Reddit Posts Classification

Colin Mondi

Agenda

01 Overview / Problem Statement

02 Process

03 Analysis Findings

05 Model Development / Results

06 Conclusion

Introduction

Problem Statement:
Subreddit feeds often include user posts that violate community norms, rules, and polices. Although there is a process in place to remove these types of posts, the turnaround time can be slower than expected giving more viewers time to read the content. To fix the issue Reddit is looking for a way to pre-determine which subreddits are most likely to contain these types of bad posts.

Introduction

Introduction:

The first step in the process is to determine which subreddit an individual post originated from using post content alone, leading to new discoveries regarding patterns/trends in topic discussion boards. This capability not only identifies bad content but can predict beforehand where it is most likely to occur. Specific subreddits can be closely monitored therefore increasing turnaround time in post removal. But going back to the first step, is there a way to even link individual posts to subreddits based on words alone?

For this project I chose two sub-reddits to test against, both with topics on strategies for fantasy sports. ***r/fantasyfootball*** includes content on fantasy football while ***r/dfsports*** includes content on daily fantasy challenges across all sports (not just football).

The goal is to predict which subreddit an individual post originated from only using post content alone.

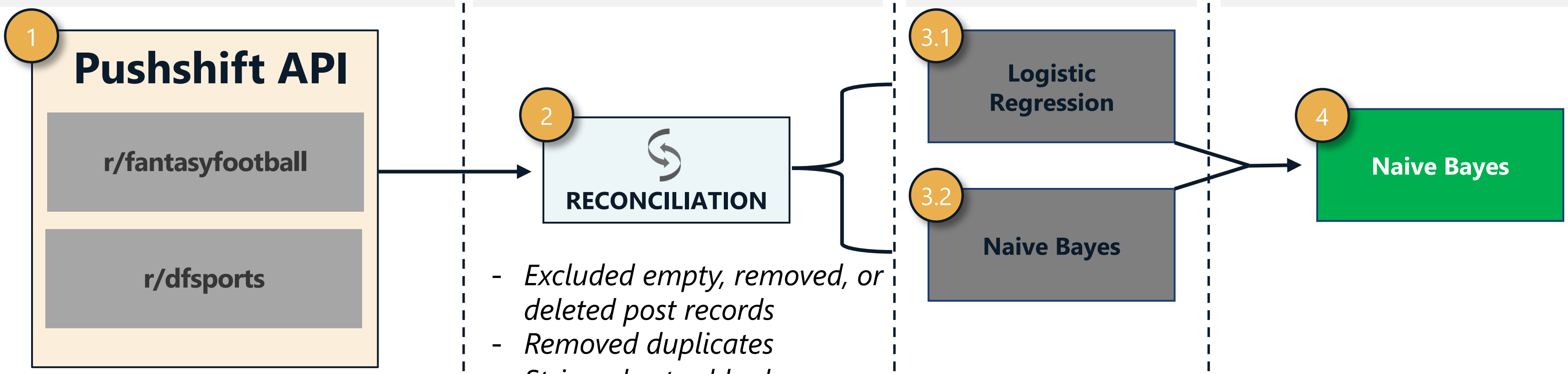
Process

Extracted subreddit data via API Call

Cleansed, intersected and analyzed combined data

Developed and tested two models

Compared model performance; determined best option



1 **Pushshift API**

r/fantasyfootball

r/dfsports

2 RECONCILIATION

3.1 Logistic Regression

3.2 Naive Bayes

4 Naive Bayes

Profiled independent data sets

- *Excluded empty, removed, or deleted post records*
- *Removed duplicates*
- *Stripped extra blanks*

