Analysis Written Report

The BEST Fit - Philip, Olivia, Leo, Allison

2025-04-10

Analysis + peer review

Draft report

Introduction and data

Our project utilizes 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 over two years (from 2013 to 2015). Mashable Inc. is a digital media website founded in 2005 and as of November 2015, it has over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. The data set in total, has 39644 observations, each representing an individual article. Our project motivation is that with the rise of the internet, we're interested in seeing how different factors influence or decide what goes "viral" on social media. For media companies specifically, this could aid in revealing patterns in what attracts readers to certain articles. Thus, our project aims to answer the

Key Research Question:

How do different article attributes (ex. Polarity, Positive/Negative Sentiments, Number of Images, etc.) relate to its virality on social media?

Key Variables:

rate_positive_words - Rate of positive words among non-neutral tokens in the article content. Values range from 0.0 to 1.0, with a mean of 0.6822 and standard deviation of 0.1902. This metric captures the positive emotional tone of the article.

Rate_negative_words - Rate of negative words among non-neutral tokens in the article content. Values range from 0.0 to 1.0, with a mean of 0.2879 and standard deviation of 0.1562. This metric captures the negative emotional tone of the article.

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.

n_tokens_title - Number of words in the article title. Values range from 2 to 23 words, with a mean of 10.40 and standard deviation of 2.11. This measures length of headlines.

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.

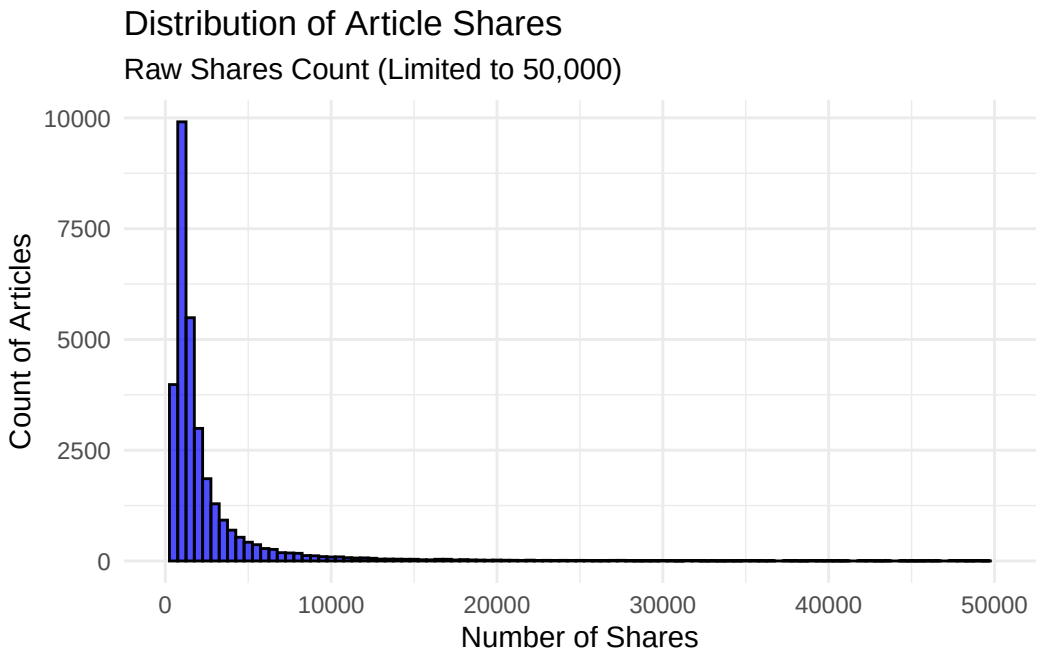
kw_avg_avg - Average shares of average keywords in the article. Values range from 0.0 to 43,567.66, with a mean of 3,135.86 and standard deviation of 1,318.13. This measures the expected popularity of the article's keyword selection.

