Classification

BY DAVID MARMOR

Problem statement

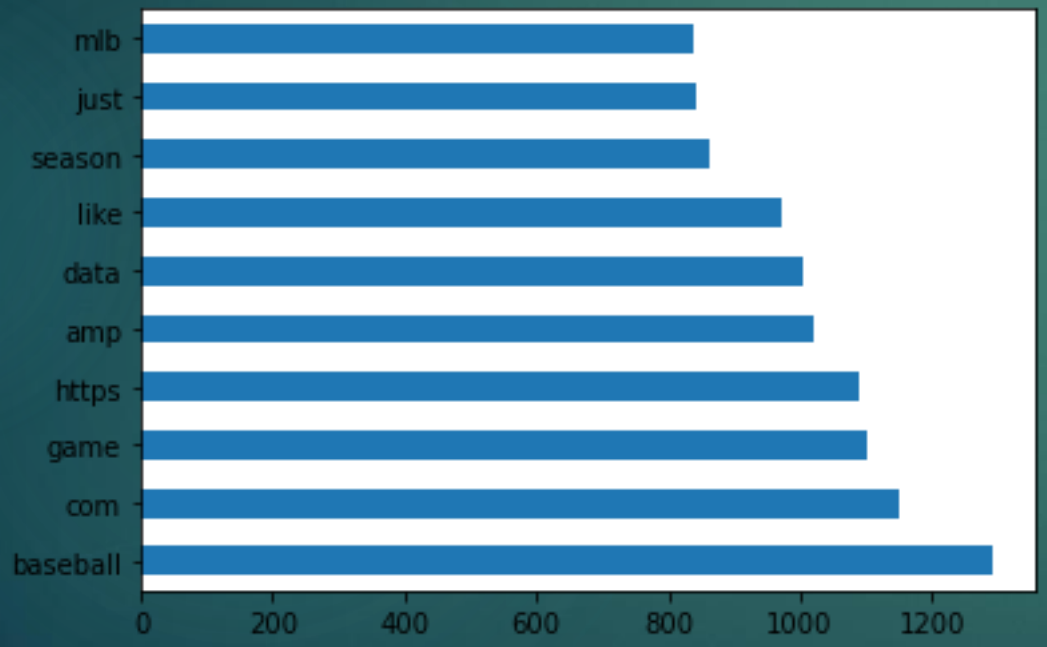
- We are trying to help Major League Baseball(MLB) with fan outreach
- Do all fans interact with the sport in similar ways
- Can we build a model to predict which baseball subreddit a post comes from?
- A good model means MLB may want to tailor their ads.
- Sabermetric vs mlb subreddits

Data Cleaning

- Got rid of deleted, removed, and blank posts.
- That left 3,901 posts.
- Slight data imbalance: 50.7% from the mlb subreddit.
- That is our baseline.
- Target variable mapping (mlb 0 sabermetrics 1)

EDA

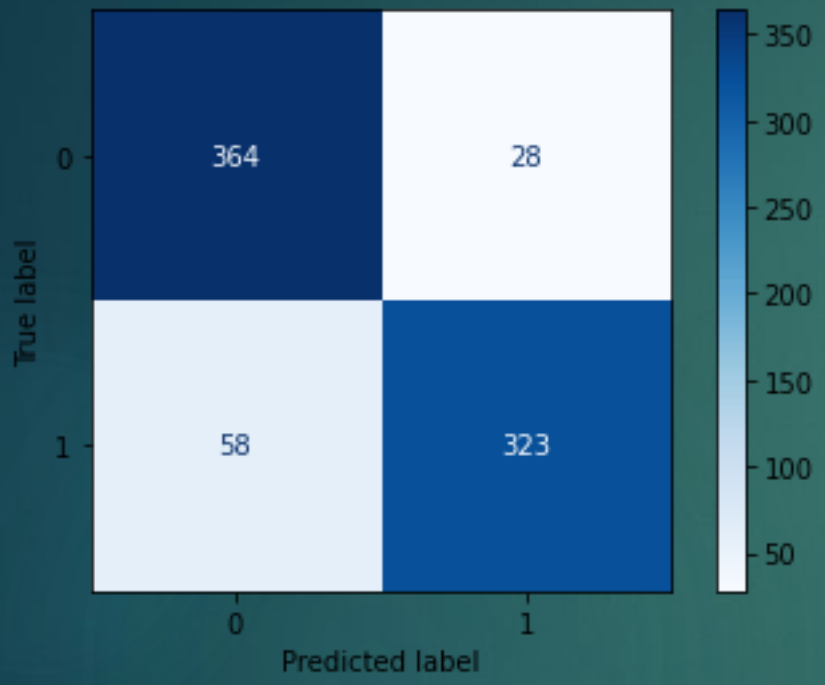
Most common words with stop words removed.



