Lab 2 : Online News Popularity

Agnese, Minazzo; Etienne, Ndedi; Tingting, Li

1 Introduction

The shift in news consumption patterns, as detailed in a Pew Research study[1], shows over 80% of Americans now access news via digital platforms, marking a departure from traditional print media and highlighting the internet's growing influence. As per the Organization for Economic Cooperation Development (OECD) findings[2], digital platforms have revolutionized news distribution and consumption. In this digital era, independent online news sites like Mashable are enhancing their market presence. A key strategy involves assessing whether a digital platform can maintain or increase its market power, with article categorization playing a crucial role in boosting shares and popularity. Certain categories, like entertainment, believed to be more popular, warrant specific attention. Our statistical analysis, leveraging Mashable's data, aims to answer a pivotal question:

> *Does publishing entertainment-focused articles on weekends influence their popularity on Mashable, as measured by share counts?*

In this study, we aim to answer the above question through an online news popularity dataset derived from the Mashable platform.

2 Data and Methodology

Sourced from the UCI database, the dataset for this study contains news popularity features obtained from the Mashable internet news platform by Fernandes[3]. The dataset includes articles published from January 7th, 2013, to January 7th, 2015, with a total of 61 columns and 39,644 rows. Fernandes conducted an analysis of the HTML code of the articles, extracting 47 features. These features encompass various aspects such as the number of keywords and word count. Additionally, he generated other statistics for the articles using natural language processing (NLP), including measures of text subjectivity and sentiment polarity. In this study, our focus is on increasing the share of articles in the "Entertainment" channel, and thus, the data set was narrowed down to 61 columns and 7,057 rows.

For this study, we split our data into 2 parts: an exploration set with 30% of the data for initial exploration and model building, and a confirmation set with 70% of the data for model confirmation and generating the final report.

As our main focus is the increase in the number of shares, we examined the distribution of our dependent variable (Y), revealing a significantly right-skewed distribution (skewness: 12.04). To address this and improve the fit for our Ordinary Least Squares (OLS) model, we applied a log transformation to the shares variable. This adjustment significantly reduced the skewness to 1.44, resulting in a distribution that more closely approximates normality, albeit with a slight right skew.

[1] https://www.pewresearch.org/short-reads/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

[2] https://www.oecd.org/daf/competition/competition-issues-concerning-news-media-and-digital-platforms-2021.pdf

[3] K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

As a news website, Mashable leverages article shares as a metric for determining popularity. Our initial belief is that the timing of article publication likely impacts the share count. Based on the table below, we extracted the daily article count and corresponding share statistics. The table shows a clear trend: there is a notably higher volume of articles published during weekdays (Monday to Friday) than on weekends (Saturday and Sunday). Interestingly, despite this, articles published on weekends tend to have lower share statistics compared to those published on weekdays.

To further analyze this, we compared weekdays and weekends. Surprisingly, the average article count on weekends constitutes only 36% of the average article count on weekdays. Nevertheless, the average shares (3357.18 VS 3012.22), median shares (1600 VS 1100), and mean of log shares (7.63 VS 7.26) for weekend articles surpass those of weekday articles. This study investigates if publishing entertainment articles on weekends boosts their share counts.

Shares Analysis for Entertainment Channel by Day of Week and Weekday/Weekend Summary

day_of_week	number_of_articles	total_shares	mean_shares	median_shares	mean_log_shares
Monday	959	2907474	3031.78	1100	7.27
Tuesday	885	2431611	2747.58	1100	7.23
Wednesday	920	2867045	3116.35	1100	7.25
Thursday	857	2560022	2987.19	1100	7.24
Friday	695	2234571	3215.21	1200	7.33
Saturday	258	775179	3004.57	1600	7.57
Sunday	366	1319699	3605.73	1600	7.68
---	---	---	---	---	---
Weekdays	4316	13000723	3012.22	1100	7.26
Weekends	624	2094878	3357.18	1600	7.63

Therefore we created our regression model as follows, where β_1 represents the increase in log of shares if articles are published on weekends, ϵ is the residual that captures the difference between the predicted value and the real world value.

$$log\_share = \beta_0 + \beta_1 \cdot (is\_weekend) + \epsilon$$

3 Results

Table 1 shows the results of three representative regressions. By comparing the adjusted R-squared and the result of the F-test, the best-performing model is Model_3. To ensure that our model is sound, and that we could use statistical guarantees, we tested the 5 Classical Linear Model (CLM) assumptions.

