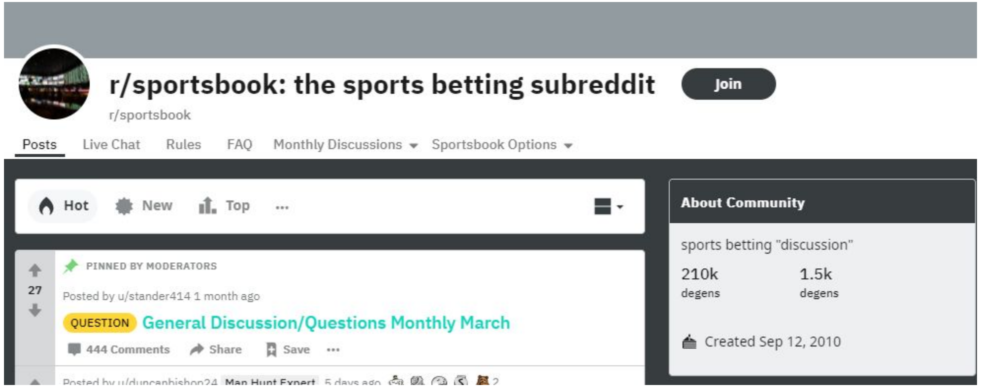
Project 3 - Web APIs and NLP

DSIR-22221
Patrick Cudo

Problem Statement:

This project explores comments collected from Reddit.com to predict what particular subreddit those comments were collected from. Through this process we can gather some insight on what words to focus on in our predictions. Once these words are identified they can be used for further analysis.

Chosen Subreddits



r/sportsbook: the sports betting subreddit

r/sportsbook

PINNED BY MODERATORS

27

Posted by u/stander414 1 month ago

QUESTION General Discussion/Questions Monthly March

444 Comments

Posted by u/duncanbishop24 Man Hunt Expert 5 days ago

About Community

sports betting "discussion"

210k	1.5k
degens	degens

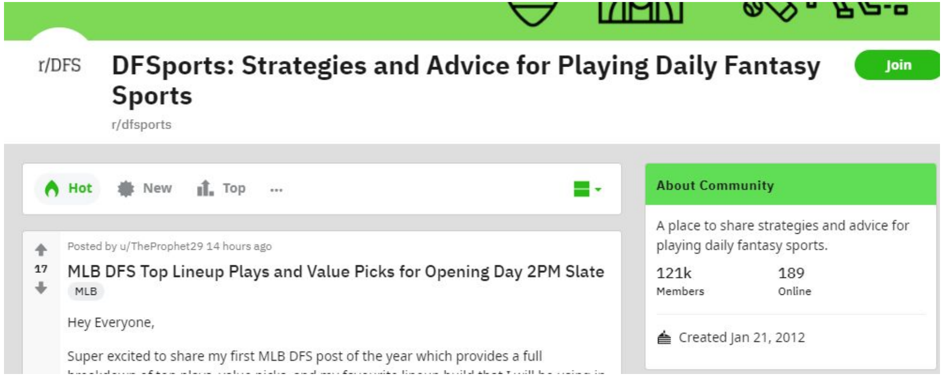
Created Sep 12, 2010

Sportsbook

- The sports betting subreddit
- 210,000 members

DFSports

- Strategies and advice for playing daily fantasy sports
- 121,000 members



r/DFS DFSports: Strategies and Advice for Playing Daily Fantasy Sports

r/dfsports

17

Posted by u/TheProphet29 14 hours ago

MLB DFS Top Lineup Plays and Value Picks for Opening Day 2PM Slate

MLB

Hey Everyone,

Super excited to share my first MLB DFS post of the year which provides a full

About Community

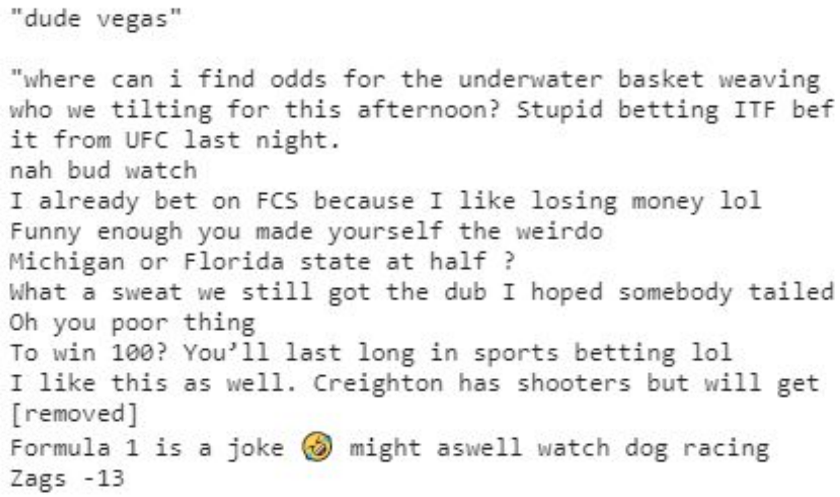
A place to share strategies and advice for playing daily fantasy sports.