Model information

- We will use ada boosting, random forests, logistic regression, and naive bayes
- The posts were vectorized using CountVectorizer and TfidVectorizer
- The vectorizer that had the higher accuracy score for each model was used.
- We will be analyzing the posts.
- All models grid search over hyper parameters.

AdaBoost



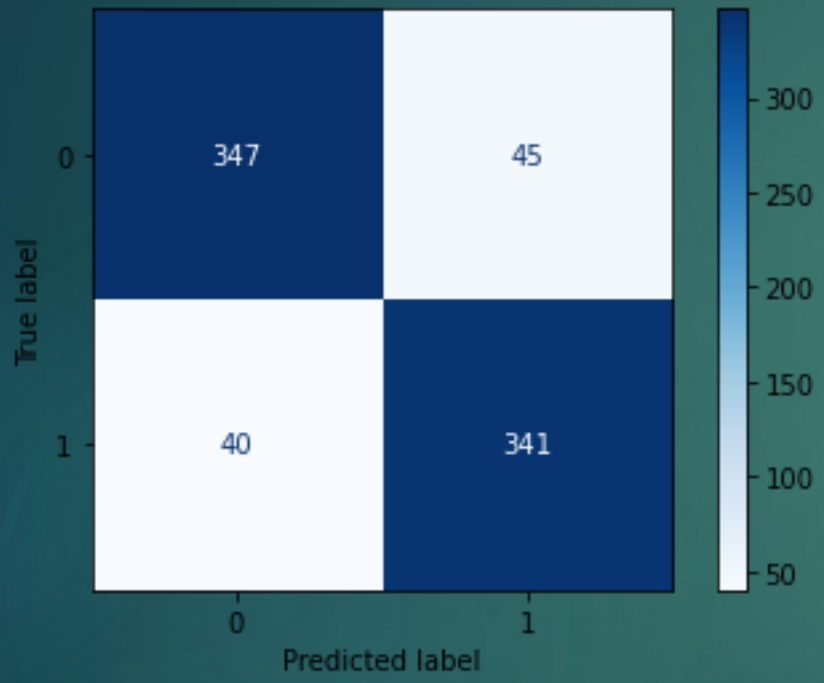
{'cvec__max_df': 0.9, 'cvec__max_features': 1000, 'cvec__min_df': 2, 'cvec__ngram_range': (1, 1), 'cvec__stop_words': 'english', 'model__n_estimators': 100}

Metrics

Metric	Testing Data
Accuracy	.889
Recall	.848
Precision	.920
Specificity	.929
Negative Predictive Value	.863

The accuracy on the training set was .938. The model is overfitting.

Random Forests



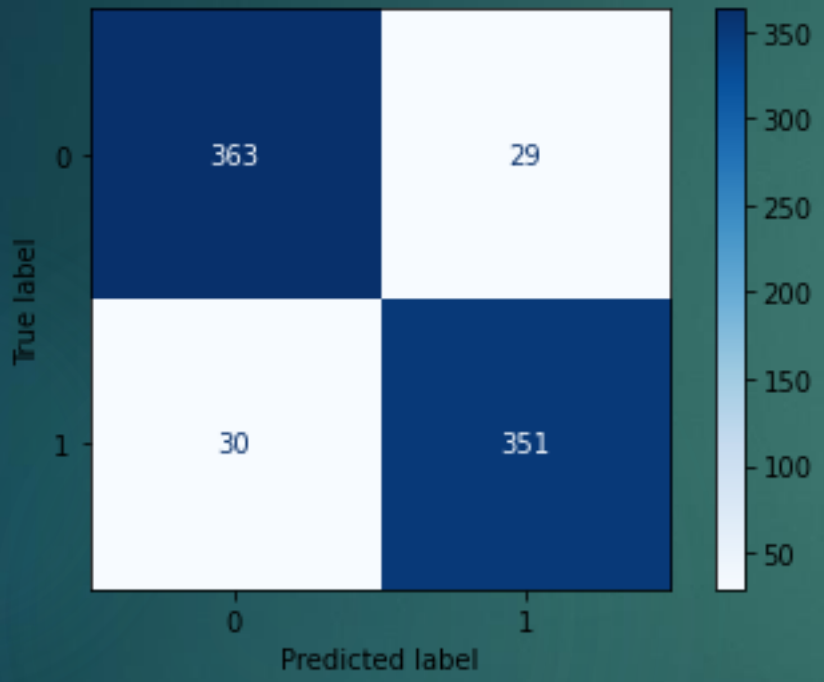
{'cvec__max_df': 0.9, 'cvec__max_features': 2000, 'cvec__min_df': 2, 'cvec__ngram_range': (1, 2), 'cvec__stop_words': 'english', 'model__max_depth': None, 'model__n_estimators': 50}