- *Identified key patterns/trends*

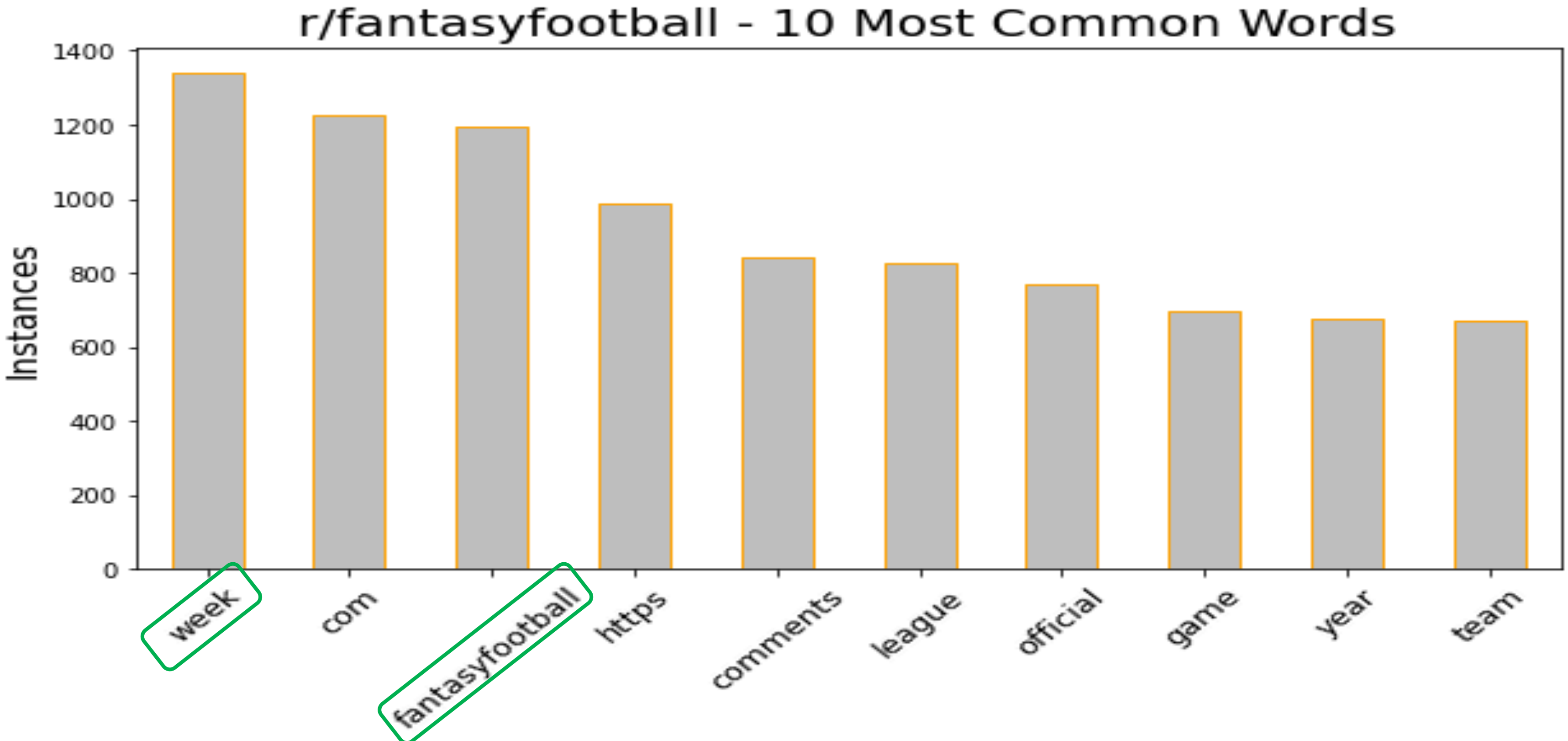
Post-Count:
- **5,000** r/fantasyfootball
- **5,000** r/dfsports

Post-Count:
- **3,912** r/fantasyfootball
- **1,696** r/dfsports

Analysis - r/fantasyfootball Posts

r/fantasyfootball is focused on fantasy football which is a *weekly* competition, while the majority of r/dfsports content focuses on *daily* fantasy challenges.

