqing Lian

Xinrong Lian

Zelong Chen

Table of Contents

Executive Summary

Problem Statement

As a crowd-sourced business review and social networking website, Yelp has hundreds of millions of reviews. However, it only provides a holistic view by giving review ratings. In this project, we separated overall ratings by different categories to gain an insight into which categories will influence the stars most. For example, a 4 star review: "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks." This example includes different topics such as environment, service, food and price. By generating the topics and sentiment intensity score from the user's review, we can use linear regression to generate coefficients of each topics and understand the relationship between stars and topics.

Description of Solution

We focused on the reviews from Las Vegas restaurants. Firstly, cleaned the data and tokenized reviews by sentences. Secondly, use TF-IDF generate word features and NMF to reduce the dimension. Then grouped the word features under different topics. We labeled sentences with extracted topics and then generate sentiment intensity score, use those scores as predictors and each user's star as response to analyze the relationship between stars and topics through linear regression. By analyzing the coefficients of linear regression we can find out which topic influence the stars most.

Summary of Key Findings

By extracting topics we found out that when people giving reviews they care about the overall experience, price, food, service and if the restaurant is worth a try. After applying linear regression with mean square error of 1.034 and aic of 12958, it shows that among all the topics service influence stars most, the second influence is food. We also noticed that the restaurants near Strip (89109) showed a significant difference from others. Compare with service, food contributes more to the overall stars in postal code 89109.

Proposed Approach

Proposed approach to this problem involves the following steps:

1. Data Selection
2. Data Cleaning
3. Topic Extraction
4. Topic Labeling
5. Sentiment Analysis
6. Linear Regression +Evaluation

Data Selection

First step is to select a subset of the dataset from the Yelp Academic Dataset containing 6,685,900 reviews and 192,609 businesses. In our experiments, we will be focusing only on restaurants in Las Vegas, Nevada. We will first apply our model to one restaurant, Hash House A Go Go (5,847 reviews), then after validating our model we will then apply it to the datasets with the restaurants group by their postal code. We will do our analysis five of the postal codes in Las Vegas surrounding the famous Las Vegas Strip (89103, 89109, 89118, 89119, 89169).

Data Cleaning

Data cleaning step involves tokenizing the reviews by sentences, lowercasing, and the removal of common English stop words.

Topic Extraction

Topic extraction involves two steps, Term Frequency- Inverse Document Frequency (TF-IDF) and Non-Negative Matrix Factorization (NMF). We first feed in the corpus of all the reviews with each row being a tokenized sentence. TF-IDF Vectorizer generates a document feature vector for every tokenized sentence. Then using NMF, we are able to reduce the dimension of the TF-IDF and group the TF-IDF features under different topics. Using groupings of TF-IDF features, we then can come up with a general topic that encapsulates all of its TF-IDF features.

Topic Labeling

Using the NMF vectors that are corresponding to each tokenized sentence of the reviews, we can obtain the relevance scores of that sentence to the topics extracted by the NMF. Thus, each tokenized sentenced is labeled with the topic with the highest score.

Sentiment Analysis

After labeling every sentence, we then use VADER Sentiment Analysis to generate a sentiment intensity score of [-1,1] for that sentence. Then we aggregate the sentiment intensity scores of each user by averaging the intensity score under each topic.

Linear Regression + Evaluation

With the inputs of each user and their sentiment intensity score of each topic, we then fit a linear regression model to generate the coefficients of each topic that explains the relationship between stars and that particular topic. We mainly use Mean Squared Error (MSE) and Akaike Information Criterion (AIC) to evaluate our models. Then through cross validation of several different linear regression models such as linear regression, polynomial, ridge, lasso, we can confidently provide the model that have the lowest MSE and AIC scores. Then using the best model, we apply it to our test data and evaluate our final model with MSE and AIC scores.

