Model Metrics File

Model Features

The features we used were all feature engineered from the underlying data. We focused on the review summary and the review text. The review summary is a one line description of what the reviewer wants to title the review, and the review summary is an open ended section where reviewers can go in depth about their experience with a product. Listed below are all the features used in model production. These features were engineered for both review text and review summary sections.

- character count – number of characters in each section
- word count – number of words in each section
- sentiment score – Textblob's sentiment scoring for each section between -1 and 1. 1 being the most liked and -1 being the most disliked
- ease of reading – Flesch's ease of reading scale, where the higher the number the easier the text is to read
- vader score – sentiment analysis specifically for social media, these scores were divided into three categories, negative, neutral, and positive.

Hyperparameters

For each model we first ran the algorithms with the default settings, before using Grid Search to tune for the best hyperparameters.

Random Forest:

```
RandomForestClassifier(min_samples_leaf=10, n_estimators=20, n_jobs=-1)
```

Logistic Regression:

```
LogisticRegression(C=0.01, random_state=40)
```

Naïve Bayes:

```
GaussianNB(var_smoothing=0.08111308307896872)
```

Performance Metrics

We reported out precision, recall, and f1-score and used this as a baseline to see if improvements were being made during hyperparameter tuning. We also collected ROC-AUC curves that were cross-validated for more accuracy. This curve gives us some intuition on how much our model can encompass and show potential areas for improvement. We also used confusion matrices to show the true-false rate for each model. The most important metric for this study would be recall because we want to encompass as many spam reviews as possible. If our model overfits in this scenario it is not as costly or devastating as those lost reviews will not have a great effect on the purchasing decision of a customer