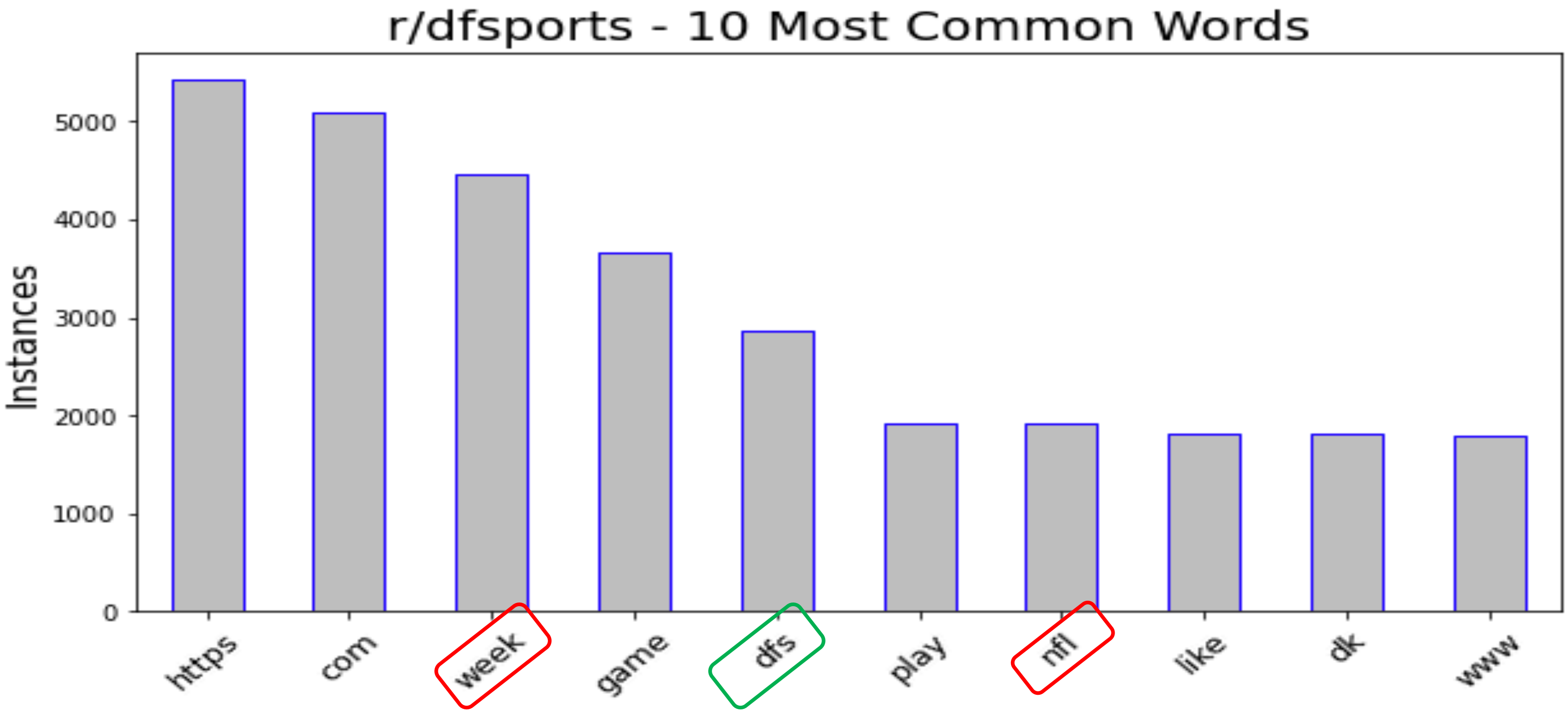
- The word "week" occurs most often in the r/fantasyfootball feed; this is a good sign regarding distinguishing subreddit posts (weekly versus daily)
- It's also good that the subbreddit name itself, "fantastfootball”, is the third highest occurring word



Analysis - r/dfsports Posts

r/dfsports includes content on multiple sports, while r/fantasyfootball includes content on football only, specifically the NFL.