Experimental Results

Mean square error and AIC are the primary evaluation criteria used. MSE and AIC have been evaluated for the training set and testing set over the machine learning models used (linear regression, polynomial, ridge, lasso). Our project's goal is to predict a yelp review star for a specific restaurant and find out which review topics contribute most to the star rating process. Our dataset has 5847 reviews and we split those reviews into 80% training and 20% testing data. NMF feature vectors are generated for training and testing. Training data is used for various regression models training and testing data is evaluated on the models that have been trained and out evaluation metrics show the promising results. For the reference and easy understanding of how MSE and AIC are calculated, the following shows,

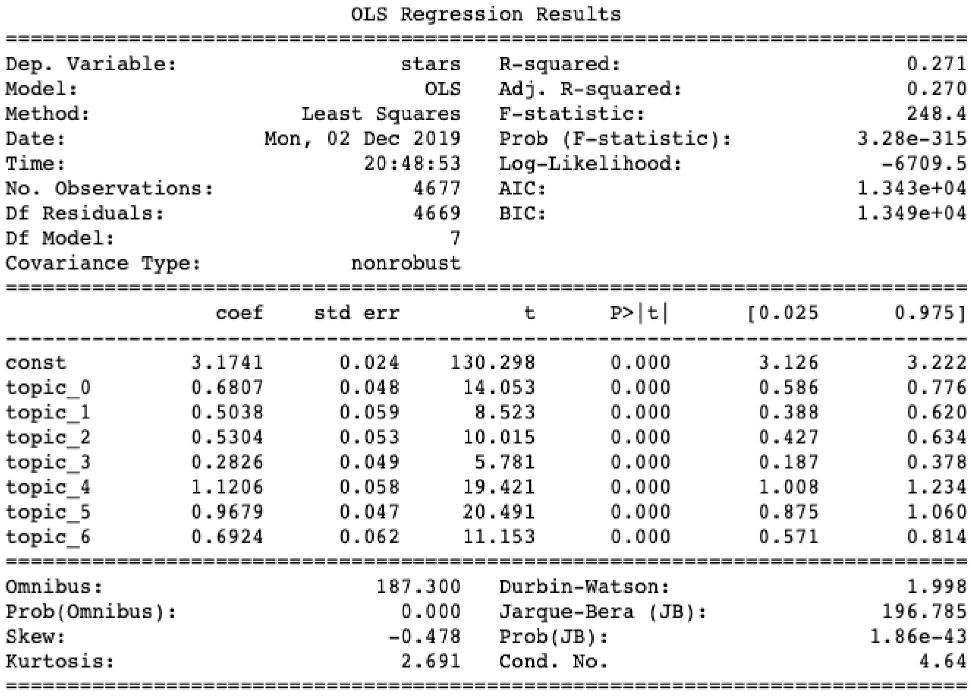
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

$$\text{AIC} = \log\left(\frac{\text{RSS}}{n}\right) + 2k$$

In our project, the y is the target variable which is the stars category for MSE calculation. For AIC, n is the number of observations and k is the number of variables.

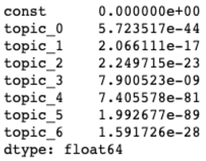
As why regression modeling has been utilized, it is that since we are predicting how many stars is going to be given based on the relationship between the star and topics we generated. We performed all the approaches that has been said in the last section. In the first, also considered as our baseline algorithm, we experimented with the simple linear regression with **two settings.** First one has all topics and intercept and we removed non-topics and intercept. We got respective MSE, AIC value, and p-values for each predictor. The result will be shown later below.

