Making the News on r/news:

Predicting story engagement on the front page of the internet

by

Casey La Honta

Table of Contents

1. Introduction

Within the last ten years, social media has revolutionized the flow of information among users all over the world. One of the most popular social media services, Reddit, allows users to interact on various boards covering a wide range of topics. The r/news board is one of the most popular on the site, with around 23.4 million subscribers. The page averages over 30,000 weekly users and often draws over 10,000 comments per day.

Despite the heavy flow of information and news, it is difficult to know exactly which stories will engage users. Since users click into news stories via headlines on the board, the information conveyed in the headline is invariably important to driving this engagement. For news providers and services looking to drive traffic to their own sites, grabbing the attention of users is key.

The intended use of this model would be for news services to optimize their headlines in order to drive engagement from Reddit's millions of users. Reddit users will also benefit from viewing more relevant information, and Reddit itself could certainly use the model to not only improve user experience but also to drive traffic to their own site.

2. Dataset Information

The data set for this project comes from a web scraper known as the Python Reddit API Wrapper (PRAW), which retrieved 232,806 stories from the first six months of 2021. Data for these stories included identifiers such as Post ID and URL, as well as engagement measures such as Score and Total No. of Comments.

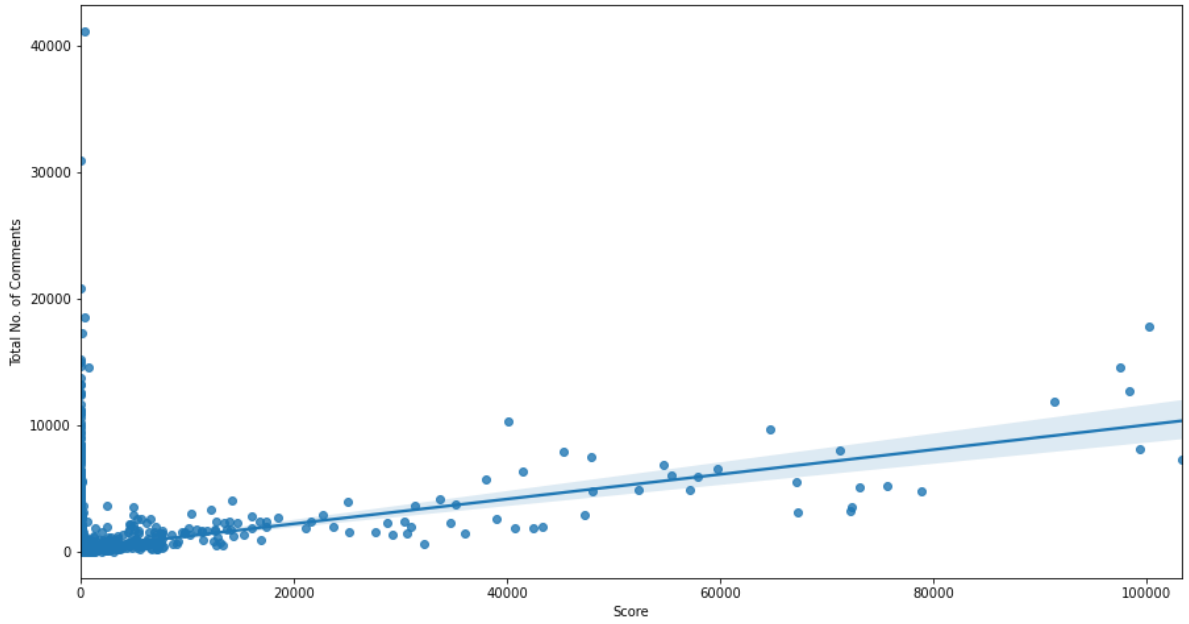
A feature known as `Engagement` was created in order to properly perform this project as a classification, which is detailed in the next section. By setting a cut point of 10 for this score, a data set of 11,240 stories was produced which would be used for the exploratory data analysis and model building. Additionally, setting this cut point produced a higher quality dataset that no longer included advertisements or completely irrelevant stories.