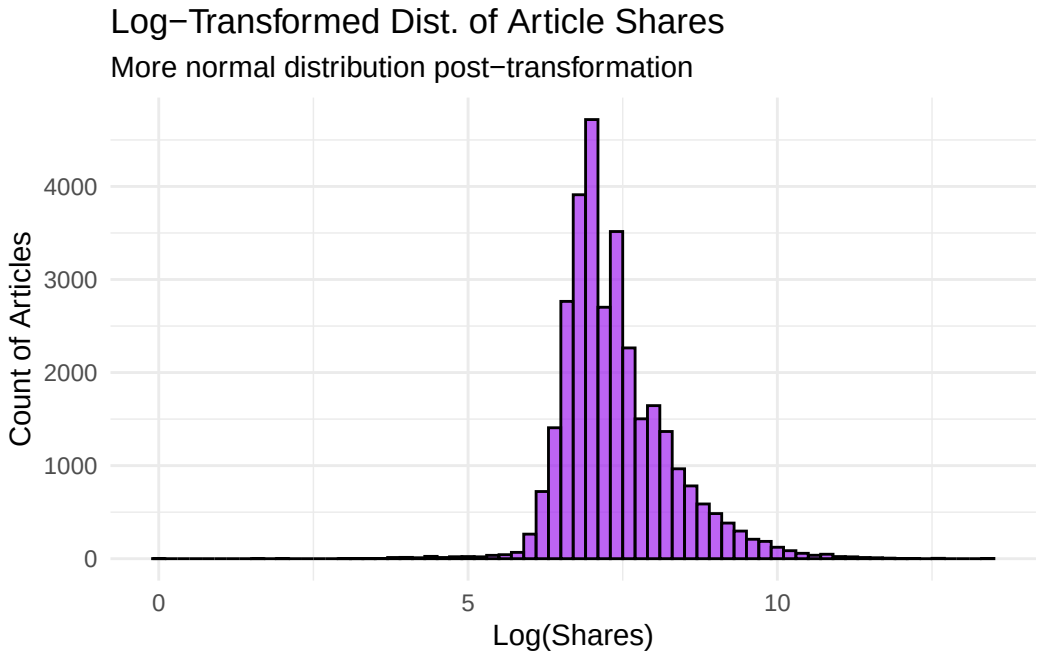
global_subjectivity - Measures the overall subjectivity of the article text. Values range from 0.0 (completely objective) to 1.0 (completely subjective), with a mean of 0.4434 and standard deviation of 0.1167. This quantifies how opinion-based versus fact-based the content is.

title_sentiment_polarity: A 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 a standard deviation of 0.2654.

Key EDA

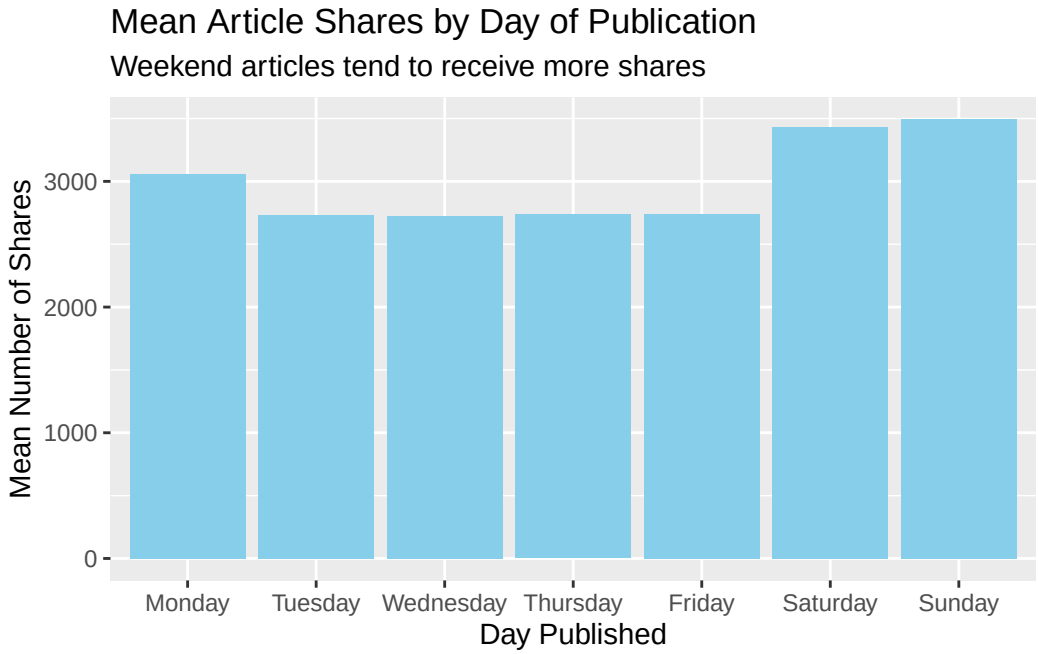
Response Variable - our initial EDA of the response variable revealed that it had a heavily right skewed, unimodal distribution. Thus, we imposed a log transformation, which was more symmetric and normally distributed.



