Malicious Compliance or Pro Revenge?

classifying Reddit posts using NLP in Python

Tiffany Baker
9/25/2020

As defined…

Malicious Compliance:

People conforming to the letter, but not the spirit, of a request.

ProRevenge:

People going out of their way and going above and beyond to get revenge

The approach

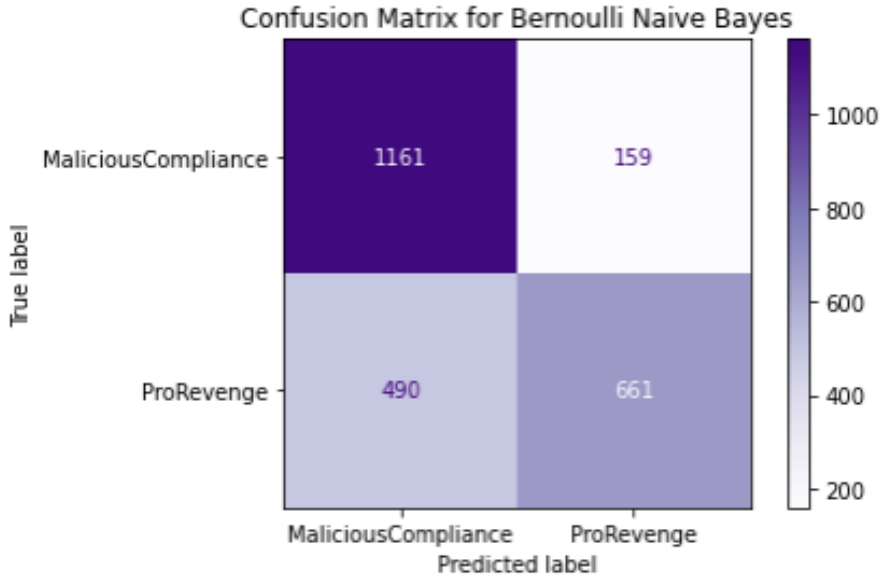
- Acquire around 5000 useful posts from each subreddit
- Develop a bespoke stopwords dictionary comprised of the tokens that are found in the top 500 tokens of each subreddit after processing through the standard English stopword dictionary
- Determine which classifier works best

Bernoulli Naïve Bayes

- Specificity 87.8%
- Accuracy 73.7%

- Vectorized based on token appearance, not frequency

```
Confusion Matrix for Bernoulli Naive Bayes
[[1161  159]
 [ 490  661]]
```