121k	189
Members	Online

Created Jan 21, 2012

Data Collection, Cleaning and EDA

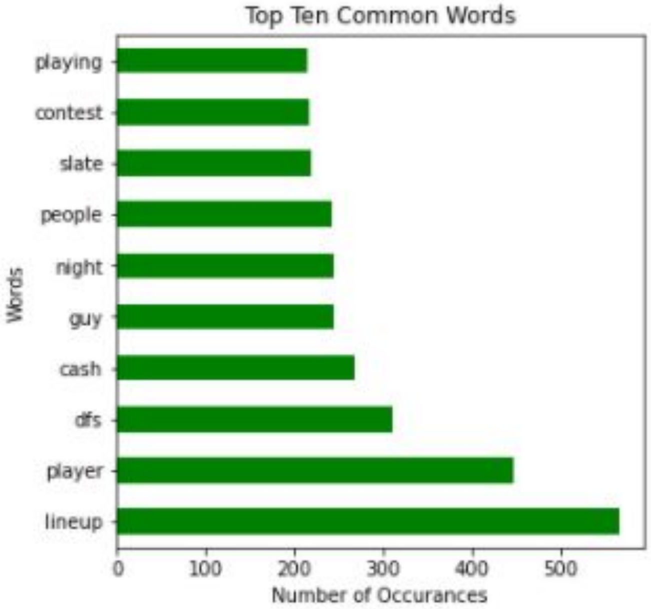
- Data collected using pushshifts' API - used to search for Reddit comments on chosen subreddits
  - 6,400 total comments; 3,200 for each subreddits
  - Used automated python file to pull
- With these two subreddits, users did not submit many images, lots of text
  - Dropped all [removed] comments
  - Removed links, not entire post with link
- Example of text collected shown on right.



```
"dude vegas"

"where can i find odds for the underwater basket weaving
who we tilting for this afternoon? Stupid betting ITF bef
it from UFC last night.
nah bud watch
I already bet on FCS because I like losing money lol
Funny enough you made yourself the weirdo
Michigan or Florida state at half ?
What a sweat we still got the dub I hoped somebody tailed
Oh you poor thing
To win 100? You'll last long in sports betting lol
I like this as well. Creighton has shooters but will get
[removed]
Formula 1 is a joke 🤣 might aswell watch dog racing
Zags -13
```

Preprocessing: Stop words, lemmatize and stem

- Sklearn ‘english stop words used as base added then added to it.
- Took common most words in both subreddits and found common words found in both subreddits to add to base stop words.
- Once stop words list was finalized, create .csv file for both lemmatized and stemmed words