Model summary:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  stars   R-squared:                       0.271
Model:                            OLS   Adj. R-squared:                  0.270
Method:                 Least Squares   F-statistic:                     248.4
Date:                Mon, 02 Dec 2019   Prob (F-statistic):          3.28e-315
Time:                        20:48:53   Log-Likelihood:                -6709.5
No. Observations:                4677   AIC:                         1.343e+04
Df Residuals:                    4669   BIC:                         1.349e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.1741      0.024    130.298      0.000       3.126       3.222
topic_0        0.6807      0.048     14.053      0.000       0.586       0.776
topic_1        0.5038      0.059      8.523      0.000       0.388       0.620
topic_2        0.5304      0.053     10.015      0.000       0.427       0.634
topic_3        0.2826      0.049      5.781      0.000       0.187       0.378
topic_4        1.1206      0.058     19.421      0.000       1.008       1.234
topic_5        0.9679      0.047     20.491      0.000       0.875       1.060
topic_6        0.6924      0.062     11.153      0.000       0.571       0.814
==============================================================================
Omnibus:                      187.300   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              196.785
Skew:                          -0.478   Prob(JB):                     1.86e-43
Kurtosis:                       2.691   Cond. No.                         4.64
==============================================================================
```

Variables p-values:

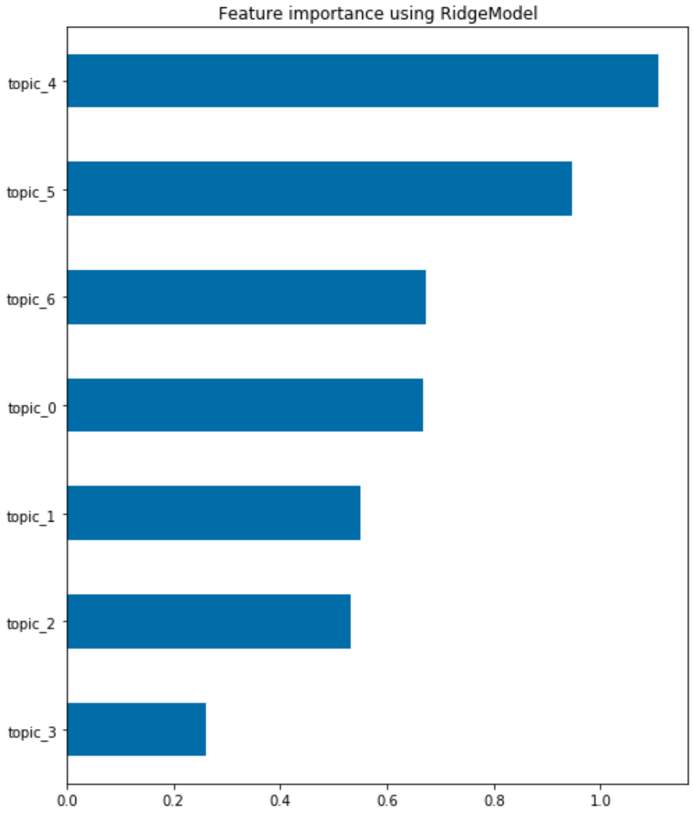


