Evaluating Article Virality based off of Article Attributes

The BEST Fit - Philip, Olivia, Leo, Allison

2025-04-28

Introduction and Data

Project Motivation and Research Question

The ways in which people interact with media and discover news have dramatically shifted in recent years, with social media often displacing traditional news outlets. The decentralized nature of social media means the reach of each article is largely dependent on its individual merits, rather than the popularity of the publication it belongs to.

Thus a question arises: What article attributes are associated with social media virality?

Dataset and Key Variables

In this report we investigate the effects of different article features on social media success using the University of California Irvine Machine Learning Repository's "Online News Popularity" data set. It includes share counts and descriptive characteristics for articles published by Mashable, a digital media website, over two years (from 2013 to 2015). The data has 39,644 entries in total, with each repsenting an individual article and its associated textual and metadata features.

Key Variables:

is_viral - Binary response variable created that evaluates whether an article is considered viral or not, based off of whether total shares is greater/less than 1,400 [INSERT CITATION HERE] (0: FALSE, 1: TRUE).

data_channel - Categorical variable denoting article topic, merged from indicators: data_channel_is_lifestyle, data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, and data_channel_is_world. This classifies content by subject area.

day_published - Categorical variable indicating publication day, merged from indicators: weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_saturday, weekday_is_sunday. Additionally includes is_weekend (mean 0.1309) to distinguish weekday from weekend publications.

title_sentiment_polarity - Measure of the title's sentiment polarity (positivity/negativity). Values range from -1.0 (extremely negative) to 1.0 (extremely positive), with a mean of 0.0714 and standard deviation of 0.2654. This indicates how emotionally charged article titles are.

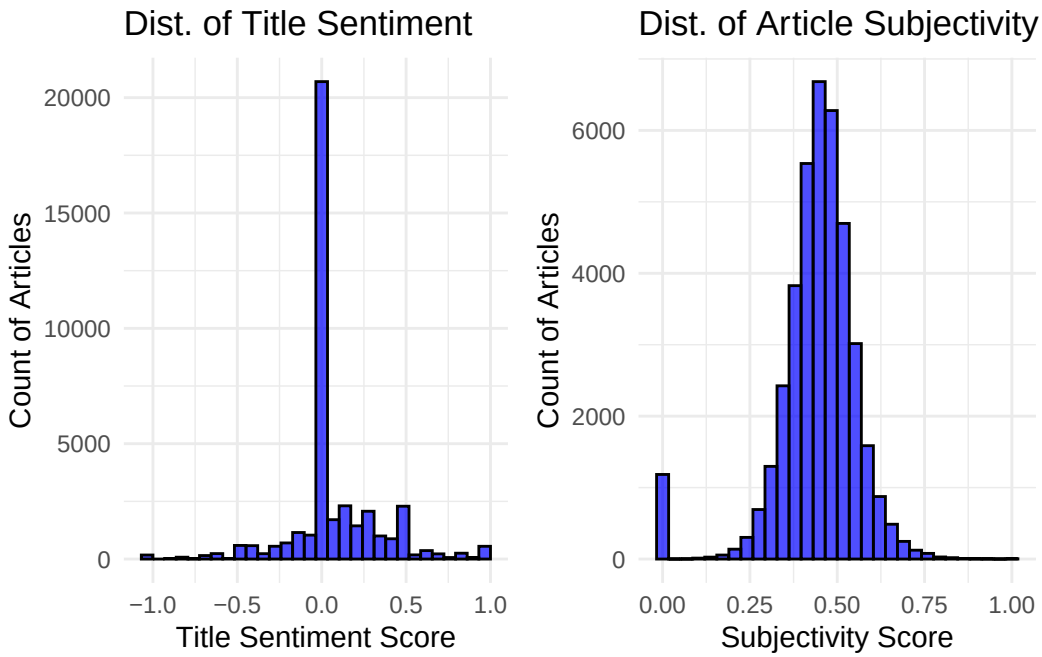
n_tokens_content - Number of words in the article content. Values range from 0 to 8,474 words, with a mean of 546.51 and standard deviation of 471.10. This quantifies the overall length of the article.

kw_avg_avg - Average shares of average keywords in the article. Values range from 0.0 to 43,567.66, with a mean of 3,135.86 and standard deviation of 1,318.13. This measures the expected popularity of the article's keyword selection.

