Project 3: Web Scraping and Classification

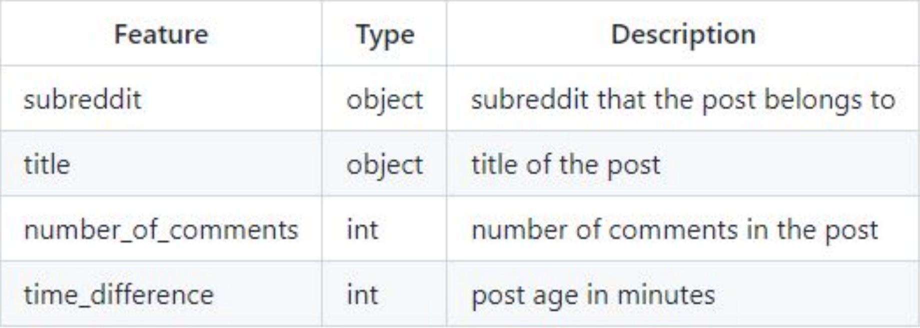
Francisco Trejo
June 7, 2022

Problem Statement

We are looking into breaking into the world of freelance data journalism and are reaching out to Nate Silver and co. at FiveThirtyEight so they can hear our pitch on how to create a Reddit post that will get the most engagement. We want to find out what characteristics of a post on Reddit will be the most predictive of the overall interaction on a post as measured by number of comments (above/below the median). With this we hope to provide a classification model to FiveThirtyEight that is satisfactory and jumpstart our career!

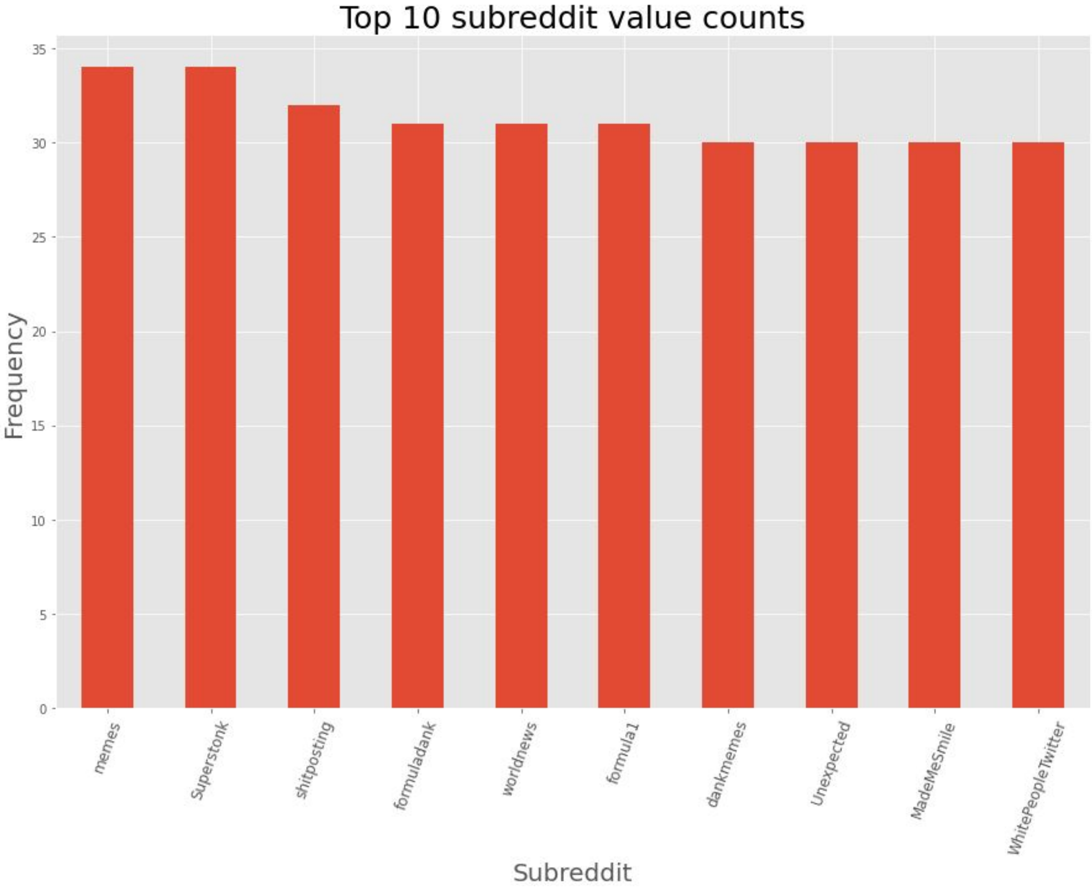
The Data

- Post on r/all that are currently “hot”
- Scraped 20,000 posts from 5 different days
- Our features were subreddit, title, and time_difference (age of post)
- above/below number of comments was our target