```
const      0.000000e+00
topic_0    5.723517e-44
topic_1    2.066111e-17
topic_2    2.249715e-23
topic_3    7.900523e-09
topic_4    7.405578e-81
topic_5    1.992677e-89
topic_6    1.591726e-28
dtype: float64
```

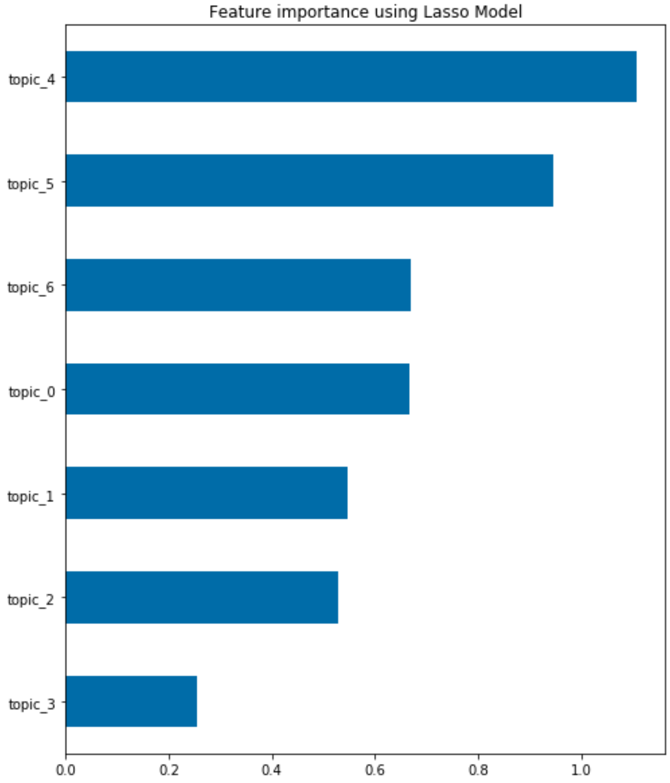
$$yhat = 3.17 + 0.68 * Service + 0.50 * Food1 + 0.53 * Worth + 0.28 * Food/Service + 1.12 * Topic4 + 0.97 * Food2 + 0.69 * Wait$$

The MSE is 1.0876 and AIC is 13434. Through model summary and variables p-values, we are not removing any variables since all p-values are way smaller than 0.05(assuming alpha level is 0.05). In the second settings of our first approach. The MSE is 6.120 and AIC is 21352. Metrics increases since it is the trade-off between interpretability and accuracy.

After the first approach, we experimented with RidgeCV, polynomial, and LassoCV regression respectively and reported the MSE values.

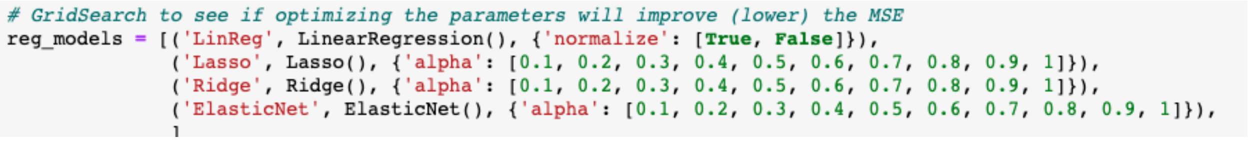


The RidgeCV average mse is 1.039 and this figure shows the feature importance of ridge model and we found topic4 is the most dominant one, which it is 1.11. And it is followed by topic5 that is 0.95.

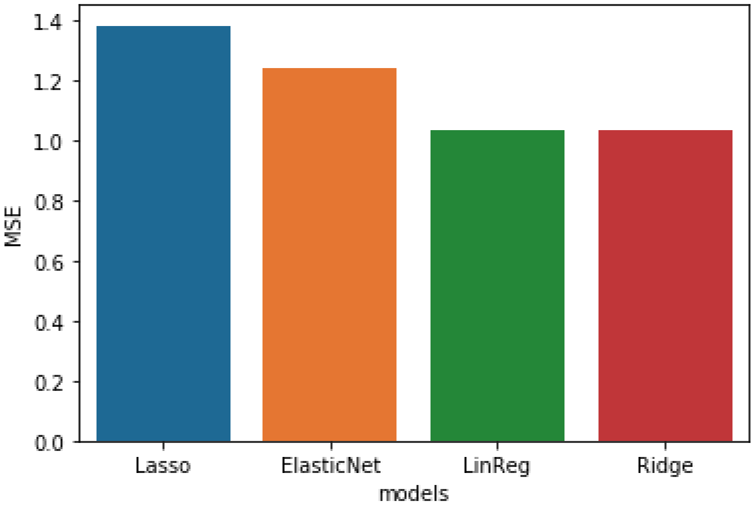


The LassoCV average mse is 1.083 and this figure shows the feature importance of lasso model and we found topic4 is the most dominant one, which it is 1.10. And it is followed by topic5 that is 0.94. It is not a huge difference between lasso and ridge in terms of feature importance.

In the last, we used cross-validation to validate performance of different algorithms(linear regression, lasso, ridge, ElasticNet) in totality and each algorithm has been tested with different choices of parameters using GridSearchCV.



