Reasons Why People Share Online Content

Alicia Verde, Eashan Thakuria, Robin Murphy, H Drake Stancil, Trystan Manning, Matthew A. Lanham
Purdue University, Krannert School of Management
verde@purdue.edu; ethakuri@purdue.edu; murph380@purdue.edu; hstancil@purdue.edu; mannint@purdue.edu; lanhamm@purdue.edu

ABSTRACT

We developed a predictive model that would help a media content provider identify article features that would increase the number of shares. The motivation for this study is that businesses rely heavily on revenue to compete and survive in the news industry. Today, most revenue comes from ads and shares/views from social media. Content providers need to know what factors lead to the largest number of shares since this leads to more earnings. We posit the content provider could use our predictive model to help them craft popular future articles. Using nearly 8,000 articles, many features about the articles' content were derived using text analytics. We use these features to develop a model that predicts shares. Our model provides estimated parametric effects of what works best to increase shares.

INTRODUCTION

As society progresses digitally, it is imperative for online news sources to understand the behaviors of its consumers and the market in which it is in. A study done by Microsoft revealed that an article has a user’s attention for about eight seconds. This makes it important to utilize predictive analytics to identify what will keep the user engaged, capture their attention longer, and ultimately entice them to share the content with others.

Companies can improve their marketing tactics once they gain an understanding of the factors that drive consumers to interact with a post, they can improve their marketing tactics. As an article receives more attention, the articles’ popularity increases overall profits, views, reposts, and other forms of content sharing, which leads to more revenue.
Source: *Microsoft*

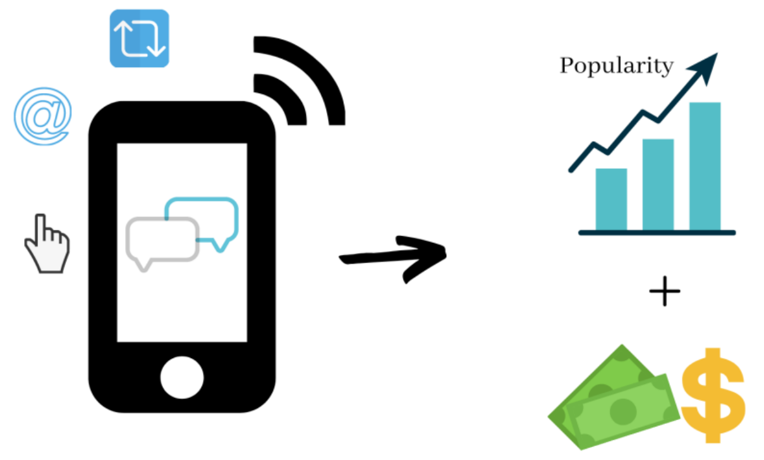


Fig 1. Views Lead to Revenue

Research Question

- What are the optimal article features that drive the most shares of an article post?

LITERATURE REVIEW

Previous literature regarding the discovery of features that significantly affect online news popularity has led to deeper insights for how businesses can maximize their viewership. Past studies have classified “popular” and “unpopular” articles based on a threshold of shares or have researched social media platforms to see what user-facing features indicate popularity. Our study is differentiated from previous papers due to our response variable lacking a rigid threshold for popular/unpopular articles. Moreover, our study emphasizes analyzing features that are visible prior to publication, and our comparison of feature significance among linear regression, k-NN, and Random Forest predictive models.