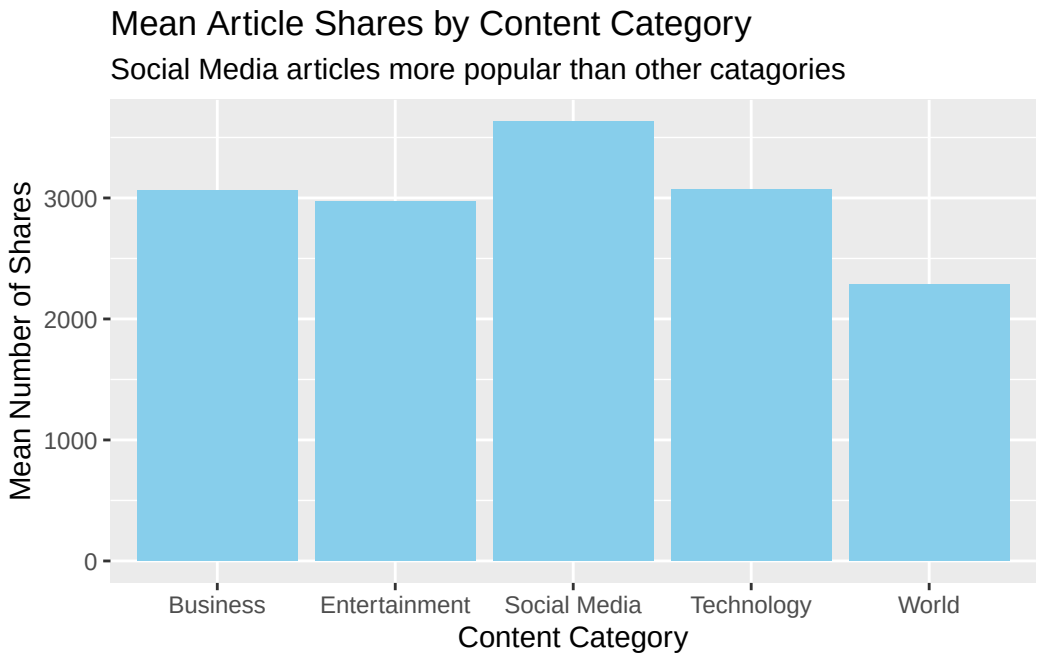


Key Variables - The key predictor variables we found from our initial exploration were Data Channel and Day Published, with the bivariate EDA we performed with our response variable, log(shares), shown below.

```
# A tibble: 6 x 2
  day_published mean_shares
  <fct>               <dbl>
1 Monday              3057.
2 Tuesday             2731.
3 Wednesday           2727.
4 Thursday            2737.
5 Friday              2741.
6 Saturday            3431.
```




```
# A tibble: 5 x 2
  data_channel  mean_shares
  <fct>               <dbl>
1 Business            3063.
2 Entertainment       2970.
3 Social Media        3629.
4 Technology          3072.
5 World               2288.
```



Methodology

From our EDA, we found that the most significant predictor variables were data_channel and day_published, however, other predictors we initially viewed (such as n_tokens_title, etc.) were not significant. Thus, we chose to fit an initial MLR model using data_channel and day_published, alongside other predictors that we hadn't attempted before (such as kw_avg_avg, n_tokens_content, etc.).