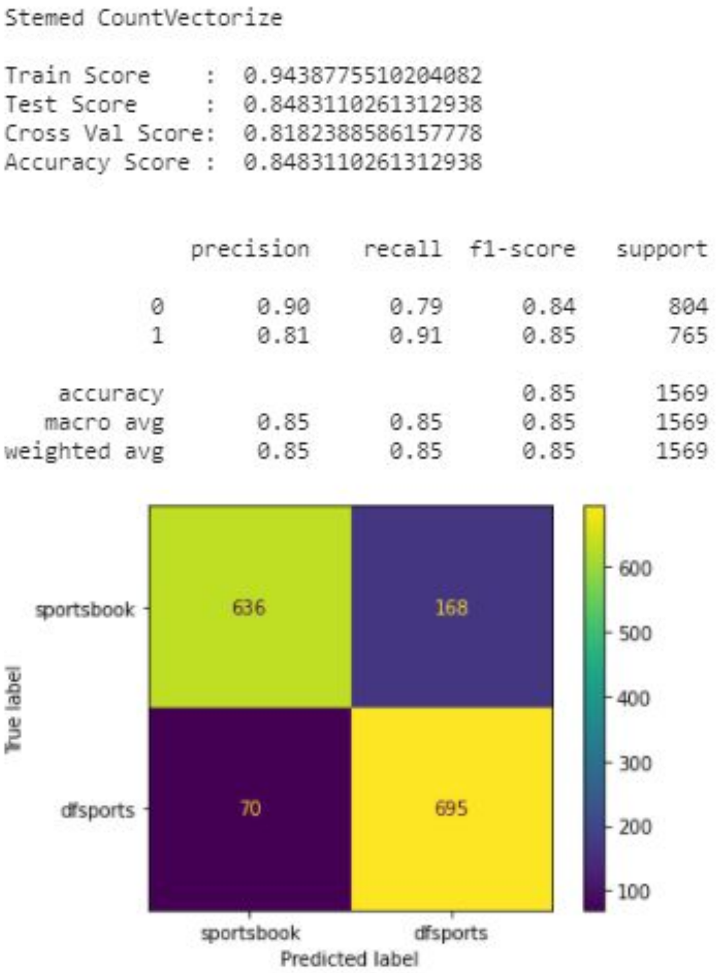


Model Performance

Base Model

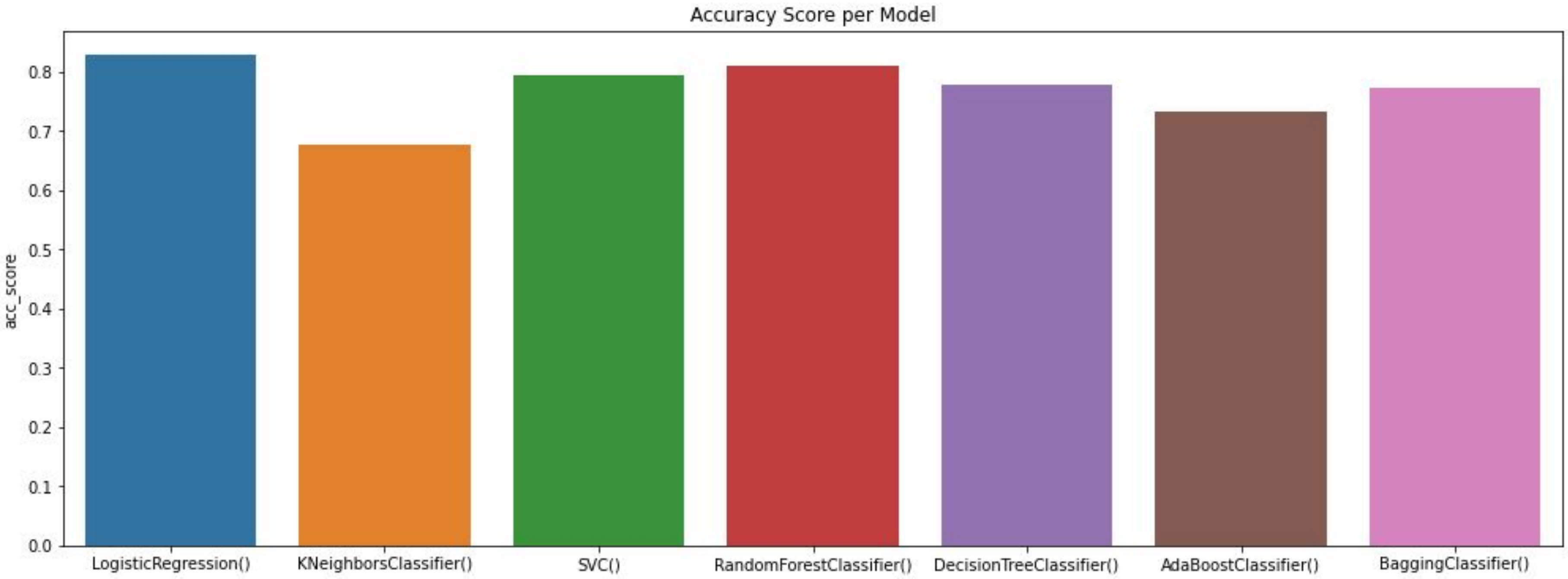
- Transformers: PorterStemmer, CountVectorizer
- Classifier: Logistic Regression
- Params: Default

Model was overfit and only default params were used. In order to increase regularization gridsearch was used to find a decreased C value as well as effective penalty.



```
Stemed CountVectorize

Train Score     :  0.9438775510204082
Test Score      :  0.8483110261312938
Cross Val Score:  0.8182388586157778
Accuracy Score :  0.8483110261312938

              precision    recall  f1-score   support

           0       0.90      0.79      0.84       804
           1       0.81      0.91      0.85       765

    accuracy                           0.85      1569
   macro avg       0.85      0.85      0.85      1569
weighted avg       0.85      0.85      0.85      1569
```

