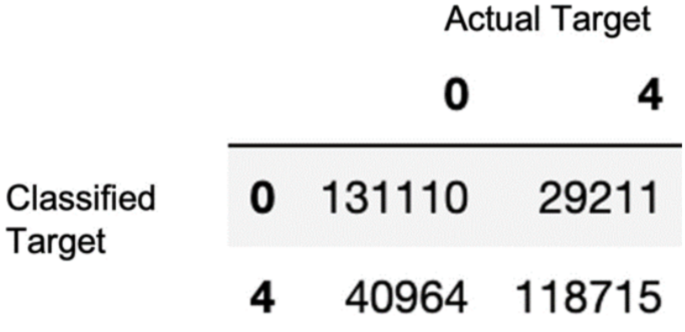
Sri Tarun Gulumuru
CSC 44800
12/14/23

Naive Bayes Model

The Naive Bayes model is the one of the most popular models for text classification. This is mainly because this model performs well with high-dimensional data which allows it to better extract features from a provided dataset. Naive Bayes is known for having a higher bias and lower variance, making it better than Logistic Regression as this model requires less data to analyze. The Naive Bayes model is based on a culmination of classification algorithms that follow the Bayes' Theorem, which is shown below:

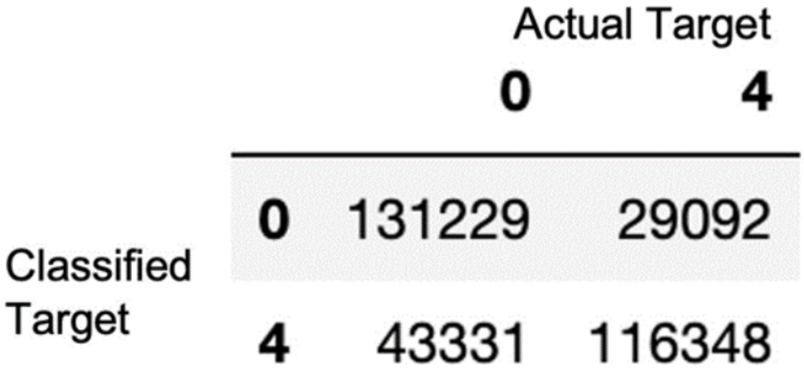
$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

As I did with Logistic Regression, I imported the dataset, selected the fields I wanted to analyze and split the dataset between 80% training data and 20% testing data. I also tested this model using Count Vectorization and TF-IDF Vectorization. For Count Vectorization, I had a final accuracy of 0.780703125, which was lower than the Logistic Regression model. The resulting classifications of this model are shown in Figure 1. As shown below, there are 131,110 true positives and 118,715 true negatives from this test, which is also less than the Logistic Regression model. Despite performing a little worse than the Logistic Regression model, the Naive Bayes model still did reasonably well. For TF-IDF Vectorization, the resulting accuracy was 0.773678125. Something I found interesting here is that TF-IDF performed a little poorly compared to Count Vectorization. This could mean that based on the type of model used, the type of vectorization may be better or worse. The classification results are shown in the Figure 2 below and show 131,229 true positives and 116,348 true negatives. Although the accuracy has lowered, the overall accuracy score is still very close to 80% which means that the model is still fairly accurate.



		Actual Target	
		0	4
Classified Target	0	131110	29211
	4	40964	118715

Figure 1: *True positive and true negative classifications of the Naive Bayes model using Count Vectorization.*



		Actual Target	
		0	4
Classified Target	0	131229	29092
	4	43331	116348

Figure 2: True positive and true negative classifications of the Naive Bayes model using TF-IDF Vectorization.