Reddit Scraping

James Sobrino

Problem Statement

Your company creates reddit bots that farm karma with the goal of selling accounts with significant karma. You are part of a project to create a bot that will take posts from r/nba and crosspost it to the team subreddit associated with the post. Eg: A post on Klay Thompson is identified by the bot and crossposted to r/warriors. Your task is to create a model that can correctly identify posts between 1 team subreddit and r/nba.

Data Scraping

PRAW: The Python Reddit API Wrapper

- Thomaz says this is better than using Pushshift API
- It made data scraping very simple, but requires a reddit account
- Data was pulled into lists and then placed into a dataframe

Data: I scraped posts, upvotes, # of comments however I ended up only using posts.

Modeling

Every model was piped using cvec and the chosen model, only the standard stop words were removed

Models: Logistic Regression, KNN Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier, Naive Bayes

Hindsight: I should have used cvec on my data first and created a dataframe. By piping and fitting to my model, I had to run my gridsearch everytime I wanted to rerun my model to update the jupyter instances.

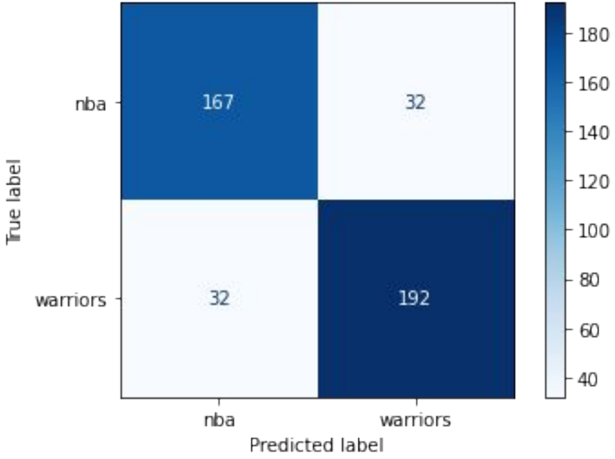
Findings - Logistic Regression

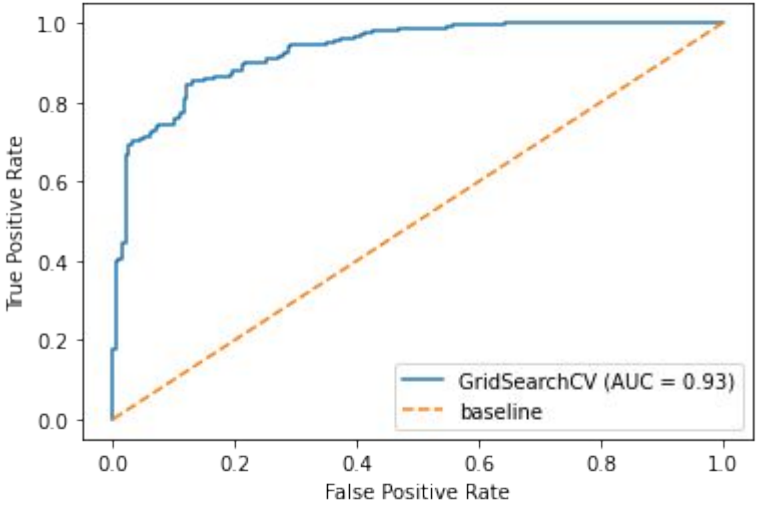
Sensitivity: 0.86325

Precision: 0.90179

Accuracy: 0.87234

ROC AUC: 0.93182





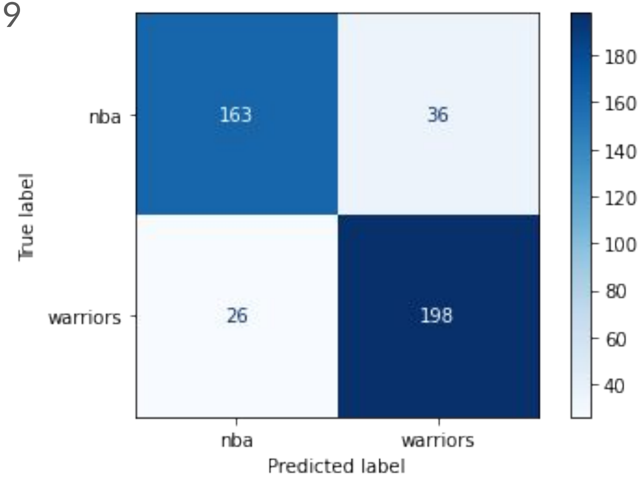
Findings - Naive Bayes

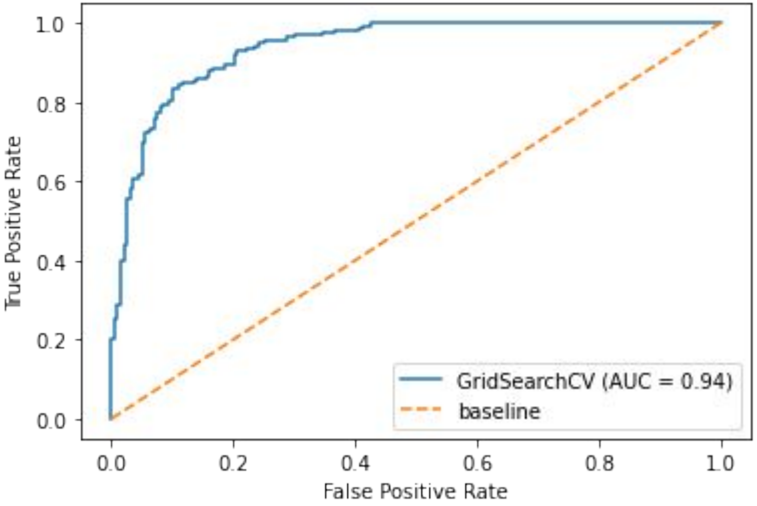
Sensitivity: 0.88936

Precision: 0.89316

Accuracy: 0.87943

ROC AUC: 0.93889





Findings

Both models performed about as well as each other. Naive Bayes performed slightly better on most metrics except for precision. However I selected Logistic regression as my model because precision was the most important metric.

False positives would lead to the posting unrelated content on the warriors subreddit, this would cause moderators to remove the posts and could possibly cause an identification of the bot, leading to moderators banning it.

Modeling Error

I realized that there would be false/false positives (posts that were predicted to be from r/warriors but were actually from r/nba, but they are posts that concern the warriors.) We want our model to select these posts so I located all false/false positives and changed them to be true positives. This increased the logistic regression precision slightly more than it did for Naive Bayes

Next Steps

Scrape more data from r/nba

These models need to be further tuned to maximize precision. Use of a Voting Classifier model could increase precision scores.

Train models for each NBA team’s subreddit; more posts = more karma

Try TFIDF Vectorizor to see if there is increased performance