This produced said results:

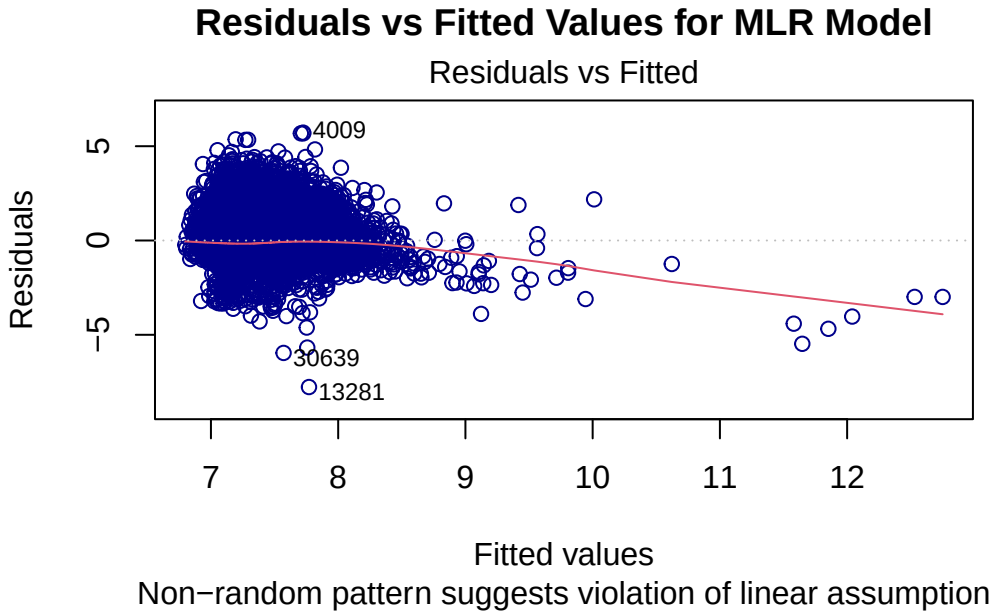
term	estimate	std.error	statistic	p.value
(Intercept)	6.8521	0.0286	239.3774	0.0000
kw_avg_avg	0.0001	0.0000	30.4256	0.0000
n_tokens_content	0.0001	0.0000	9.2061	0.0000
data_channelEntertainment	-0.1485	0.0145	-10.2424	0.0000
data_channelSocial Media	0.3056	0.0203	15.0761	0.0000
data_channelTechnology	0.1816	0.0144	12.6383	0.0000
data_channelWorld	-0.1531	0.0141	-10.8298	0.0000
day_publishedTuesday	-0.0516	0.0156	-3.3116	0.0009
day_publishedWednesday	-0.0590	0.0156	-3.7870	0.0002
day_publishedThursday	-0.0532	0.0157	-3.3904	0.0007
day_publishedFriday	0.0162	0.0168	0.9627	0.3357
day_publishedSaturday	0.2615	0.0224	11.6517	0.0000

term	estimate	std.error	statistic	p.value
day_publishedSunday	0.3113	0.0218	14.2499	0.0000
global_subjectivity	0.2492	0.0491	5.0792	0.0000
title_sentiment_polarity	0.0976	0.0189	5.1778	0.0000

```
              Metric  Value
1                 R^2 0.0903
2 Adjusted R-squared 0.0899
3               RMSE 0.8303
```

We found that all our predictors were statistically significant (p-value < 0.05), with the exception of day_publishedFriday, and thus kept them in this model.

```
              Metric  Value
1                 R^2 0.0903
2 Adjusted R-squared 0.0899
3               RMSE 0.8303
```



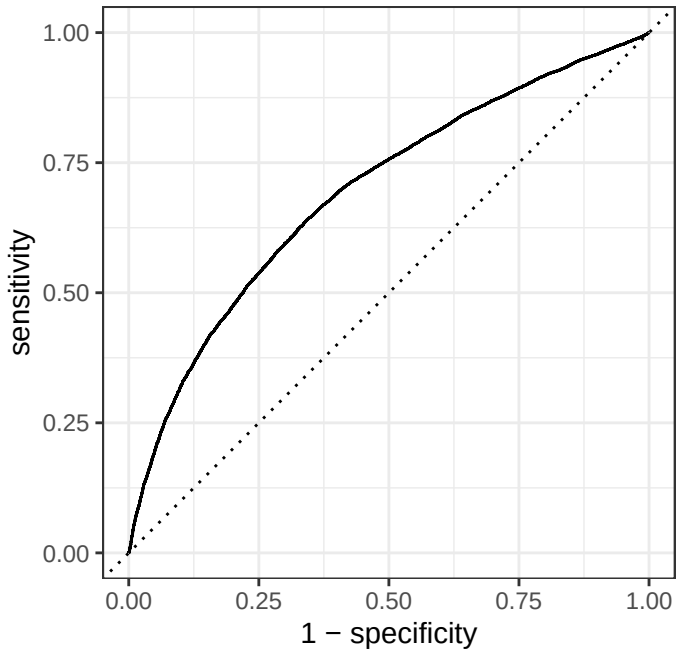
Our MLR had an $RMSE$ of 0.8303 and R^2 value of 0.0903 . Considering that our response variable was log(shares), this suggests that this MLR model is poorly fit to our data, as it can only explain about 9.03% of the variability that we can see from our response variable,