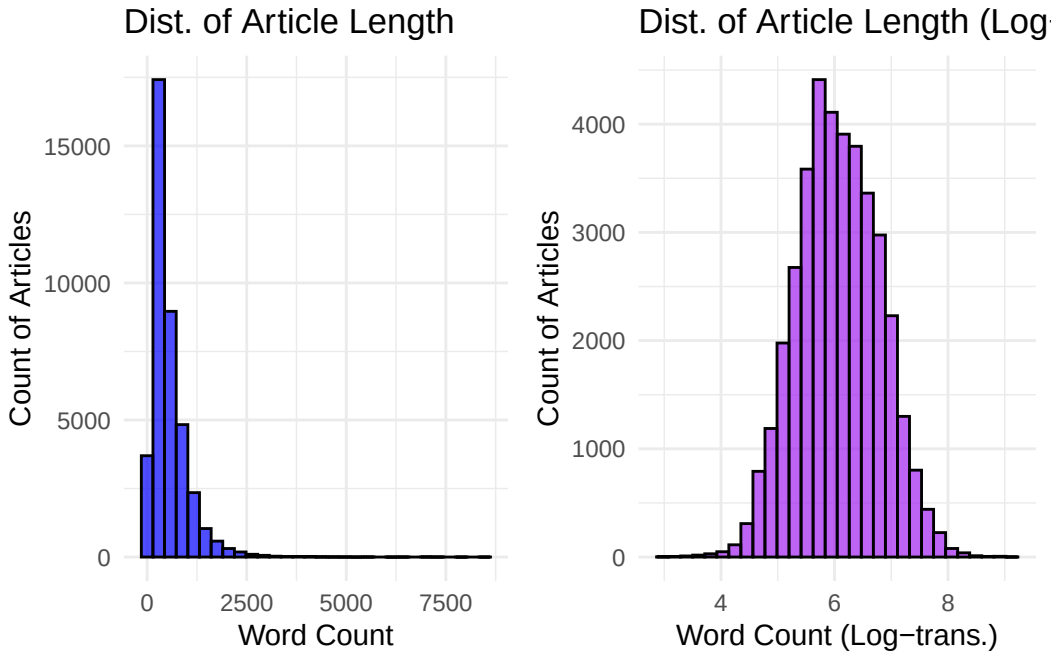
global_subjectivity - Measures the overall subjectivity of the article text. Values range from 0.0 (completely objective) to 1.0 (completely subjective), with a mean of 0.4434 and standard deviation of 0.1167. This quantifies how opinion-based versus fact-based the content is.

Univariate EDA

To understand our response and predictor variables more deeply, we first looked at their individual distributions. We found that while some variables are relatively approximately symmetric, others are heavily skewed and required log transformations to better meet modeling assumptions.