Metrics

Metric	Testing Data
Accuracy	.890
Recall	.895
Precision	.883
Specificity	.885
Negative Predictive Value	.897

The accuracy on the training set was .998. The model is overfitting.

Logistic Regression



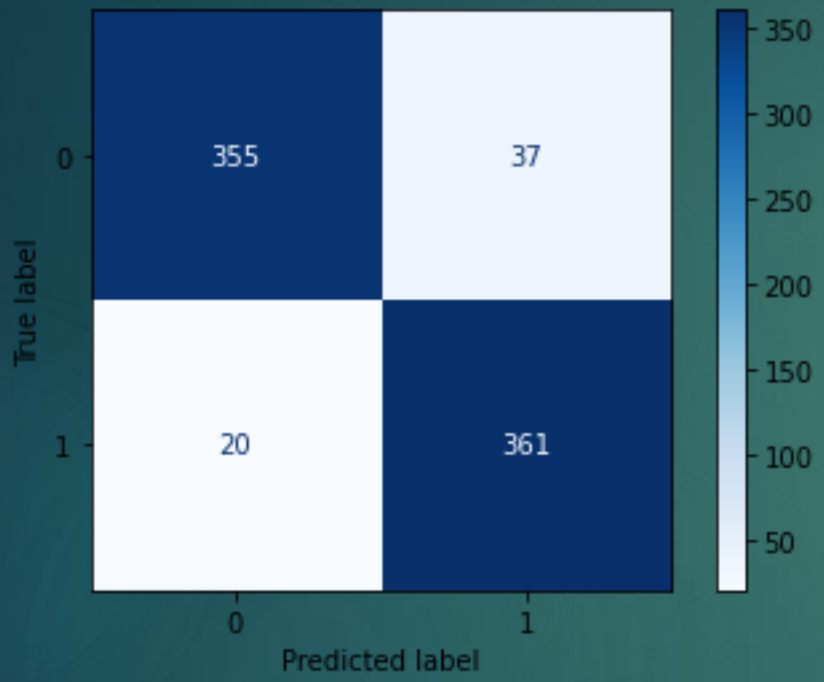

```
{'model__C': 1, 'model__solver': 'liblinear', 'tvec__max_features': 5000, 'tvec__ngram_range': (1, 2), 'tvec__stop_words': 'english'}
```

Metrics

Metric	Testing Data
Accuracy	.924
Recall	.921
Precision	.924
Specificity	.926
Negative Predictive Value	.924

The accuracy on the training set was .972. The model is overfitting.

Naive Bayes




```
{'tvec__max_features': 4000, 'tvec__ngram_range': (1, 1), 'tvec__stop_words': 'english'}
```

Metrics

Metric	Testing Data
Accuracy	.926
Recall	.948
Precision	.907
Specificity	.906
Negative Predictive Value	.947

The accuracy on the training set was .956. This model has the least overfitting. It also has the highest accuracy on the test score.

Conclusion

- There is a difference between mlb and sabermetric subreddits.
- The naive bayes model is the best model for predicting that difference. It has the best accuracy and the least overfitting.
- Because there are differences MLB should consider targeted ad campaigns.

Further Work

- Further model tunning to reduce overfit and increase accuracy
- Add post title and comments
- Next project is using the naive bayes model to identify the differences so MLB can target a variety of fans.
- Repeat this process for other baseball subreddits.

Thank You

- Questions?