log(shares), and moreover, has a very high RMSE value. This is reaffirmed by the residuals plot, which reveals that the residuals are not randomly distributed, meaning that linear regression might not be the ideal model as our data does not satisfy the linear regression criteria.

Given these limitations, we pivoted to use a logistic regression model to classify an article as popular or otherwise. We selected a threshold for popularity of "1400 shares" based on prior literature, transforming shares column into a binary response variable with 1 for articles more popular than 1400 shares and 0 for those with less. Due to the heavy tails of 'shares' and non-linear relationship between the predictors and article shares, we would expect the logistic regression to be a more effective choice.

As for predictors, we continued using statistically significant predictors data_channel,day_published, etc.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.3623	0.0782	-17.4177	0.0000
kw_avg_avg	0.0004	0.0000	23.4937	0.0000
n_tokens_content	0.0003	0.0000	10.1961	0.0000
data_channelEntertainment	-0.6580	0.0367	-17.9427	0.0000
data_channelSocial Media	0.8882	0.0560	15.8688	0.0000
data_channelTechnology	0.4839	0.0362	13.3727	0.0000
data_channelWorld	-0.5048	0.0355	-14.2039	0.0000
day_publishedTuesday	-0.1092	0.0394	-2.7754	0.0055
day_publishedWednesday	-0.1194	0.0393	-3.0376	0.0024
day_publishedThursday	-0.0675	0.0396	-1.7056	0.0881
day_publishedFriday	0.1450	0.0424	3.4206	0.0006
day_publishedSaturday	1.0247	0.0618	16.5937	0.0000
day_publishedSunday	0.8979	0.0580	15.4847	0.0000
global_subjectivity	0.5384	0.1258	4.2801	0.0000
title_sentiment_polarity	0.2244	0.0486	4.6197	0.0000




```
4126
0.4916522
```