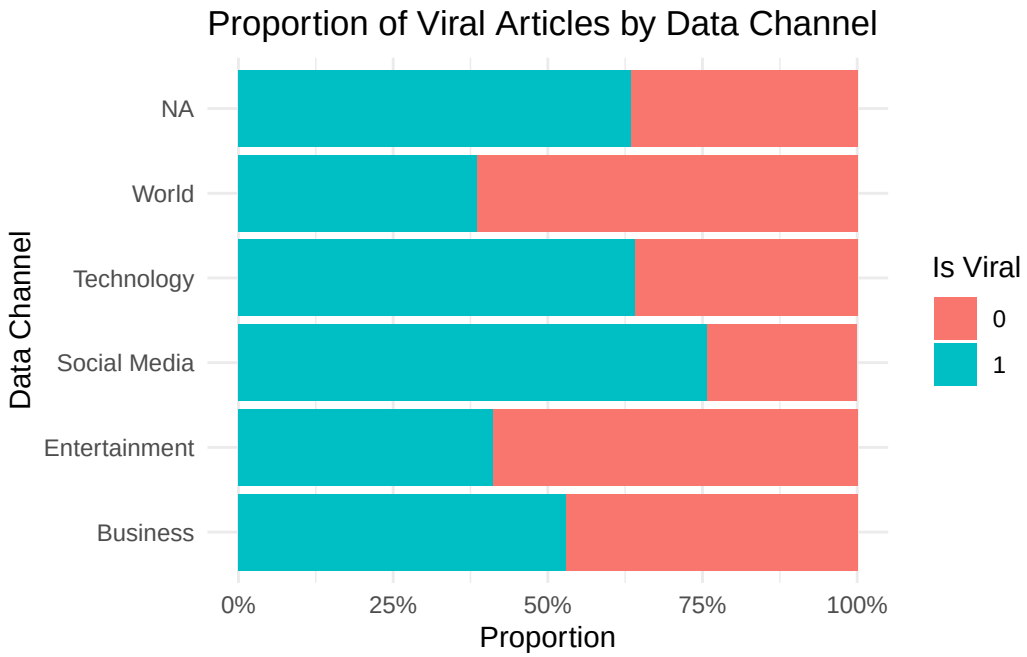


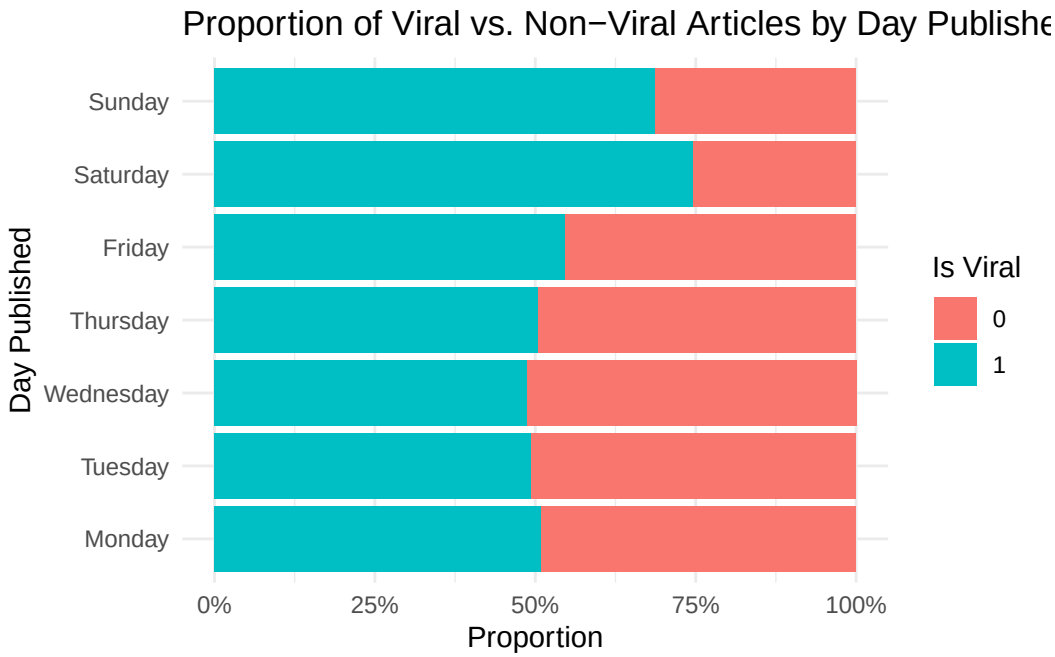
The title_sentiment_polarity variable quantifies the emotional change of article titles. The distribtuion suggests that many titles are emotionally neutral. The global_subjectivity variable is rougly symmetric with its distribution suggesting that most articles contained a balanced mix of factual and opinion-based language. Both global_subjectivity and title_sentiment_popularity display relatively balanced distributions that do not require transformation.



Both `n_tokens_content` and `kw_avg_avg` displayed heavily right-skewed distributions that warranted log transformations.To correct for this skewness and reduce the influence of extreme values, we log-transformed both variables. Post-transformation, the distributions became more symmetric and centered, better aligning with modeling assumptions.