- Given the word ”week" occurs often for r/fantasyfootball it was expected “day” (or similar representation) would occur often for r/dfsports… *but the results were the exact opposite*
- "nfl" (National Football League) is the seventh highest occurring word but no other other sports/league are listed here; another bad sign for distinguishing subreddit posts
- BUT it’s good that the subbreddit name itself, ”dfs”, is the fifth highest occurring word



Model - Development

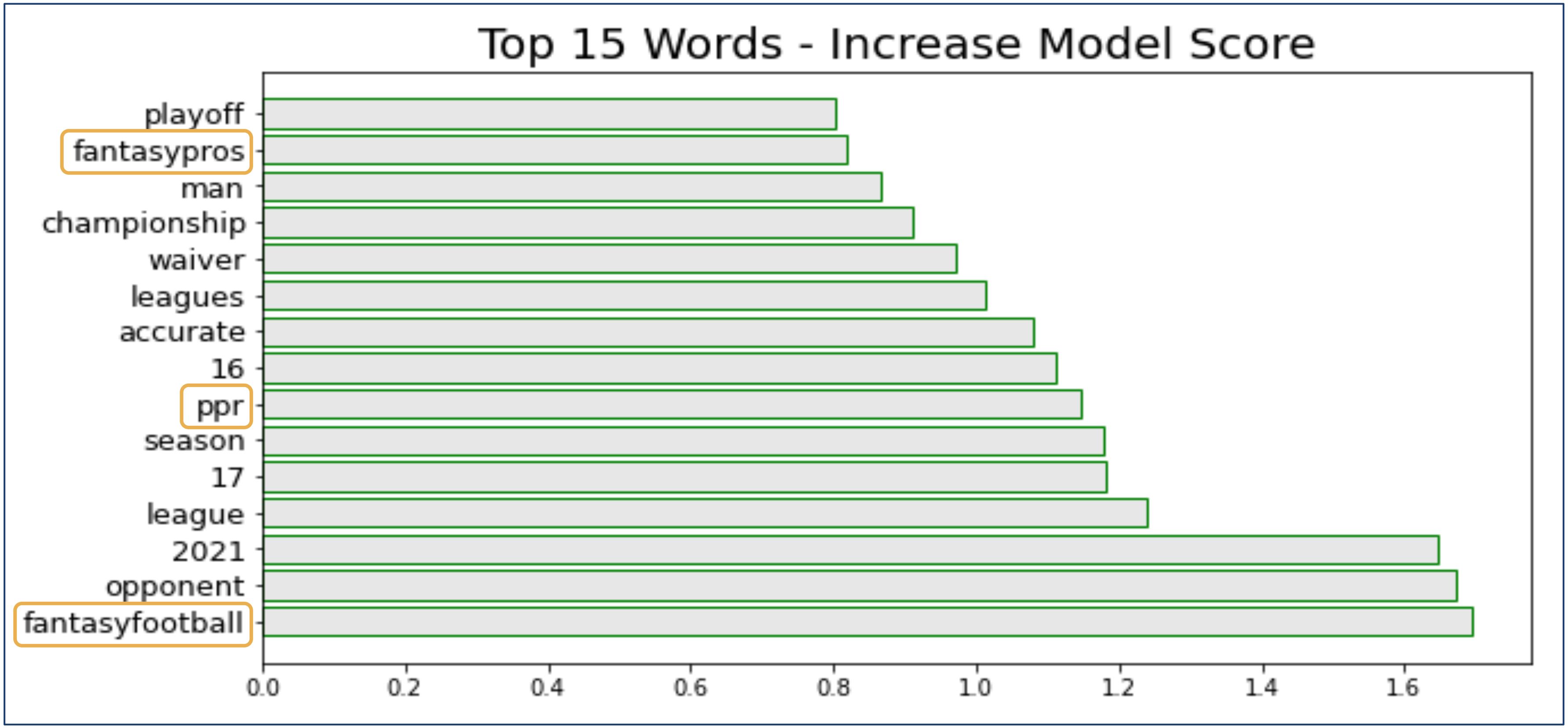
Logistic Regression - Naive Bayes

High-Level Process

- Generated a pipeline consisting of multiple components
  - ***Logistic Regression***: CountVectorizer, StandardScaler, & **LogisticRegression**
  - ***Naive Bayes***: CountVectorizer & **MultinomialNB**
- Calculated evaluation metrics including R-squared & RSME (root mean square error)
- Cross validated to evaluate pipeline accuracy
  - k-Fold classifier utilized to generate multiple groups of sample data
- Generated coefficient weights to determine the impact of individual words on model performance
- Created confusion matrix to compare actual versus predicted results

