Model Selection and Evaluation

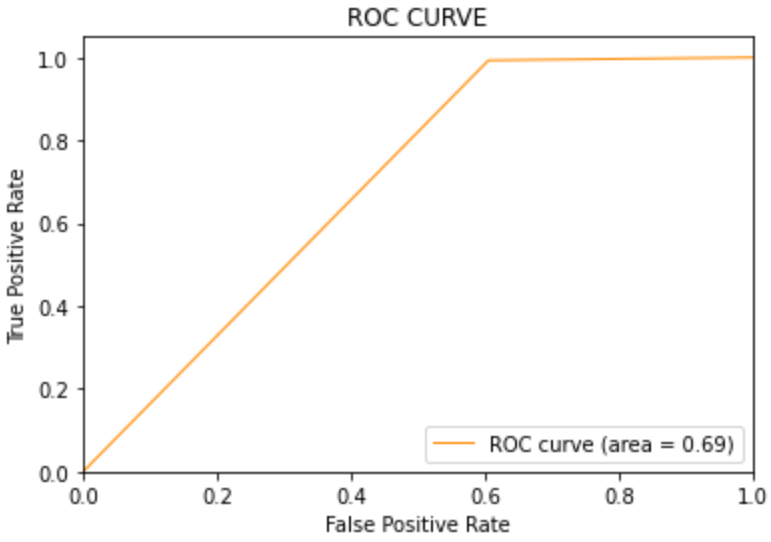
From checkpoint 1 we were originally going to pull data on popular Twitter topics in different cities. We are now going to look at the sentiment of the Tweets we pulled based on a specific topic. We are doing a sentiment analysis of the 2022 Super Bowl halftime performance. Our method of collection will still be the same pulling tweets directly from users as stated from checkpoint 1. After cleaning our data we are going to use the NLP to understand the tweet and help classify the sentiments that are expressed in the text source.

For our model of choice we used Sklearn's Logistic regression.

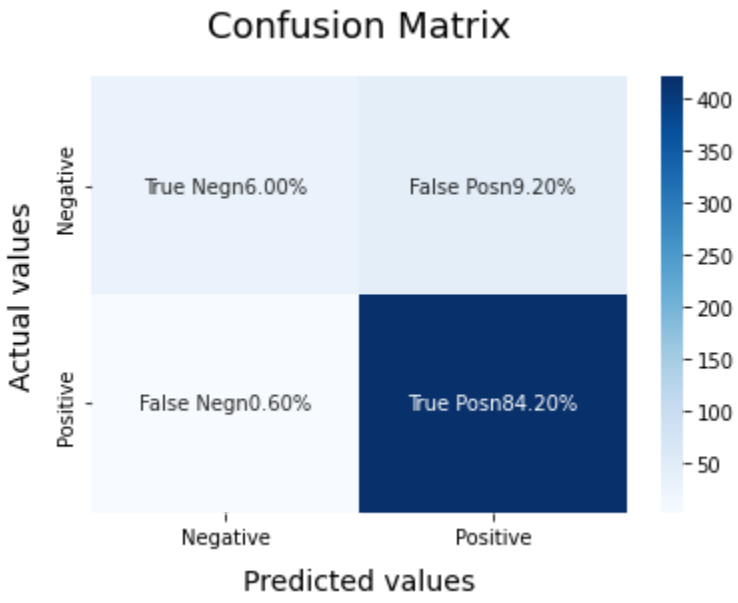
- **Justify your model choice based on how your response is measured and any observations you may have made in your EDA.**
  - Logistic regression is employed when the label of interest is a discrete value. Since we are seeking a prediction that is either positive "1" or negative "0" logistic regression was the appropriate supervised algorithm to employ. LR is also efficient, easy to interpret, and fast at classification tasks. LR performs well when the data is linearly separable, which is how our data is organized.
  - Precision: or the number of positive and negative predictions that were identified correctly for the model was accuracy of the model.
  - Recall: the measure of our model correctly identifying True Positives. Thus, for all the tweets that had positive reactions to the superbowl, recall tells us how many we correctly identified.

	precision	recall	f1-score	support
0	0.91	0.39	0.55	76
1	0.90	0.99	0.95	424
accuracy			0.90	500
macro avg	0.91	0.69	0.75	500
weighted avg	0.90	0.90	0.89	500

- **Report the model's test error rate using one of the techniques we discussed in lecture. Justify your choice.**
  - The technique we are using for the rest error rate is an ROC curve. The ROC curve is a graph showing the performance of a classification model at all classification thresholds. This model shows the relationship between the true positive rate and false negative rates. The area under the curve (AUC) is a measure of separability. The model represents the probability that an actual positive data point is positioned to the right of an actually negative data point. Using this model helped us check the accuracy of the model's performance.



- **Based on the estimated test error rate, discuss how well the model fits the data.**
  - Being that this data set was smaller than we anticipated there is still further testing that needs to be done. Working with more tweets will allow us to see if the model is skewed too far in one direction. However, for the logistic regression model the success rate was broken down into two main scores, the true positive score being 84% and the true negative score being 6.00%. There were other values that were observed, but based on the TP-score and TN-score the model accurately found the difference between tweets that had positive things and negative things to say about the superbowl halftime show



- **Use the model to make predictions for at least three cases of interest.**
    - Disclaimer: we run our model by pickling the vectorizer and classifier and deploying it on a simple flask application.

Test case	Model response
I hate the superbowl, but the halftime show was awesome!	Positive sentiment (Correct)
The superbowl halftime show this year was okay- not the best, but definitely not the worst.	Positive sentiment (Truly neutral but that is not an option)
Not really sure what the halftime show planning committee was thinking, but Eminem in the year 2022 is not good!	Positive sentiment (False positive)