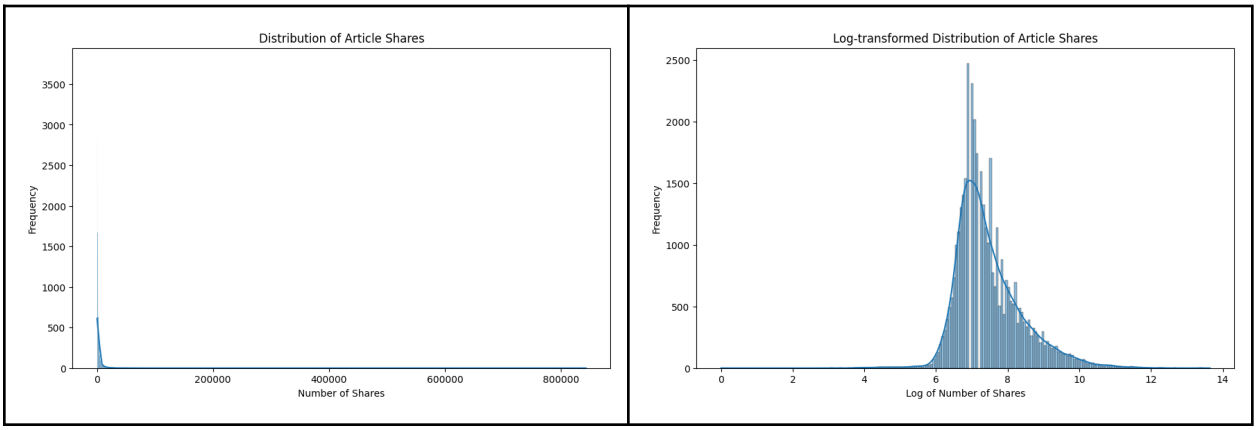
Mashable Exploratory Data Analysis

Our analysis began with an Ordinary Least Squares regression model to establish a baseline understanding of the data. The model accounted for approximately 12.7% of the variance in the logarithm of shares, with an R-squared value of 0.127. While this indicates a moderate level of predictability, it suggests that other factors not included in the model may also play a significant role in predicting article popularity.

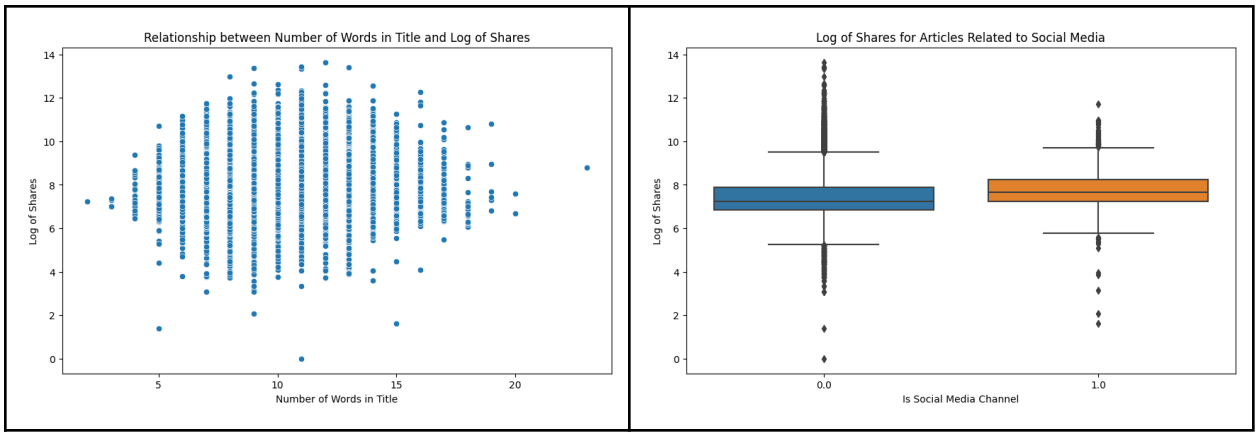
Target Variable Visualization

An initial examination of the target variable, the number of shares, revealed a highly skewed distribution with a long tail, indicating that most articles have a relatively low number of shares, while a few have a very high number. To normalize the distribution and improve model performance, we applied a logarithmic transformation. The resulting log-transformed distribution showed a more bell-shaped curve, indicating a better fit for linear modeling techniques.



Significant Predictor Relationships

Visual analyses of potential predictors were conducted to understand their relationships with the number of shares. For instance, the number of words in the title showed a pattern where neither too few nor too many words were associated with the highest number of shares. Similarly, articles related to social media generally had more shares, highlighting the importance of social media as a significant predictor in the distribution of content.



Model Performance Evaluation

The out-of-sample performance of our models provided a more realistic evaluation of their predictive capabilities. The initial OLS regression model's performance, with a test RMSE of approximately 1.2256, served as our benchmark.

Upon refining our models using decision trees and random forests, we achieved a test MSE of 0.7534 and 0.7181, respectively, after pruning and hyperparameter tuning. These improvements over the OLS model indicate a better generalization to unseen data.

In our models, several predictors stood out as significant. The **n_tokens_title predictor**, representing the number of words in an article's title, was consistently significant across different models. Additionally, whether an article was related to social media (**data_channel_is_socmed**) was a strong predictor, reflecting the platform's role in content dissemination.

Final Model Selection

The Random Forest model, with its ensemble approach, has demonstrated the best balance between complexity and predictive power. It not only captured the underlying trends in the data more effectively than the OLS model but also showed less variance in its predictions compared to the Decision Tree model. Thus, it stands out as the best predicting model for our analysis.