Model – Coefficient Weights

The following individual words improved the performance of the model -



- **'fantasyfootball'** – this word occurs often across posts in the r/fantasyfootball subreddit, but almost no cases of appearance in r/dfsports posts
- **'fantasypros'** – this word reflects a web page (fantasypros.com) that's another source fantasy football participants refer to for expert advice
- **'ppr'** – an acronym that stands for "points per reception"; a type of fantasy football scoring method where a player earns one additional point per reception made

Model - Confusion Matrix

Summary

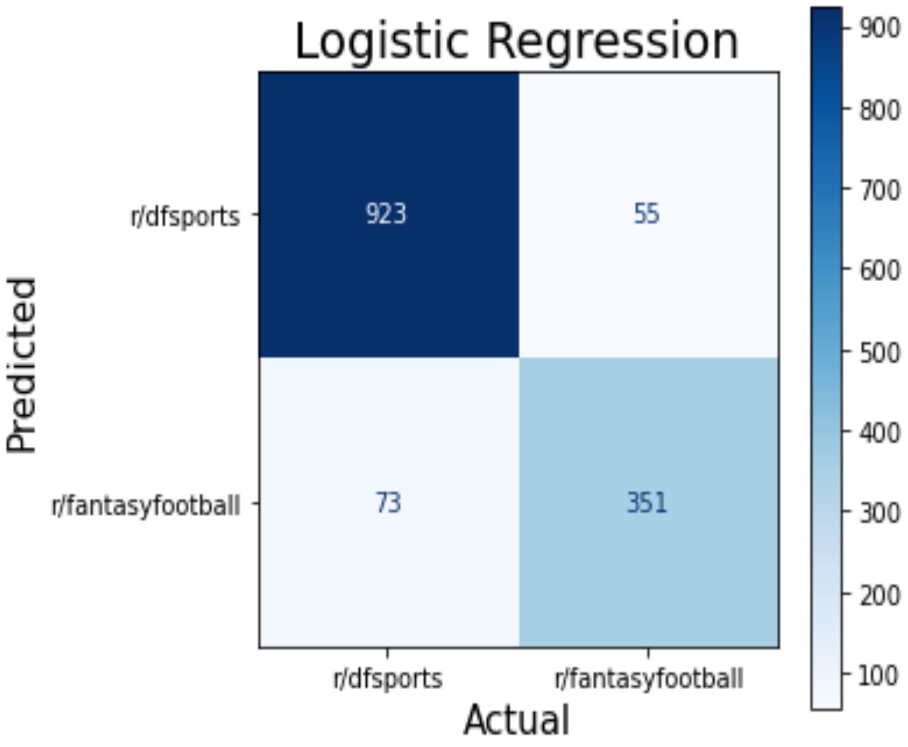
Both models had a higher accuracy rate for r/dfsports (around ~90%) in comparison to r/fantasyfootball (around ~85%).