Feature	Type	Description
subreddit	object	subreddit that the post belongs to
title	object	title of the post
number_of_comments	int	number of comments in the post
time_difference	int	post age in minutes

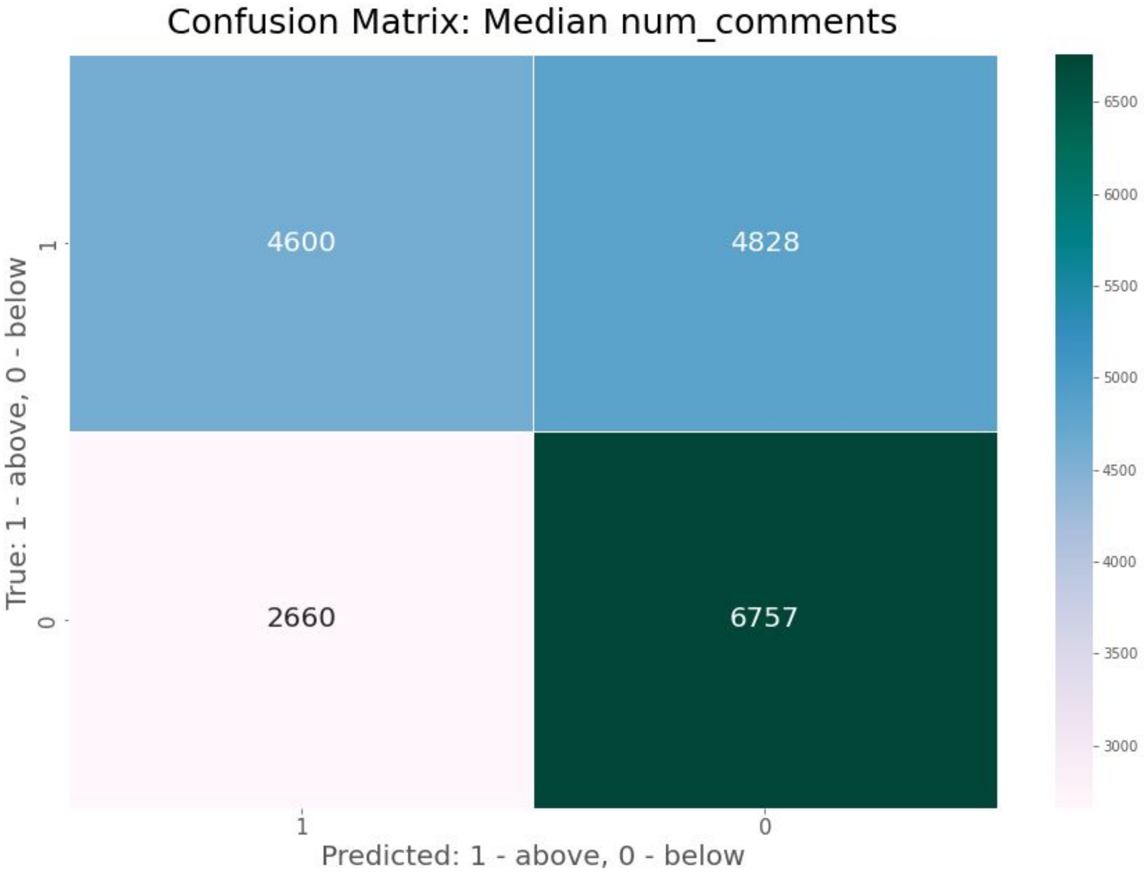
Subreddits



- Performed EDA on features
- We had 3810 unique subreddits in our data
- Feature engineering