Model Performance

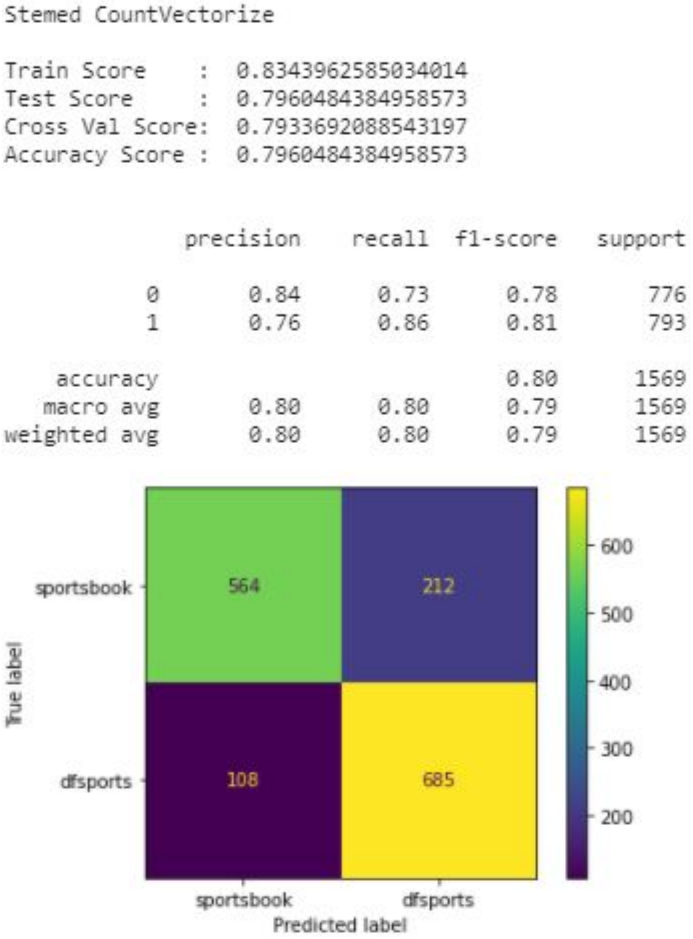


Model Performance

GridSearchCV Model

- Transformers: PorterStemmer, CountVectorizer
- Classifier: Logistic Regression
- Params:
  - CountVectorizer:
    - max_features=500
  - Logistic Regression:
    - C = .375
    - Max_iter = 1000,
    - Penalty = l2
    - solver = liblinear

Adjusting C value and selecting Ridge penalty help increase regularization and decrease the overfit previous model. At a cost, model is not very good decrease in accuracy and increase in number of negatives.



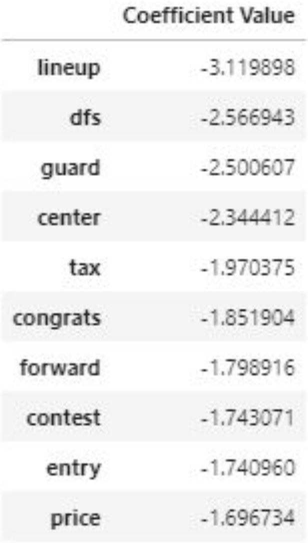
```
Stemed CountVectorize

Train Score      :  0.8343962585034014
Test Score       :  0.7960484384958573
Cross Val Score:  0.7933692088543197
Accuracy Score  :  0.7960484384958573


              precision    recall  f1-score   support

           0       0.84      0.73      0.78       776
           1       0.76      0.86      0.81       793

    accuracy                           0.80      1569
   macro avg       0.80      0.80      0.79      1569
weighted avg       0.80      0.80      0.79      1569
```

Conclusions

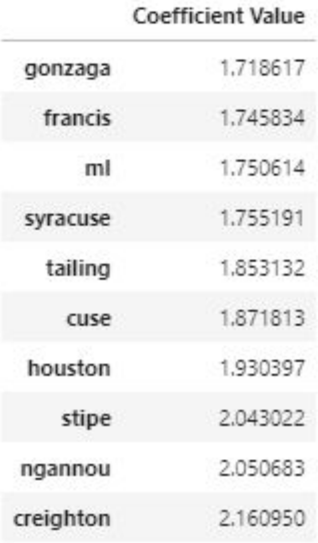
Top ten words for each subreddit are shown on the right.

Negative coefficients have strongest correlation to dfsports and positive coefficients have strongest correlation to sportsbook.

Interesting contrast between two subreddits, both gamblers?



	Coefficient Value
lineup	-3.119898
dfs	-2.566943
guard	-2.500607
center	-2.344412
tax	-1.970375
congrats	-1.851904
forward	-1.798916
contest	-1.743071
entry	-1.740960
price	-1.696734



	Coefficient Value
gonzaga	1.718617
francis	1.745834
ml	1.750614
syracuse	1.755191
tailing	1.853132
cuse	1.871813
houston	1.930397
stipe	2.043022
ngannou	2.050683
creighton	2.160950