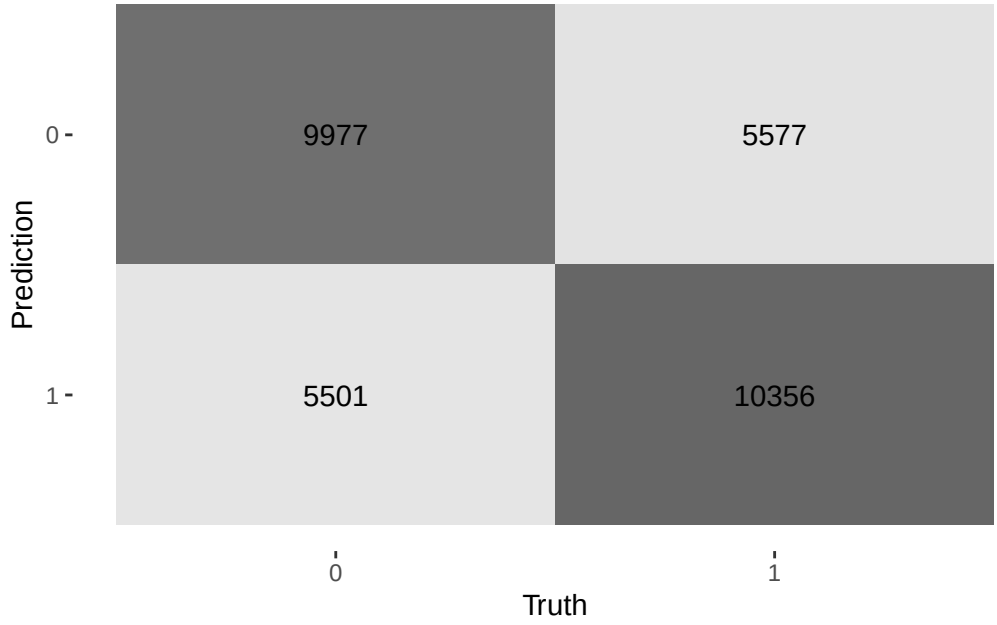


Table 3: Logistic Model Metrics Summary

Metric	Value
Accuracy	0.647
Misclassification Rate	0.353
Sensitivity (Recall)	0.650
Specificity	0.645
Precision	0.653
False Positive Rate (FPR)	0.355
False Negative Rate (FNR)	0.350
AUC	0.693

```
                          GVIF Df GVIF^(1/(2*Df))
kw_avg_avg                1.095946  1        1.046874
n_tokens_content          1.030240  1        1.015007
data_channel              1.174273  4        1.020284
day_published             1.017642  6        1.001458
global_subjectivity       1.069706  1        1.034266
title_sentiment_polarity 1.007743  1        1.003864
```

While our logistic regression model's ROC curve has an AUC of about 0.693, this is a dramatic improvement from our initial linear regression model, and thus we chose to continue using our logistic model for this data set.

Results

The final model we fitted was:

$$
\begin{aligned}
\text{logit}(p_{isViral}) = & -1.3623 \\
& + \ 0.0004 \times \text{kwAvgAvg} \\
& + \ 0.0003 \times \text{nTokensContent} \\
& - \ 0.6580 \times \text{dataChannelEntertainment} \\
& + \ 0.8882 \times \text{dataChannelSocialMedia} \\
& + \ 0.4839 \times \text{dataChannelTechnology} \\
& - \ 0.5048 \times \text{dataChannelWorld} \\
& - \ 0.1092 \times \text{dayPublishedTuesday} \\
& - \ 0.1194 \times \text{dayPublishedWednesday} \\
& + \ 0.1450 \times \text{dayPublishedFriday} \\
& + \ 1.0247 \times \text{dayPublishedSaturday} \\
& + \ 0.8979 \times \text{dayPublishedSunday} \\
& + \ 0.5384 \times \text{globalSubjectivity} \\
& + \ 0.2244 \times \text{titleSentimentPolarity}
\end{aligned}
$$

Our logistic regression model has an AUC of around 0.693, an accuracy of 0.647, specificity of 0.645, and sensitivity of 0.650. In comparison, the mis-classification rate, FPR, and FNR rates were 0.353, 0.355, and 0.350 respectively. This suggests that our model is moderately well fit for the data, as while the accuracy, specificity, sensitivity, and precision were relatively high at around 0.650, the FNR, FPR, and mis-classification rates were lower, at around 0.350. This precision means that approximately 65% of articles predicted to be viral were correctly classified, indicating the model performs significantly better than random chance. This relatively low predictive power may also be due to random noise, as many features of each article are likely uncaptured by the dataset and article virality may be influenced by sudden trends.

From our model, we can conclude that several key factors significantly influence article virality:

Content category plays a critical role in determining article popularity. Notably, Social media articles and technology articles are approximately 2.43 and 1.62 times, respectively, more likely than a similar business article to go viral. Conversely, Entertainment and World news are less sucessful catagories, with 48.2% and 39.6% lower odds of going viral, respectively. This suggests that readers are particularly engaged with content about social media and technology innovations, while being less likely to widely share entertainment and world news.

Day of publication is another important factor in article vitality. Weekend publications dramatically outperform weekday content, with Saturday articles enjoying 2.79 times higher odds and Sunday articles 2.45 times higher odds of virality compared to Monday publications. This weekend effect likely arises from increased leisure time as people take off from work or school, as weekdays show more the opposite effects with Tuesday and Wednesday's articles being 10.4% and 11.3% less likely to be viral.

While our initial set of "article sentiment" variables had relatively low predictive power, our model predicts a purely subjective article would have 71.3% higher odds of achieving viral status compared to purely objective content. Similarly, an article with purely positive sentiment would have 25.2% higher odds than a similar neutral title. This trend shows that, overall, more emotionally charged and polarizing content tends to be shared more often than neutral reporting.

Finally, while statistically important to the model, article length and keyword popularity have relatively small coefficients, making them less important to practical cases.