CS5100

Wikipedia Classification - Tanner Huynh

Corpus Cleaning

- The attack annotations averaged for greater than 50%
- Removing newline and tab tokens
- Removing extra spaces and punctuation
- Iterative Possibilities: spellchecker

Features

- TfidfVectorizer to weight occurrences
- Analyzer for word and char
- Lowercase
- Unigram model (1,1) (1,2)
- Default dictionary for stop words

Pipeline

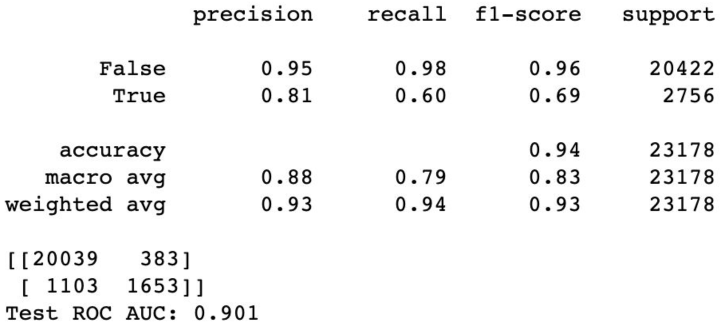
- Sklearn Pipeline
  - Word TfidfVectorizer
  - Tree classifier
- Imblearn Pipeline
  - Feature Union Word and Char Vectorizer
  - Over and under sampling
  - Linear Classifier
- Parameter grid for grid search

Sklearn Hyperparameters

- Random Forest Classifier
  - Gini and Entropy Information Gain
  - Max depth
  - Weighted
- Works best for imbalanced data because of the tree structure

Sklearn Metrics

- Improvement in recall
- Predict 60% of attacks
- More mileage can be gained with imblearn



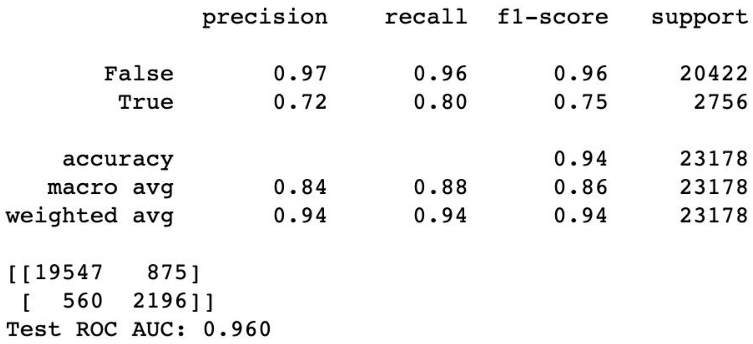
```
              precision    recall  f1-score   support

       False       0.95      0.98      0.96     20422
        True       0.81      0.60      0.69      2756

    accuracy                           0.94     23178
   macro avg       0.88      0.79      0.83     23178
weighted avg       0.93      0.94      0.93     23178

[[20039   383]
 [ 1103  1653]]
Test ROC AUC: 0.901
```

Imblearn Hyperparameters

- SMOTE (Synthetic Minority Oversampling Technique)
  - Nearest k-neighbors
  - Modified sampling strategy
- Random Undersampling of majority
  - Modified sampling strategy
- Logistic Regression Classifier

Imblearn Metrics

- Improvement in classifying attacks
- Less FP than Sklearn model
- Better recall!
- Better accuracy with more data



```
              precision    recall  f1-score   support

       False       0.97      0.96      0.96     20422
        True       0.72      0.80      0.75      2756

    accuracy                           0.94     23178
   macro avg       0.84      0.88      0.86     23178
weighted avg       0.94      0.94      0.94     23178

[[19547   875]
 [  560  2196]]
Test ROC AUC: 0.960
```