Study	Popularity Threshold	Linear Regression	K-NN	Random Forest	Pre-pub. Features
Shreyas (2016)	✓		✓	✓	✓
Uddin (2016)	✓			✓	✓
Khan (2018)	✓			✓	✓
Singh (2018)				✓	✓
Chopra (2019)					
Our Study		✓	✓	✓	✓

METHODOLOGY

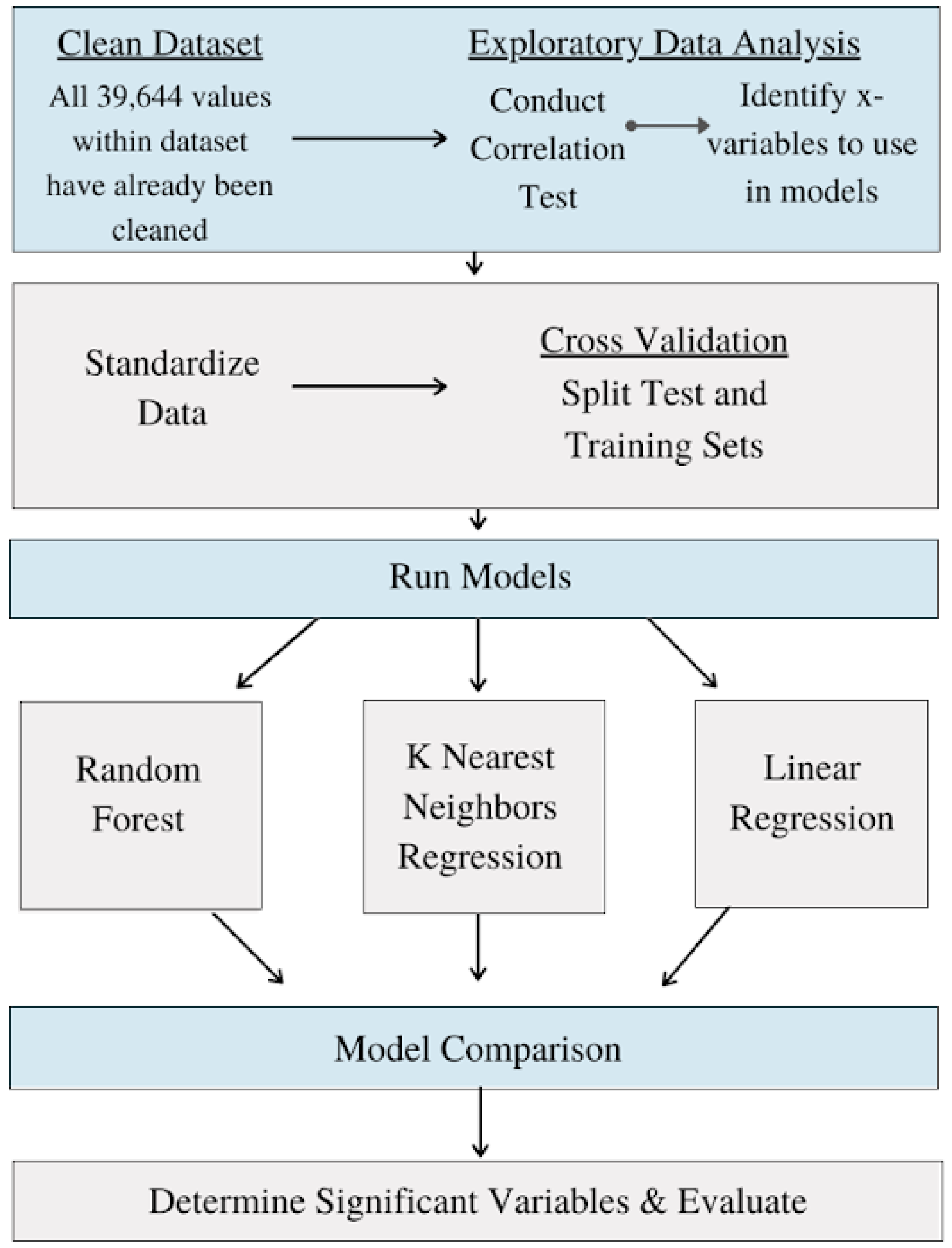