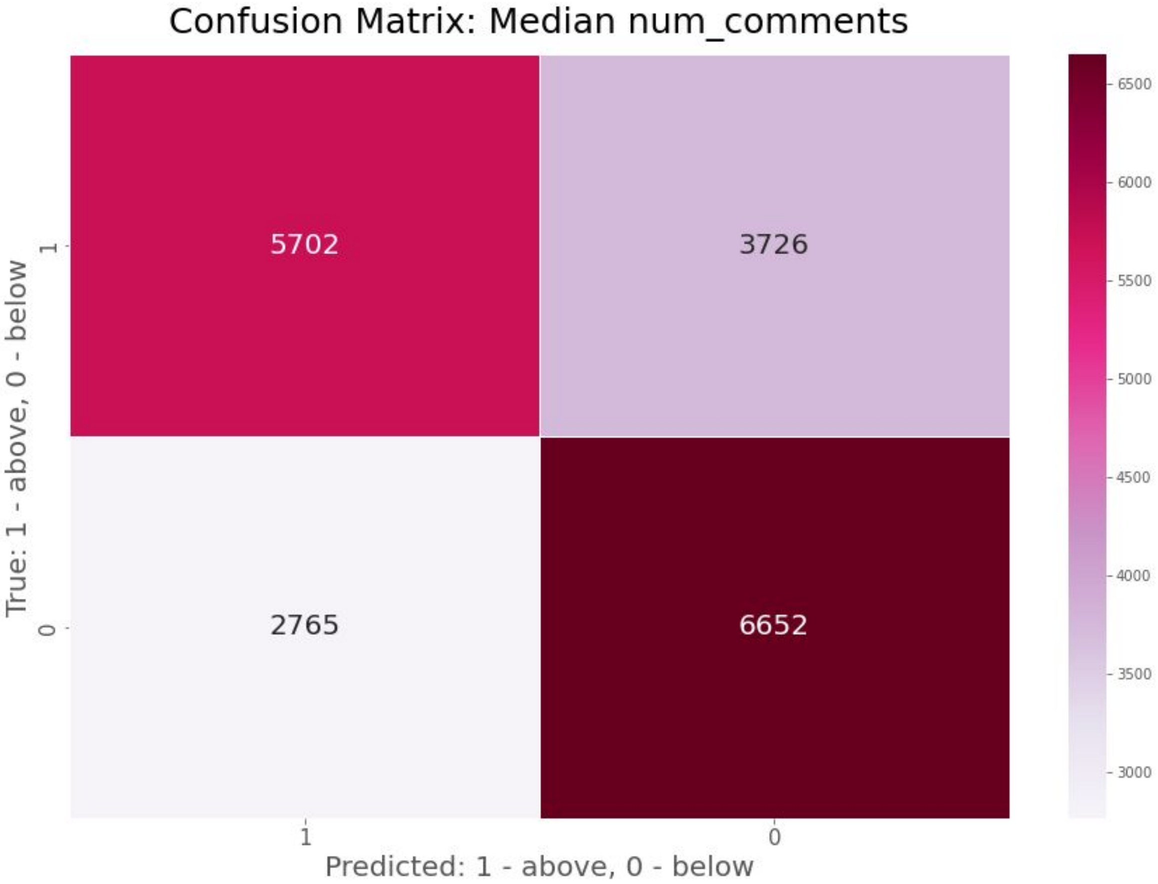
Modeling - KNN

- Post titles were lemmatized then vectorized
- Lemmatized KNN model accuracy was 60%
- Confidence interval of 0.6016 ± 0.0326



Modeling - Random Forests

- Post titles were lemmatized then vectorized
- Random Forest model accuracy was 66.68%
- Confidence interval of 0.6556 ± 0.0120



Confusion Matrix: Median num_comments

Conclusion/Recommendations

Our model identified the following 5 features as the most important: time_difference, subreddit_appears_more_than_once, subreddit_in_top_10, number, and year. We discovered that for our "hot" and “not-hot” posts average post age was 10 and 7 hours respectively. We can gauge that when writing a post it takes time to get engagement but if our post is not "hot" after 10 hours we may want to reconsider our post. For the following 2 features, subreddit_appears_more_than_once and subreddit_in_top_10, we want our post to be from a subreddit that fits these parameters. The last 2 features, number and year, are tokens/words that appeared in our posts. Number is a tokenized word for any post that contained numbers so we want to make sure that our post has numbers in it to better increase our chances of a high interaction post.

We believe that our finding can benefit those at FiveThirtyEight but we do recommend a substantial amount of additional data be collected to better improve our model. Reddit has about 48 million active monthly users so new posts are made constantly. Collecting additional post attributes is also recommended, there may be additional attributes that our model can use to show potentially better predictive power than the features that were collected.