Bivariate EDA

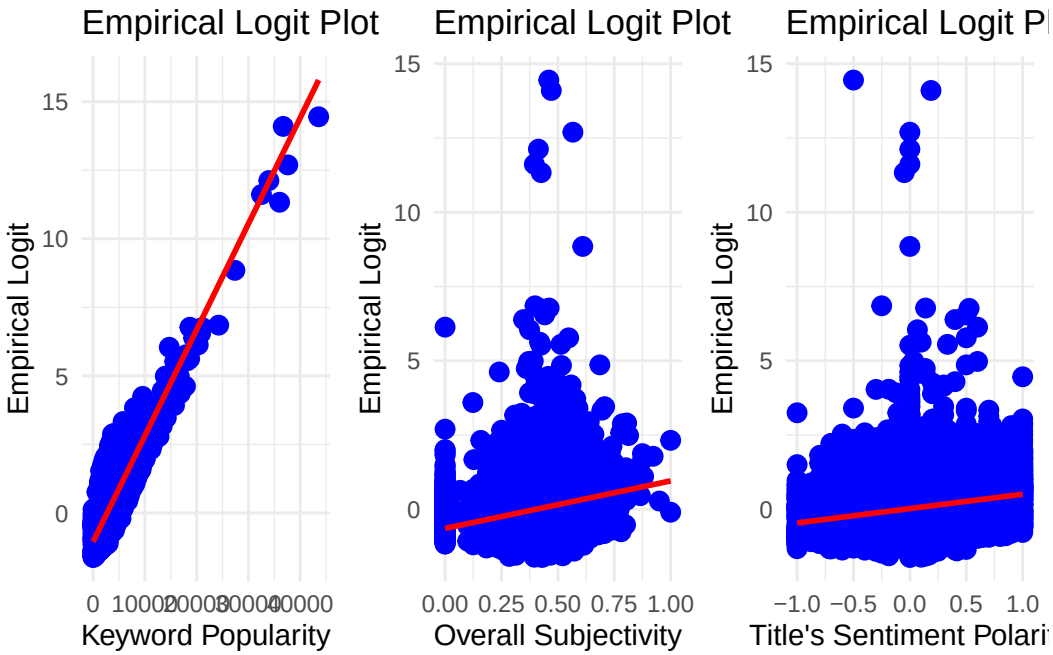




The first graph explores how virality varies across different content categories, as captured by the data_channel variable. Article published under the Social Media category have the highest proportion of viral outcomes, with many reaching viral status, This suggest that content tailored for or about social platforms may be particularly good for engagement. Articles that

are categoized under Technology and Business also show strong performance, with over half of the articles in each category going viral, On the other hand, articles in the Entertainment and World categories are less likely to go viral, falling below the 50% mark. This graph underscores how topic area influences content reach potentially because of the differences in audience behavior or platform algorithms.

The second graph examines the relationship between the day an article is published and its likelihood of going viral. Articles published on weekends are substantially more likely to be viral compared to those published during the week. Saturday stands out with the highest porportion of viral content, followed closely by Sunday. In contrast, weekday article tend to have lower virality rates, with viral and non-viral occurring in nearly equal proportions. These findings suggest that timing plays a role in determining an article's reach, likely reflecting differences in user engagement patterns across the week.



The empirical logit plots offer further insight into the relationship between key predictors and the binary outcome is_viral. The plot for average shares of average keywords reveals a positive relationship with virality. As the average popularity of an article's keywords increases, the likelihood of the article going viral rises sharply, with the empirical logit showing an upward linear trend. This suggest that keyword selection plays a key role in driving article engagement and affirms kw_avg_avg is a key predictors in modeling vitality. In contrast, the plot of overall subjectivity indicated on a slight positive relationship. Although the fitted line trends upward, the data are widely dispersed, and the effect appears weak and inconsistent, implying subjectivity alone is not a reliable driver of viral outcomes. A similar conclusion can be drawn from the plot of title sentiment polarity. While there is a marginal positive

slope, suggesting that more positive titles generate a slightly greater chance of going viral, the overall relationship is weak. The plots highlight a clear distinction in predictive power among the variables and support prioritizing keyword metrics over emotional tone or subjectivity when modeling article vitality.