3. Data Wrangling

Since the headlines were gathered using Reddit's API, there were fewer features that required cleaning up to prepare the data for analysis. For example, the features were the correct data type and did not contain any null values. The dataset did require the removal of duplicates (2108 stories), and the creation of an engagement score.

For this project, our target variable for prediction was at first vaguely defined as engagement. Two features present in the data set represented different aspects of user engagement: `Score` and `Total No. of Comments`. The former represents the aggregate of all upvotes (+1) and downvotes (-1) on a Reddit story, and is a key metric when understanding user interface with the Reddit site. Reddit's boards are configured for high-scoring stories to remain higher up in the feed, while lower-scoring stories instead drop off of the page. Therefore, a high score extends the visibility and time for interaction by users, and often leads

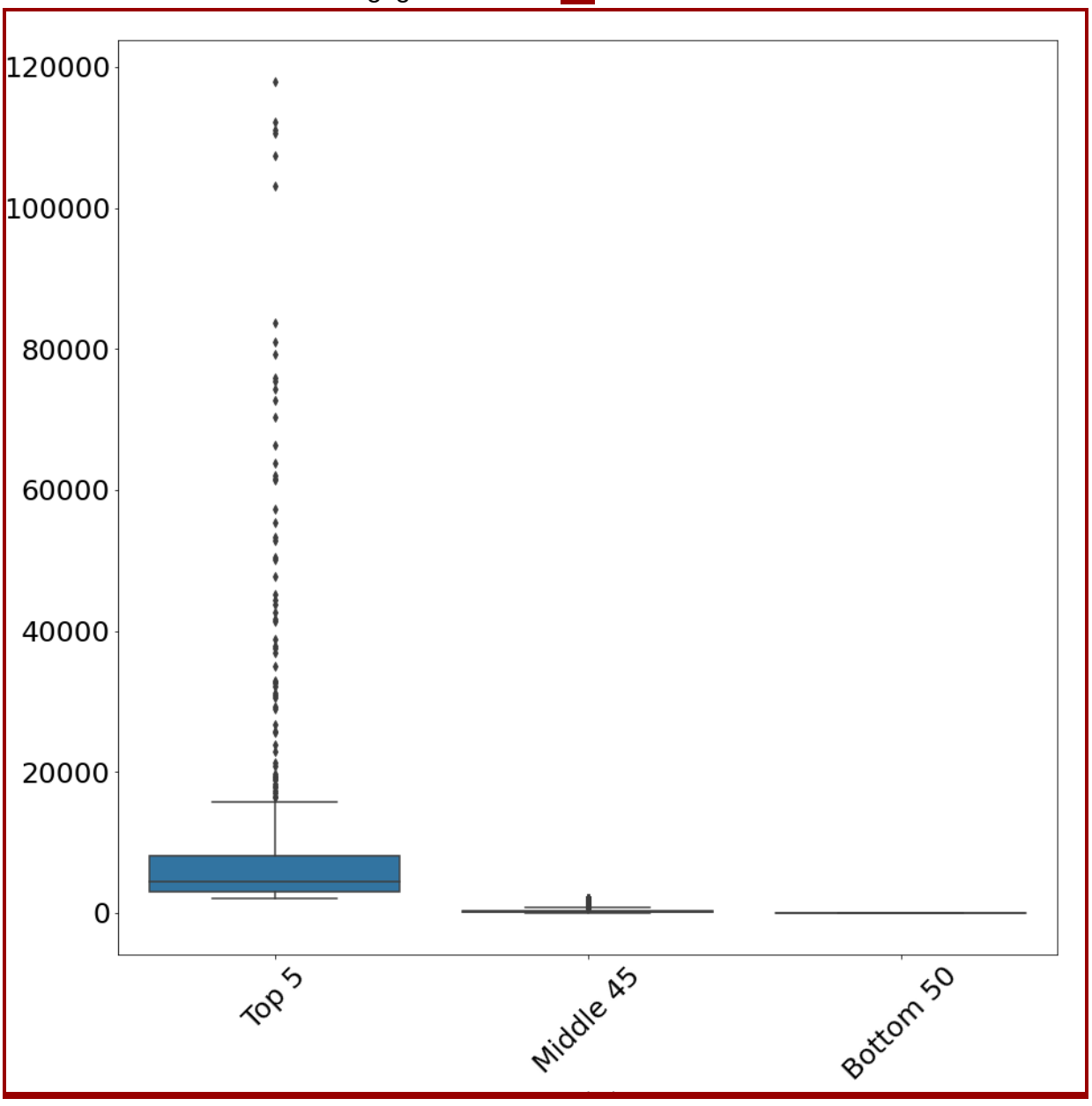
to higher engagement. Theoretically, this would lead to a correlation between the score and the number of comments, if both are accurate representations of user engagement:



So much for that theory. This low correlation indicates that there is only a weak relationship between score and number of comments. This weak relationship necessitated the adding up of score with number of comments to create a new metric known as `Engagement`, which would allow the developed model to focus on a simple target variable.

4. Exploratory Data Analysis

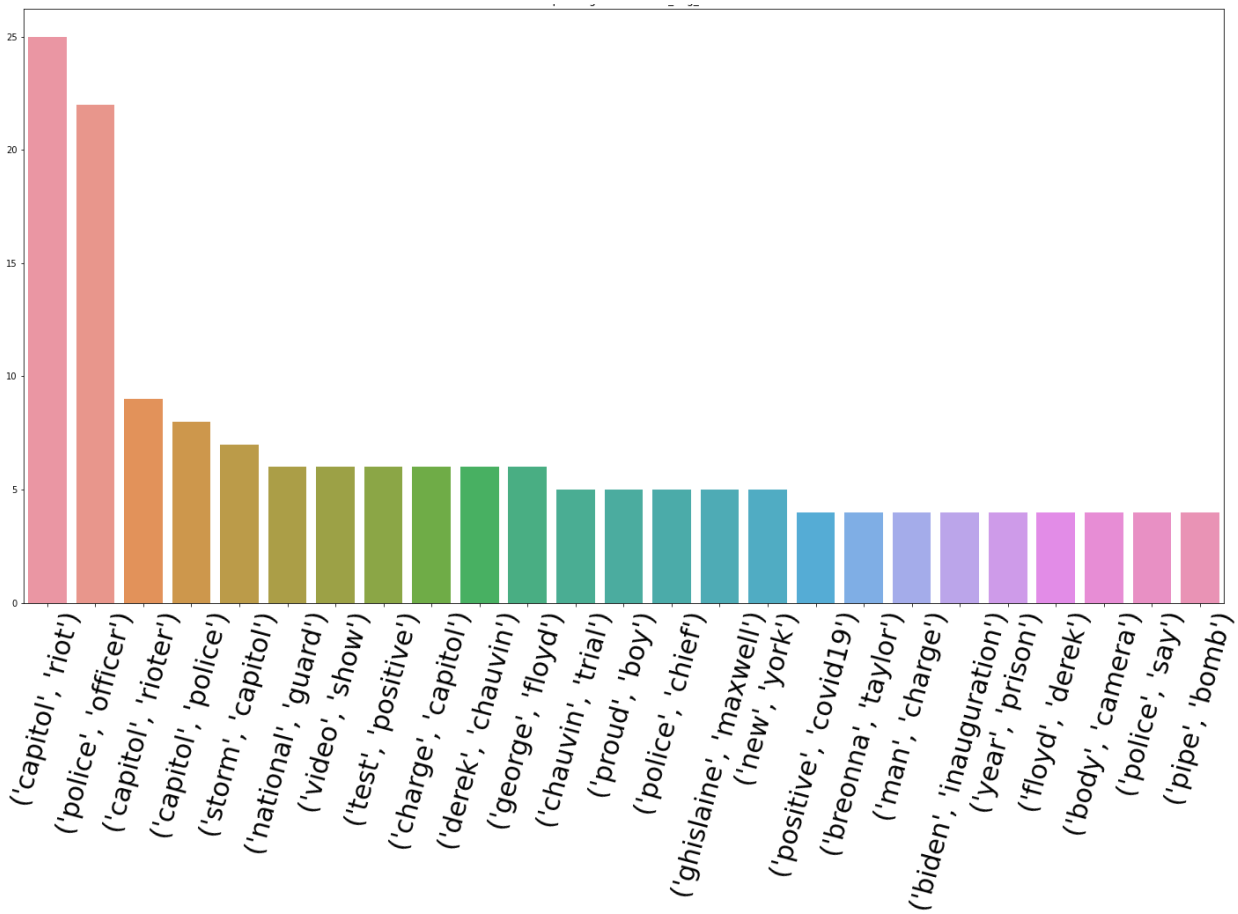
For this project, our target variable was the `Engagement` metric created from `Score` and `Total No. of Comments`. News stories were then banded into Top 5%, Middle 45%, and Bottom 50% in terms of Engagement score:



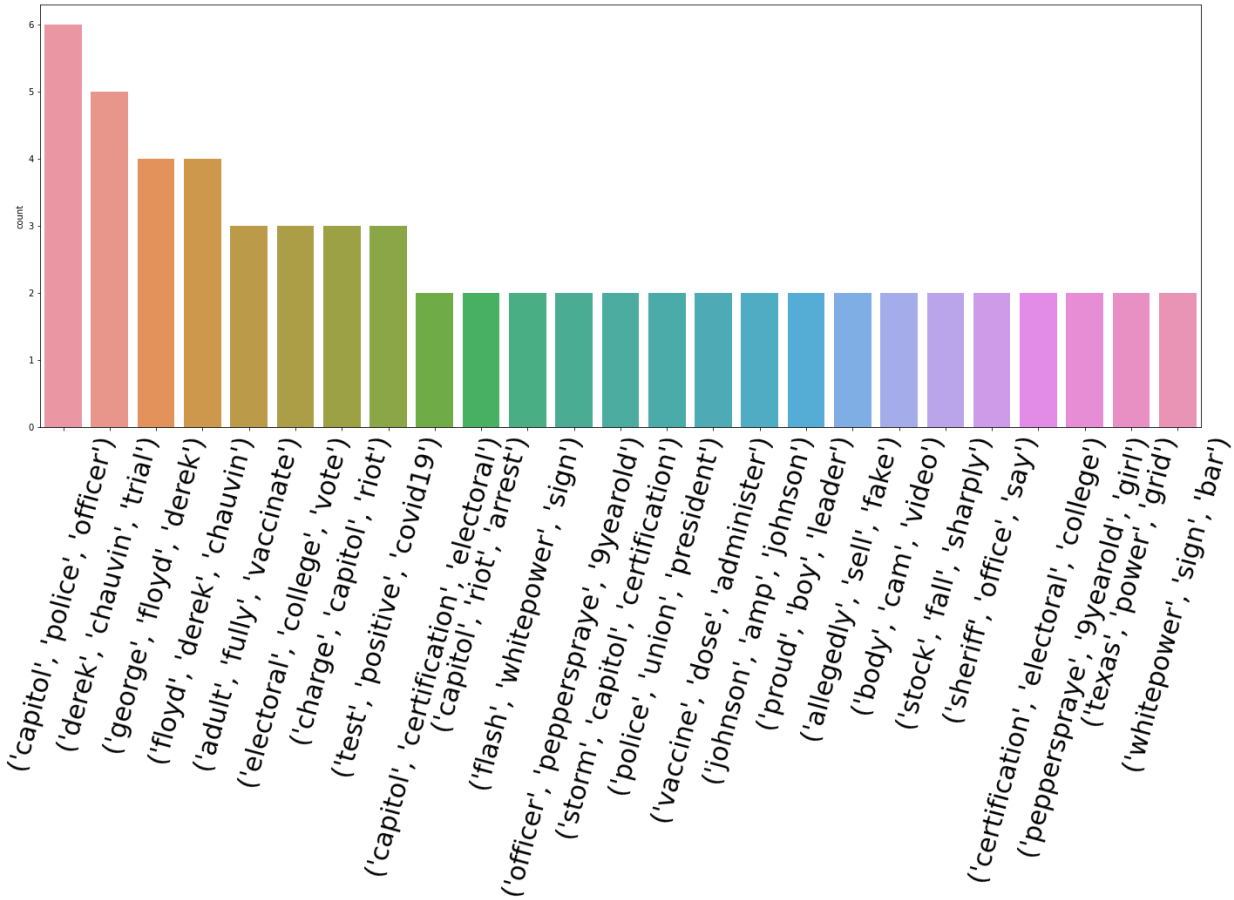
These bands were picked in order to balance two different needs for this project. On the one hand, the target population needed to have a larger number of cases in order to build a better-quality model, meaning a smaller top band (such as the top 1% of stories) would consist of too few stories. On the other hand, the populations needed to be distant enough from each other for the binary classification to be meaningful. The larger the top band (for instance, if it

