

Project 3

Web APIs and NLP

By Tiffany Houston



Overview

- Problem Statement
- Data Gathering
- Cleaning / EDA
- Modeling Process
- Key Findings
- Conclusions



What's the
problem?



Hunting and Gathering

- Find “significant” forum
 - Imported subreddits via UTC ID number
 - 5000 posts(rows) pulled
 - ~90 total columns combined
-

Cleaning / EDA

Created one dataframe for all posts

Dropped duplicates and null values

Replaced whitespace with “_”

Got rid of insignificant data



Pre-Modeling:

Made Subreddit column binary

Defined X and y variables

Baseline prediction:

$$0 \text{ (r/NBA)} = 47.3\%$$

Multinomial Naive Bayes

Training Score Prediction: 72.7 %

Testing Score Prediction: 68.9 %

Logistic Regression with CountVectorizer

Training Score Prediction: 76.7 %

Testing Score Prediction: 70.2 %

Logistic Regression with TFIDF Vectorizer

Training Score Prediction: 86.7 %

Testing Score Prediction: 72.5 %

Random Forest

Training Score Prediction: 99.6 %

Testing Score Prediction: 71.8 %

Score Summaries / Key Findings

Multinomial Naive Bayes

Testing Score Prediction:

68.9 %

vs

Logistic Regression w/ TFIDF

Testing Score Prediction:

72.5 %

- Overfit initial models
- The machine does **NOT** know best!



Conclusion / Summaries



2



1



Thank you!