```
# GridSearch to see if optimizing the parameters will improve (lower) the MSE
reg_models = [('LinReg', LinearRegression(), {'normalize': [True, False]}),
              ('Lasso', Lasso(), {'alpha': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]}),
              ('Ridge', Ridge(), {'alpha': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]}),
              ('ElasticNet', ElasticNet(), {'alpha': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]}),
              ]
```



Eventually, it is observed that the MSE is at its lowest(1.034) when ridge is our regression model.

Discussion

Initial analysis of the Las Vegas restaurant of Hash House A Go Go provided us several interesting results:

1. Simple Linear Regression had a mean squared error of 1.087 and AIC of 13435.64 and an equation of

$$yhat = 3.17 + 0.68 * Service + 0.50 * Food1 + 0.53 * Worth + 0.28 * Food/Service + 1.12 * Topic4 + 0.97 * Food2 + 0.69 * Wait$$

   a. Through the equation we can interpret the intercept to be customers' expectations of the restaurant before dining in. The coefficients tells us that Topic4 and Food2 are the most important topics to customers dining at this specific restaurant.

2. However, when we try to remove the intercept and unexplained topic so that the regression model can be explained by our interpretable topics we got a mean squared error of 6.117 and AIC of 21352.176 and an equation of

$$yhat = 2.41 * Service + 2.61 * Food1 + 3.11 * Worth + 2.82 * Food/Service + 3.40 * Food2 + 2.47 * Wait$$

3. This demonstrates the trade-off between Accuracy and Interpretability. Where our accuracy decreased by approximately six folds to gain interpretability.

After analyzing the data form postal code 89103, 89118, 89119, 89169 and the first 60000 rows from postal code 89109, the summary of the key insights are as below:

1. Extracted the influential topics form reviews which are overall experience, price, food, service and if the restaurant is worth a try.

2. Among five postal codes in Las Vegas (89103, 89109, 89118, 89119, 89169), in area 89103, 89118, 89119, 89169 the most important factor is service, and the second important topic is food.

3. In area 89019, where Strip is located, the most important factor is food instead of service, and the coefficient of "food" is almost twice as much as others.

4. The coefficients in different postal codes are similar, except the last two. We noticed that for the topic 5 in area 89103, 89118, 89119, 89169 the coefficient is 1.66~1.83 while for postal code 89109 is 0.75. Similarly, the coefficients of food in the first four postal codes are from 0.80~1.06 while the coefficient of food in 89109 area is 1.65. It shows factors that influence customer ratings vary by region.

Conclusion, Future Work, and Lesson Learned

Yelp reviews and ratings system are important as people make decisions on what to do and where to eat. It always has been the most trusted social network source. We think that understanding the relationship between stars and review topics can help users and business owners in each own benefits. For example, if an business owner would like to know how to attract more customers, it would be fundamentally meaningful if he or she understands what contributes the most to the business in terms of topics.

In this project, through data analysis and linear regression, we generated coefficients of each topic and understand the relationship between stars and topics with mean square error of 1.034 and aic of 12958. We also showed that ridge performs better by conducting k-fold cross-validation and parameter searching. In the end, though the aic of simple linear regression and aic of ridge regression have no big difference, we selected ridge as it has smaller mean square error value. By analyzing the equations we found out that service and food had the greatest impact on the stars, with the degree of impact varying by region.

In our future work, we can try out the performance of another dimensionality reduction algorithm of Latent Dirichlet Allocation (LDA) to reduce the dimensionality of our TF-IDF document features. We might be able to get other groupings of features that can produce clearer topics. Another alternative that we can try, is to tune the performance of VADER Sentiment Analysis or change to another program to better understand slang and modern phrases. As the current version we are using could not detect the correct sentiment intensity scores for phrases like “This place freaking rock” as a positive sentiment.

Work Cited

[1] https://www.yelp.com/dataset

[2] https://scikit-learn.org/0.15/auto_examples/applications/topics_extraction_with_nmf.html

[3] https://github.com/Vishwacorp/yelp_nlp/blob/master/2_text_processing.ipynb

[4] https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

[5] https://codeday.me/bug/20170703/30475.html

[6] https://github.com/ahegel/yelp-dataset/blob/master/Predicting%20Star%20Ratings.ipynb