were the top 10% of stories), the closer in value the two bands would become, making the development of a classifier more difficult.

Next step in the exploratory analysis was to review which terms occurred frequently in the Top 5 band. The purpose of this exercise was to get an idea of which topics were more likely to drive user engagement. The following represent frequencies of two-word sequences:



And three-word sequences:



This analysis does indicate that certain storylines are engaged with more often than others. This is useful information, as the objective of this project is to construct a model that will help to predict levels of engagement. However, it will be important to not overfit the model on particular storylines that are unlikely to repeat in later times. For example, the word sequences from this analysis identify stories such as the January 6th riot and the Derek Chauvin trial as high-engagement stories. While these stories are notable in the time period of January-May 2021, they are unlikely to provide predictive power for future headlines. These are important considerations when preparing the data for analysis.

5. Pre-processing

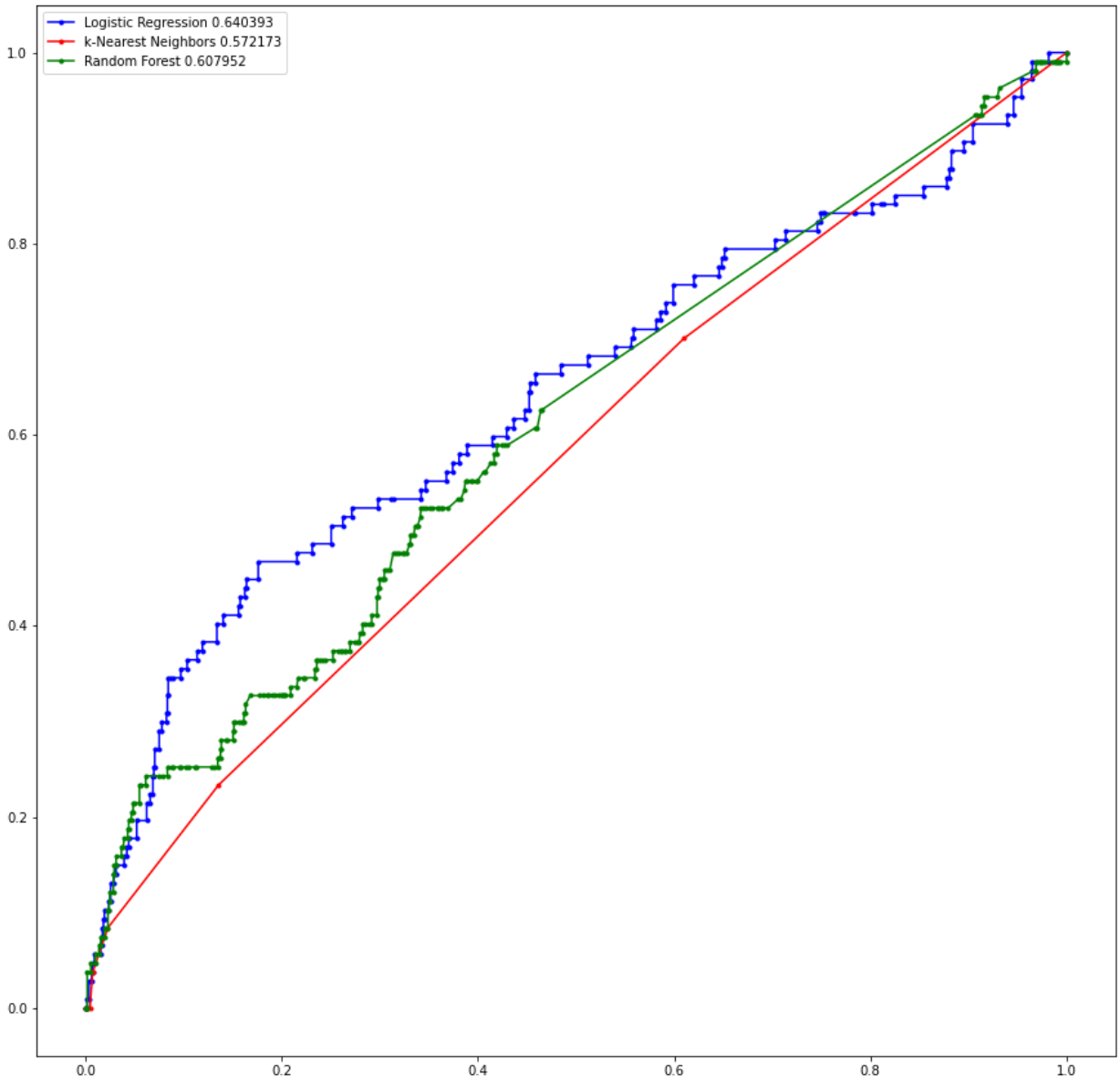
There were a number of steps taken to prepare the data. First, the target variable had to be converted into a binary classification, as this would allow for proper model development. To achieve this, the Top 5 (*n* = 543) and Bottom 50 (*n* = 5293) bands were converted into 1 and 0, respectively. The Middle 45 band was dropped from the data set.