Methodology

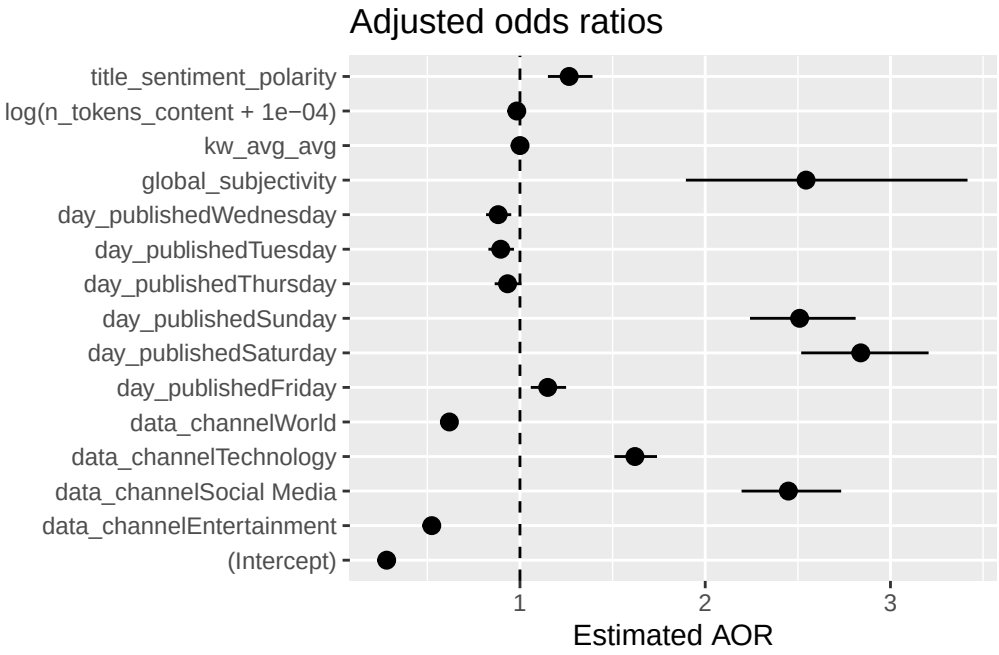
Based on our initial EDA and empirical logit visualization, we selected data channel, day published, article subjectivity, title sentiment polarity, average keyword popularity, and (log) article content length to fit an initial logistic model. Our model attributes are shown below:

Term	Estimate	Std.Error	z-statistic	p-value
(Intercept)	-1.2713	0.0782	-16.2534	0.0000
kw_avg_avg	0.0004	0.0000	23.0515	0.0000
log(n_tokens_content + 1e-04)	-0.0173	0.0070	-2.4803	0.0131
data_channelEntertainment	-0.6464	0.0366	-17.6423	0.0000
data_channelSocial Media	0.8957	0.0559	16.0287	0.0000
data_channelTechnology	0.4828	0.0361	13.3593	0.0000
data_channelWorld	-0.4796	0.0354	-13.5503	0.0000
day_publishedTuesday	-0.1097	0.0393	-2.7913	0.0052
day_publishedWednesday	-0.1253	0.0393	-3.1910	0.0014
day_publishedThursday	-0.0694	0.0395	-1.7574	0.0788
day_publishedFriday	0.1395	0.0423	3.2957	0.0010
day_publishedSaturday	1.0436	0.0616	16.9403	0.0000
day_publishedSunday	0.9202	0.0578	15.9141	0.0000
global_subjectivity	0.9338	0.1502	6.2160	0.0000
title_sentiment_polarity	0.2354	0.0484	4.8601	0.0000

Our initial fit gives all of the predictors significant p-values ($p<0.05$) and most predictors relatively high magnitude z-statistics, indicating that all variables in the model have statistically significant relationships with the likelihood of content going viral.

Coefficient Analysis

When fitting our initial model, we also visualized the adjusted odds ratios to ensure that all predictors were statistically significant.



From this initial visualization, most of the 95% confidence intervals for our predictor coefficients included 1, suggesting that the majority of our predictors added to the model. While a couple predictors (such as day_published_Thursday and log(n_tokens_content) were close to 1, we decided to still keep them in the model as they were either one level of several of a categorical variable, or because we found a possible interaction effect from the earlier EDA and felt it was at least a valuable predictor to consider in our model.

Interaction Effects

Next, we considered the addition of potential interaction effects between article length and data channel, and between global subjectivity and data channel. The hypothesis for this experiment were:

$$H_o : B_{n-tokens-content*data-channel} = B_{global-subectivity*data-channel} = 0$$
$$H_A : B_j \neq 0 \text{ for atleast one } j$$

Table 2: Drop in Deviance Test Results

Model	Log-Likelihood	Deviance Statistic (G)	df	p-value
Null Model	-20023.03	NA	NA	NA
Interaction Model	-19927.15	191.76	5	0

Examining the output of the deviance test, the p-value is very low, at around 0. This indicates that the data provides sufficient evidence that at-least one of the newly added interaction terms is a statistically significant predictor in whether an article will go viral or not, after accounting for data channel, day published, global subjectivity, title sentiment polarity, average key word popularity, and main body length for a given article. Therefore, we will keep the interaction effects in the final model.

Model Evaluation and Comparison

We further compared our two models through their ROC and AUCs.

