Reddit Posts: What Makes Them Popular?

MARIO ARTHUR-BENTIL

FIVETHIRTYEIGHT

Data Collection and Analysis

- 3000 unique posts
- Variables for determining popularity: title, subreddit, number of comments, time, domain, score
- Used BeautifulSoup for web scraping
- Median number of comments: 22

Data Modeling

- Comment magnitude: High vs Low (based on median)
- Baseline Accuracy: 50%
- First Model: Random Forest
- All variables used: 99.1%
- Best solo variable indicator: Score (67.3%)
- Popular news topics ('Black Panther' and gun control) made less of an impact: (50.7% and 50.5% respectfully)

Data Modeling

- CountVectorizer (most popular words): ‘yes’, ‘gif’, ‘fraud’
- RandomForest Score: 50.9%
- Second model: KNN
- KNN Score: 50.6%

Conclusion and Suggestions

- A model using multiple variables works best, but if we were to use only one it would be ‘Score’
- Random Forest model performs better than KNN in this case
- Should try out other classification models (e.g. SVM)