Various transformations were performed on the text data, in order to ensure that the features would produce the best model. Each headline was first converted into a list of case-corrected, lemmatized English words. Lemmatization refers to the process by which a word is converted into a single item, to control for instances where there are many forms of the same word. As one example, many conjugates of a verb ("elects," "elected," "electing") would be converted into their base form ("elect"). The features were coded as a 1 for the presence of the word in the headline, or a 0 for the absence of it.

A tool called Count Vectorizer was then used to separate the lists of headline words into distinct features. To avoid producing an unworkable dataset, only features present in 5 or more headlines were retained for the model building. These processing steps led to a final data set of 5836 headlines with 2534 distinct features.

With the appropriate features selected and the target properly coded, I began the model selection. For each of the classification algorithms, I performed sklearn's GridSearchCV to determine the best hyperparameters. I fit and classified using Logistic Regression, Random Forest, and *k*-Nearest Neighbors Classifier, all from sklearn's libraries. I used AUC score to compare the models, as it is threshold independent and helps us understand how well the models perform. The results and ROC curves are shown below:

Algorithm	AUC score	Parameters
Logistic Regression	.640	C = 0.1, max_iter = 5000, penalty = 'L2'
kNN	.572	n_neighbors = 22
Random Forest	.608	max_depth = 6, n_estimators = 100, max_features = 'auto'


Logistic Regression 0.640393
k-Nearest Neighbors 0.572173
Random Forest 0.607952
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Logistic Regression clearly won out during the first round of model selection, with various words and terms standing out as the most important:

Top Features			Top Phrases		
2337	trump	0.93199	**1516**	national guard	0.25402
398	capitol	0.85238	**407**	capitol rioter	0.24119
302	biden	0.67308	**2266**	test positive	0.23696
2167	stop	0.57740	**1005**	ghislaine maxwell	0.23249
2272	texas	0.56380	**404**	capitol riot	0.22711
1243	judge	0.44283	**303**	biden inauguration	0.22688
1587	officer	0.40302	**1736**	police shooting	0.20422
1311	lawsuit	0.37314	**982**	gamestop share	0.19991
1460	minute	0.36625	**695**	derek chauvin	0.18096
2014	sell	0.34875	**696**	derek chauvin trial	0.17815

The fact that these features rose to the top of the importance ranking, however, brings us to an issue that can be best explained through the lens of our business case.

6. Business Case and Feature Selection

The primary issue with using these features to predict story virality is that the most important features are all representative of discreet storylines. Features like these will have little predictive value to a news service or social media site. Although “capitol rioter” may be a useful predictor of a story’s engagement in early 2021, it is unlikely to retain its predictive value in the future. The goal of this project is to identify words or phrases that predict engagement without relying on a specific storyline, so more steps need to be taken in order to increase the model’s validity.

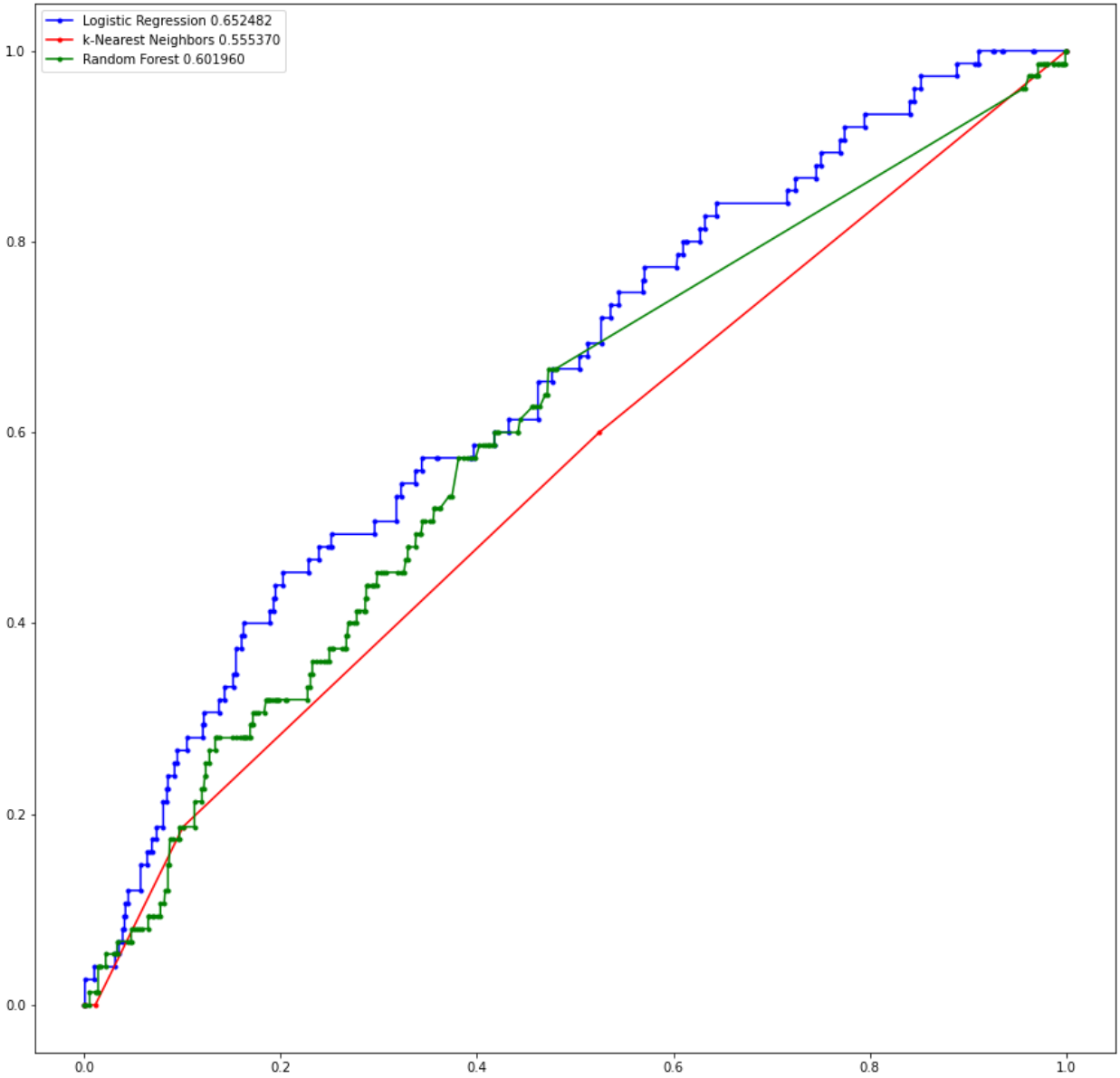
With this in mind, “major” stories needed to be identified within the dataset. The following stories were considered “major” stories based on their uniqueness and frequency in the dataset:

