Part 2: Predict star ratings from review text (Naïve Bayes Model vs Logistic regression)

2.1

**Multinominal Naïve Bayes Model**

accuracy score = 0.4712

The accuracy score is too low

Confusion matrix:

Predicted

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 star | 2 star | 3 star | 4 star | 5 star | Producer accuracy |
| 1 star | 55 | 14 | 24 | 65 | 27 | 29.73% |
| 2 star | 28 | 16 | 41 | 122 | 27 | 6.838% |
| 3 star | 5 | 7 | 35 | 281 | 37 | 9.589% |
| 4 star | 7 | 0 | 16 | 629 | 232 | 71.154% |
| 5 star | 6 | 4 | 6 | 373 | 443 | 53.245% |
| User Accuracy | 54.455% | 39.022% | 28.689% | 42.789% | 57.833% |  |

Analysis: The prediction for 1,4,5 stars are relatively more accurate than 2,3 stars.

Conclusion: This implies that we should make some adjustments to the data itself for training and testing data or try a different model to increase its accuracy.

**Logistic regression**

Accuracy score : 0.4868

The accuracy score is still very low

Confusion matrix :

Predicted

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 star | 2 star | 3 star | 4 star | 5 star | Producer accuracy |
| 1 star | 76 | 43 | 19 | 23 | 24 | 41.081% |
| 2 star | 37 | 71 | 63 | 35 | 28 | 30.342% |
| 3 star | 10 | 32 | 105 | 148 | 70 | 28.767% |
| 4 star | 11 | 22 | 90 | 456 | 305 | 51.584% |
| 5 star | 11 | 22 | 90 | 285 | 509 | 61.178% |
| User Accuracy | 52.778% | 41.279% | 38.884% | 48.152% | 54.28% |  |

The accuracy between different stars are relatively spread out compare to the Naive Bayes Classifier. While, 1 star and 5 stars have the highest accuracy among the others. Further examinations should be considered on 1 and 5 stars could be examined.

Would the increase in logistic regression be any different than the increase in Naive Bayes Classifier after filtering out some unreliable review?

Original:

Naive: 0.4712

Logistic: 0.4868

New:

Naive: 0.4805

Logistic: 0.4951

Difference:

Naive: 0.0093

Increase in percentage = 1.97%

Logistic: 0.0083

Increase in percentage = 1.71%

Naive increase more in terms of the difference and percentage, however it’s not statistically significant (0.2%) .

2.2

Predicted

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 star | 5 star | Producer accuracy |
| 1 star | 126 | 58 | 64.478% |
| 5 star | 25 | 813 | 97.017% |
| User Accuracy | 83.444% | 93.341% |  |

Analysis: The prediction for accuracy on 1 stars is lower than 5 stars.

For False positive reviews, we were able to identify common wordings in the reviews, such as ‘good’,’nice’,’impressive’ in it, generally positive, however these words are used in a sarcastic manner or are combined with other wordings that might result in a negative meaning . The machine learning correlates wordings with 5 star ratings training data set, resulting in these False positive reviews.

For False negative reviews, there are negative words, such as ‘sad’ in it. The machine learning correlates negative wordings with 1 star ratings, resulting in these false negative reviews.

|  |
| --- |
| The percentage of False positive reviews (93.341%) is lower than the percentage of False negative reviews(83.44%). |

2.3

The top ten token words in the training data for 5 star reviews are mostly positive adjectives, yet, some are totally unrelated or neutral words that does not necessary correlates to the star ratings, such as ‘brunch’ and ‘gem’.

While the top ten token words in the training data for 1 star reviews are mostly negative adjectives, such as ‘disgusting’, ‘filthy’ , however some are totally unrelated or neutral words that does not necessary correlates to the star ratings,such as ‘staff person’ bring the number 1 on the list, this might imply that a lot of the dissatisfaction of the restaurant in the data set are correlated to the staffs or service in the restaurants.

2.4

The logistic regression model gives out a better accuracy compared to Naive Bayes Classifier by nearly 50%. We examined these two model’s accuracy rate with everything, including dataset, default test and split rate being constant.

Naive Bays Classifier uses Bayes Theorem and have strong independence condition among features, it considers each of the features in the text to contribute independently to the probability for the classification. However, in our dataset, this model is not suited because the review per se should be reviewed as a whole feature by itself since positive wordings doesn’t necessary imply positive feelings or positive reviews, vice versa.

While, Logistic Regression combines one or more independent variables into the indicator variable, it would be suited better for our data. Thus, having a significant difference in the accuracy level between these 2 models.

However when we’re comparing these 2 models for all star ratings, the accuracy score would have such a big difference, despite logistic regression having a 0.01 higher accuracy score than Naive Bayes Classifier.

Thus, it would always be better to use the logistic regression model on our data set.

Part 3: Filter unreliable reviews

Naive Bayes Classifier

Old Accuracy = 0.4712

Accuracy = 0.4804719283970708

The accuracy increased by 0.01 after filtering some unreliable reviews. However, the accuracy score is still very low. This might mean that we need to weight the analysis on 1 to 5 stars together into lower ranking ,or consider another an alternative for evaluating the reviews of the restaurants are taking the 2 cases (i.e 1 and 5 stars) to predict the stars from the text, it is proven to be more accurate which will be demonstrated in the following tasks.