First, considering the i.i.d. assumption, our dataset comprises all articles published from 2013 to 2015, making it population data within this timeframe. While there may be instances of news influencing each other as creators imitate the most popular articles, we can still consider our data to mostly meet the i.i.d. assumption.

Table 1: Regression Results

	Output Variable: Log of Shares		
	(1)	(2)	(3)
Weekend	0.373***	0.356***	0.341***
	(0.037)	(0.037)	(0.037)
Subjectivity of Content		0.376**	0.386***
		(0.118)	(0.117)
Number of Keywords Squared		0.032***	0.026***
		(0.007)	(0.007)
Avg. Shares of Referenced Articles in Mashable		0.00001***	0.00001**
		(0.00000)	(0.00000)
Number of Links Squared			0.003**
			(0.001)
Number of Images Squared			0.002
			(0.001)
Avg. Shares of Average Keywords			0.0001***
			(0.00003)
Constant	7.260***	6.830***	6.445***
	(0.014)	(0.072)	(0.118)
Custom Note 1			
Custom Note 2			
Observations	4,940	4,940	4,940
R^2	0.018	0.034	0.060
Adjusted R^2	0.017	0.033	0.058
Residual Std. Error	0.927 (df = 4938)	0.920 (df = 4935)	0.908 (df = 4932)
Note:			Note: *** p<0.001, ** p<0.01, * p<0.05.

Second, regarding the assumption of no perfect collinearity, the VIF test was performed. The results indicate that none of the variables used in Model_3 have particularly high VIF values. All the VIF values are close to 1, suggesting that there is no evidence of problematic multicollinearity among the predictors.

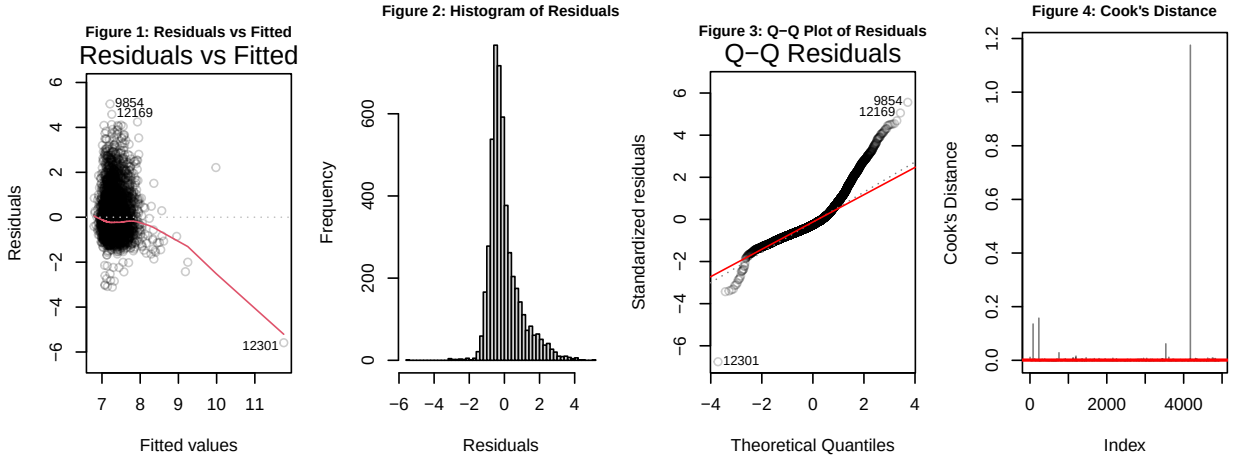
Third, concerning the assumption of linearity in the conditional expectation, we plotted the predicted values and residuals of Model_3. As shown in Figure 1, even with the quadratic transformation of our feature in the model, it is still overestimating for low values and underestimating for high values. This suggests that the assumption of linearity in the conditional expectation is not met

Fourth, regarding the assumption of constant error variance, we performed the Breusch-Pagan test. The test statistic (BP) is 230.98 with a p-value less than 2.2e-16, indicating strong evidence of heteroscedasticity in Model_3. Additionally, from Figure 1, as the fitted values increase, the variance of the residuals significantly differs on the left side. We can conclude that the data doesn't follow homoscedastic errors. However, this issue was addressed by performing heteroscedasticity-robust standard errors.