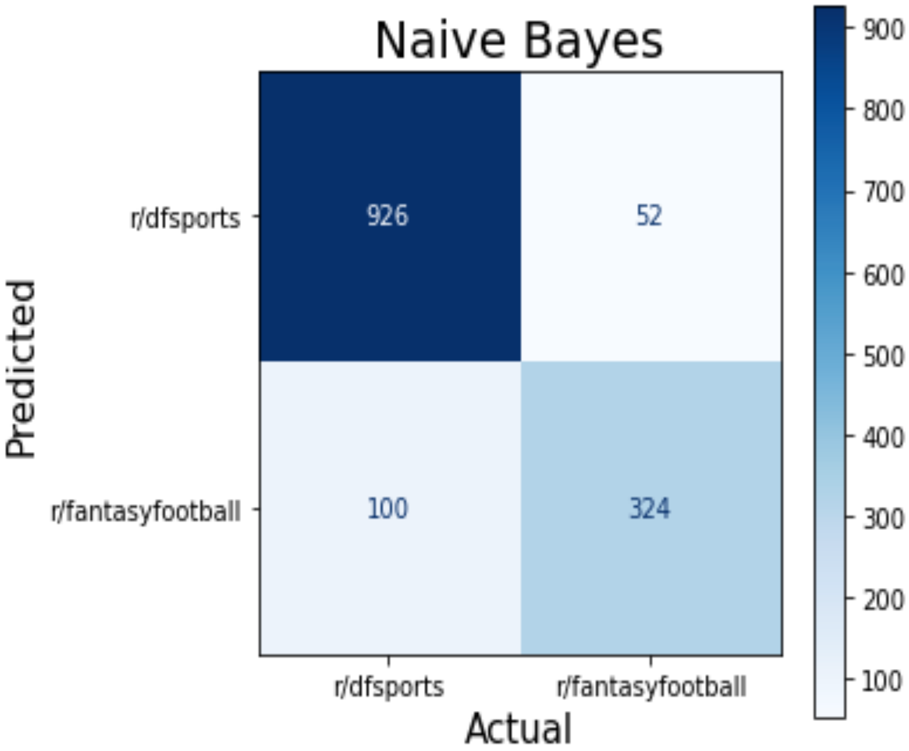
Logistic Regression

- ***dfsports:*** 92% correct, 8% incorrect
- ***fantasyfootball:*** 84% correct, 16% incorrect
- 8% difference in % correct comparing subreddits

Naive Bayes

- ***dfsports:*** 91% correct, 10% incorrect
- ***fantasyfootball:*** 87% correct, 13% incorrect
- 4% difference in % correct, comparing subreddits





Model - Results

Logistic Regression

- **R^2 score - test:** ~0.90
- **R^2 score - train:** ~0.96
- **k-Fold Cross Val score:** ~0.86
- **RMSE:** 0.31
- **Confusion Matrix – Accuracy Rate:**
  - *r/fantasyfootball*: 84%
  - *r/dfsports*: 92%

Naive Bayes

- **R^2 score - test:** ~0.89
- **R^2 score - train:** ~0.89
- **k-Fold Cross Val Score:** ~0.89
- **RMSE:** 0.32
- **Accuracy:** 0.89
- **Sensitivity:** 0.76
- **Specificity:** 0.94
- **Confusion Matrix – Accuracy Rate:**
  - *r/fantasyfootball*: 87%
  - *r/dfsports*: 91%

Summary Findings

- Logistic Regression has slightly better **R^2** (0.90 vs 0.89) and **RMSE** (0.31 vs 0.32) values
- Logistic Regression is overfit due to the higher train score (0.96) in comparison to both the test score (0.90) and k-Fold cross val score (0.86)
- Naive Bayes has almost identical scores for all three (0.89) indicating the model is fit fairly well
- Logistic Regression has a better accuracy rate for predicting r/dfsports posts (92% vs 91%)
- Naive Bayes has a better accuracy rate for predicting r/fantasyfootball posts (87% vs 84%)
- Naive Bayes has a more balanced test result regarding accuracy rate for predicting both subreddit posts

Conclusions

Summary:
Both models successfully predicted the correct subreddit a post originated from for majority of cases.

Recommendation:
Although the Logistic Regression model has slightly better R-squared score and RMSE value, the difference in both metrics is too close to draw any conclusions. But the train, test, and k-Fold cross val scores for the Logistic Regression model indicate overfitting, while this is not the case for the Naive Bayes model. Given this, I believe the Naive Bayes model is the better option for the task at hand.