Story	**Labels**
Capitol riot	“capitol”
Derek Chauvin trial/Killing of George Floyd	“chauvin”, “floyd”
Ghislaine Maxwell trial	“ghislaine”
Killing of Breonna Taylor	“breonna”
Kyle Rittenhouse trial	“rittenhouse”
2021 Inauguration	“inauguration”

Overall, 763 stories were identified and removed from the dataset. The Top 5% of stories contained 94 “major” stories (dropped) and 448 “non-major” stories (retained). Now, models trained on the dataset would avoid overfitting to early-2021 major news stories.

7. Model Selection

The classifiers were trained once more, this time using the new feature set:



Logistic regression was once again the top performing classifier, and actually gained a few points of predictive value. Additionally, the new feature set:

Top Features			Top Phrases		
1496	officer	0.15141	**1625**	police officer	0.07031
2138	texas	0.10271	**2133**	test positive	0.03164
2203	trump	0.09526	**1646**	positive covid19	0.03051
1622	police	0.09183	**1598**	pipe bomb	0.02630
1315	mask	0.07265	**961**	governor sign	0.02513
1625	police officer	0.07031	**925**	gamestop share	0.02502
200	arrest	0.06350	**849**	federal judge	0.02501
1176	judge	0.06265	**311**	body camera	0.02357
1503	official	0.05752	**1624**	police chief	0.02354
3	100	0.05401	**1631**	police shooting	0.02334

These new lists present a much more promising list for writing an engaging headline. There are not many surprises in the new lists, with stories about police, outgoing President Donald Trump, and public officials commanding the most importance.

8. Conclusion

This model is designed to be a tool in the belt of news services and social media sites, as well as providing information that can help educate news consumers. Though the model is not perfect in any sense, it provides insight into what stories and topics are important to news consumers. This product could also be combined with other tools to form an overall strategy for a news organization, to drive engagement and traffic. While world events are unpredictable and complex, this model helps provide a useful window through which to interpret news reporting.

There are a number of next steps that could follow the development of this model. Since this model focused only on news headlines, there are many opportunities to research other aspects of news stories that may influence engagement. Variables such as the source of a news story, the day of the week or the time the story broke, and the region of the country or world from where it originates are some examples of the other influences that may send a news story to the top of r/news. The fact that this dataset was sourced from r/news is another limitation of the study worth mentioning. There are countless other sites where consumers interact with news, and analysis of what drives traffic and engagement on those sites is also useful for predicting engagement.