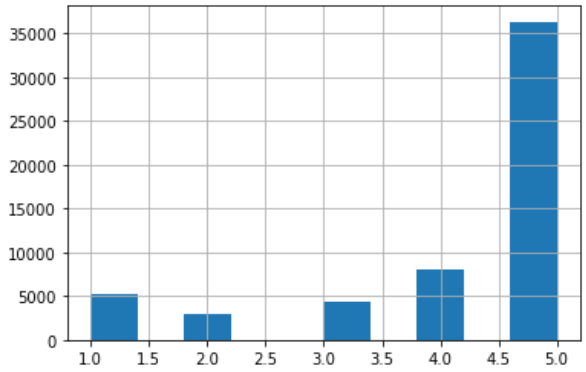
MISM6203 Project 4

Michael Conan

Product Review Classification

Machine learning is very capable of using unstructured text to make predictions. The review data provided includes the product review summary and detailed text as well as the score for the product, from 1 to 5. The review data is very skewed, with the majority of the reviews providing reviews with the top score of 5. This is not a problem for our classification problem, but we should keep this in mind when we are designing and training our models. To predict the score of reviews based on the text, we will first apply a bag of words technique to create a word count matrix, then we will apply Logistic Regression and XGBoost classification models .



The model that performed the best on the data was Logistic Regression, specifically using a L1 ration of 0.5 and regularization strength of 0.003, which yielded 69% accuracy. Not only did this model outperform the XGBoost for regression and classification, which only yielded accuracy as high as 63%, but the variable importance shows that more words were given weighting in the model, which means it is more likely to generalize well to new data.

While the model is only near 70% accurate at predicting the sentiment of the review, the method we used to calculate accuracy only considers whether the prediction was exactly correct or not, and does not consider the degree of error. Because our data is ordinal, meaning it is on a linear scale, predictions can be more or less accurate even if they are wrong. Because of this factor, we should consider the 70% pure accuracy rate to be relatively strong performance. With our data skewed toward a certain category, it is important that we also consider the precision and recall scores, as models may achieve strong accuracy by over-predicting the dominant category. For our best penalized logistic regression, the precision score is 62% and recall is 69%, which appears to be a relatively balanced model.

The words that had the most influence on the Logistic Regression model include “best”, “love”, “delicious”, and “perfect”. The Logistic Regression model uses coefficients to predict each category, so there are actually distinct importance scores for each word. One way to assess their combined importance is to sum the absolute value of the coefficients, which is a reasonable assessment of the impact a specific word has on the overall score.

word	imp_score0	imp_score1	imp_score2	imp_score3	imp_score4
best	-0.158179	-0.062162	-0.073198	0.048116	0.297952
love	-0.136387	-0.091709	-0.036383	0.022655	0.294954
delicious	-0.129686	-0.093132	-0.033292	0.067376	0.240886
loves	-0.132212	-0.041328	-0.018697	0.030749	0.213918
excellent	-0.105680	-0.054246	-0.048709	0.055406	0.205214
perfect	-0.084454	-0.075045	-0.091934	0.099853	0.203319
highly	0.000000	-0.025437	-0.028707	0.000000	0.190630
wonderful	-0.055992	-0.077144	-0.024744	0.022818	0.187626

word	value
best	0.639607
love	0.582087
delicious	0.564372
perfect	0.554605
excellent	0.469255
disappointed	0.462252

Our client can use this model not only to predict the score of reviews that did not include them, but also to identify trends in the comments customers are leaving and how they feel about the products as a whole.