Fifth, regarding the assumption of normally distributed errors, by examining the histogram(Figure 2) and Q-Q plot(Figure 3) of residuals, the distribution appears right-skewed with fat tails. Furthermore, we calculated the skewness as 1.36 and kurtosis as 6, indicating that it doesn't meet the assumption of normally distributed errors.

In the evaluation of influential data points, Figure 4 illustrates Cook's distance for each observation in the regression model. Most points are clustered near zero, however, there are a few notable exceptions that stand out due to their higher Cook's distance values and warrant closer examination.

Based on the above evaluation of the CLM assumptions, we find that we satisfy assumptions 1 and 2, indicating compliance with the assumption of a large sample linear model. However, assumptions 3 to 5 are not met, suggesting that our model may be biased and exhibit significant variance.



Across the three models, the key coefficient 'is weekend' was highly statistically significant, ranging from 0.34 to 0.37. As the coefficient for 'is weekend' doesn't change too much across the three models, we can conclude that the estimate of the effect on "is weekend" is unbiased. This indicates that for articles in the entertainment channel on Mashable, the shares will increase by 34% if published on the weekends compared to weekdays.

However, for all our models, despite the p-value being highly significant, the adjusted R-squared is quite low. Even in the best-performing model (based on adjusted R-squared and F-test), Model_3, the adjusted R-squared is only 0.058. This suggests that our model explains only 5.8% of the variance in our Y variable and our model has quite low predictive power. As observed throughout the tests used to check for the Classical Linear Model (CLM) assumption, a significant level of variability is present.

The highly significant p-value for the model prompts us to reject the null hypothesis, affirming that the model does provide a better fit than merely considering the mean of the dependent variable. Yet, the low adjusted R-squared prompts us to consider additional explanatory variables or alternative modeling techniques that

may capture more of the variance in the dependent variable. It also highlights the importance of cautious interpretation of significant p-values, particularly when the overall model fit is limited. Therefore, future model refinement or the exploration of non-linear relationships might be warranted to enhance predictive accuracy and model utility.

4 Limitation

Violation of IID: Although we argued earlier that the data mostly meets the assumption of being independent and identically distributed (IID), it must be acknowledged that articles on Mashable could potentially be interdependent due to factors such as Mashable's recommendation algorithms, shared authors, and references to each other. Additionally, the dataset spans a 2-year timeframe, implying that the sample might be drawn from different distributions, especially considering that online news tends to become more popular over time.

Omitted variables: The popularity of news is a complex topic influenced by various features. The dataset's limitations in capturing factors affecting the popularity of Mashable's entertainment articles could lead to unexplained variance in their share counts. Key omitted variables include authors, reader demographics, the exact timing of publication, social media trends, article quality, and concurrent news events. These omissions can significantly influence share counts but are not accounted for; thus, they not only introduce bias to our estimates but also diminish the predictive power of our model.

Model limitations: while a significant p-value indicates that the model is better than no model, the low adjusted R-squared suggests that other unknown factors might be influencing the share counts. Potentially we can acknowledge that while the weekend might be a significant predictor, it is not the only factor, and the true relationship may be more complex. Additionally, According to Figure 4, the data has some outliers which might introduce potential biases in the coefficients.

5 Conclusion

This study examines whether articles published on weekends or weekdays benefit from higher shares on Mashable's entertainment channel. We found that shares increase by 34% if an article is published on the weekends compared to weekdays. However, the low adjusted R-squared values across our models suggest that while weekend publication is a significant predictor of popularity, it is far from being the sole determinant. More uncaptured features should be considered in future models.

In conclusion, we recommend a cautious interpretation of the results and advocate for the use of more sophisticated models that can accommodate the multifaceted nature of online news dissemination. Future research should aim to incorporate a broader range of factors.