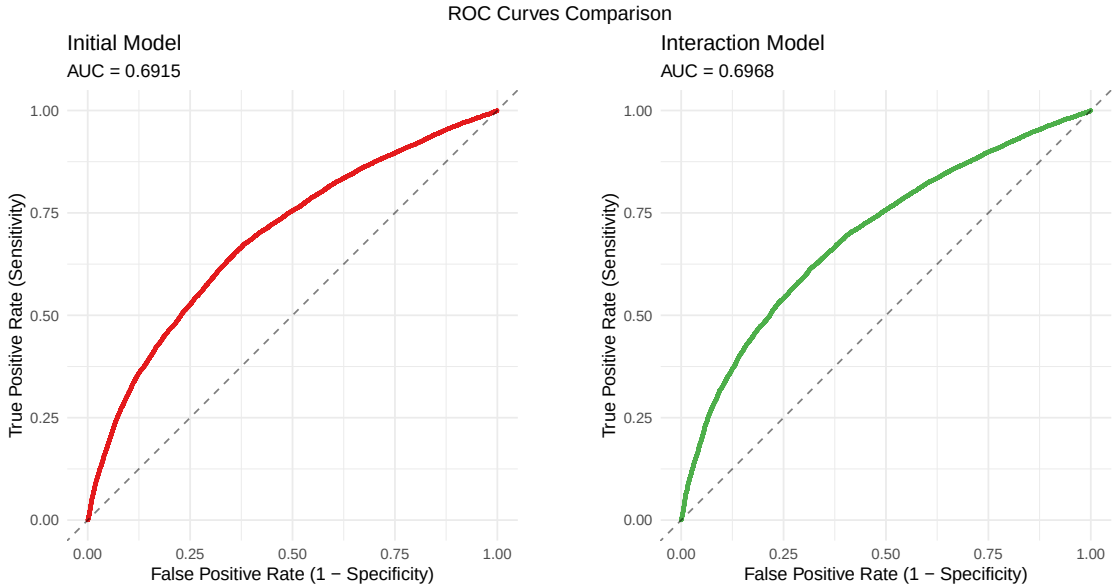


Table 3: AUC Values for Both Models

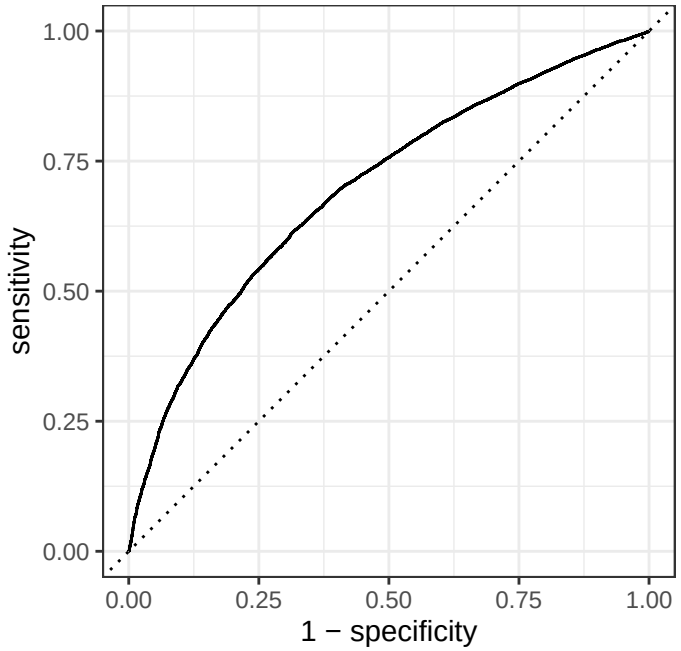
	.metric	.estimator	.estimate	model
roc_auc	binary	0.6915	Initial Model	
roc_auc	binary	0.6968	Interaction Model	

From the ROC curves, we can see that 1) Both models have an ROC curve above the random threshold, approaching the top left corner, indicating some predictive power in classifying an article and 2) The Interaction Model (AUC = 0.6902) demonstrates marginally better predictive performance than the Initial Model (AUC = 0.681), confirming our belief that the interaction effects are meaningful predictors. 3) Based on the curve, the optimal threshold for our model should target sensitivity ~ 0.65.

Selecting the point closest to the ROC curve to sensitivity 0.65 yields a threshold of approximately 0.464.

```
Optimal threshold for classification: 0.493
```

Model Performance




```
    35211
0.4932468
```