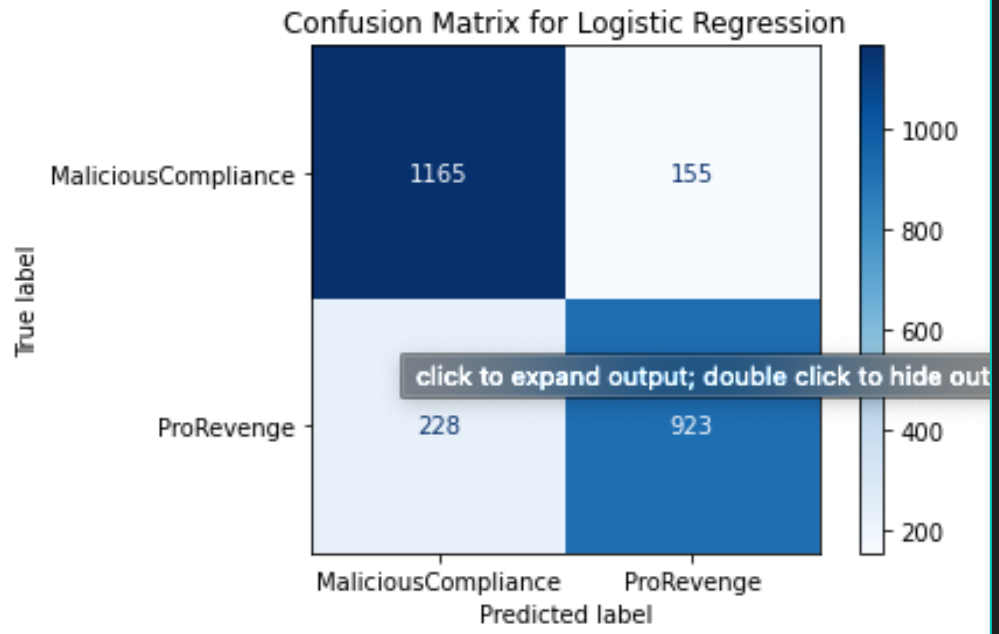


Logistic Regression- SAGA

- Specificity 88.3%
- Accuracy 84.5%

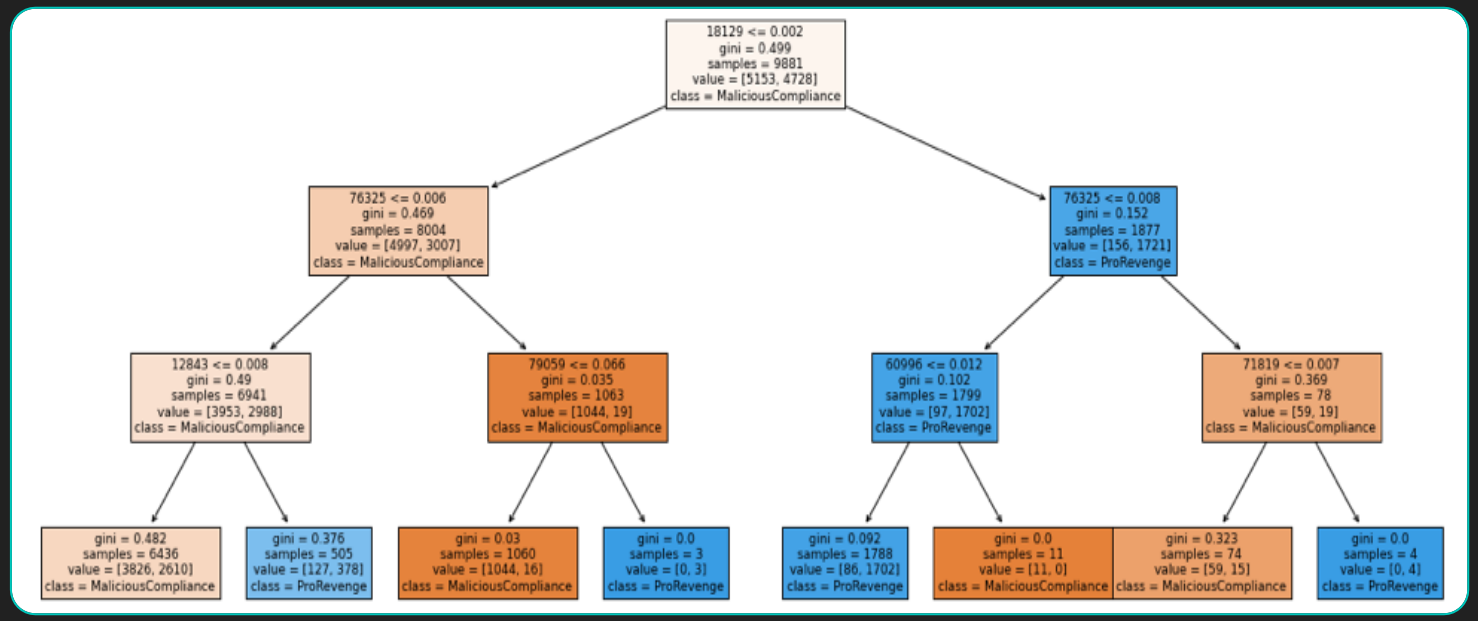
- Used SAGA solver since all features are binary
- LBFGS solver returned same
- Kept L2 regularization in model (Ridge)

```
Confusion Matrix for Logistic Regression
[[1165  155]
 [ 228  923]]
```



Decision Tree

- Use gini impurity
  - Gini impurity is the probability of classifying data incorrectly if the data is randomly selected and randomly assigned based on class distribution
- Score on train data 71.1%
- Score on test data 73.5%
- Score defined as the mean accuracy on the test data and labels



The results?

- Logistic regression provided the best bang for the buck
- Not only are the results promising, but model also has plenty of flexibility for further tuning
- Given a post from either subreddit, classification is possible with respectable accuracy