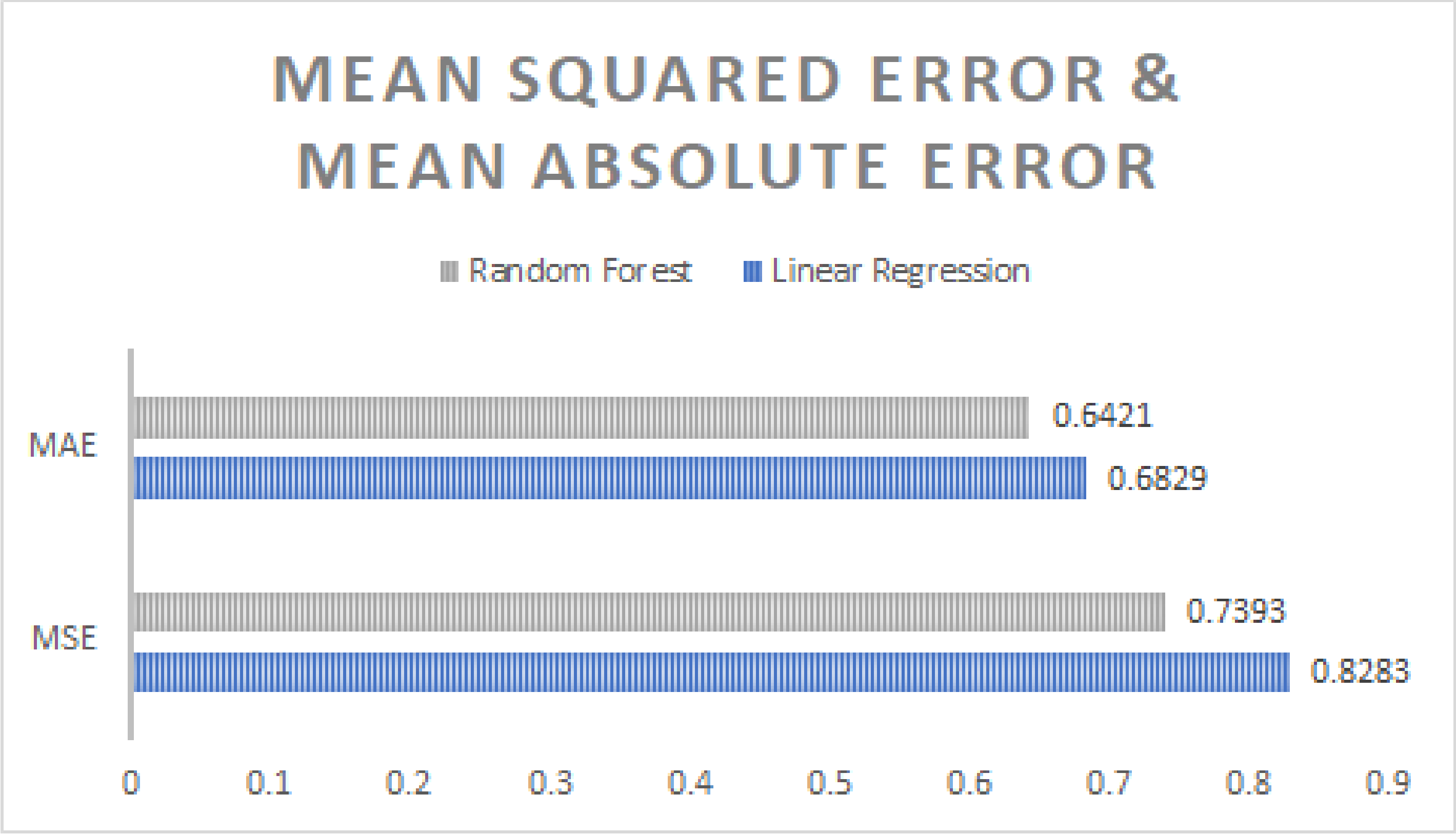

Clean Dataset
All 39,644 values within dataset have already been cleaned
Exploratory Data Analysis
Conduct Correlation Test
Identify x-variables to use in models
Standardize Data
Cross Validation
Split Test and Training Sets
Run Models
Random Forest
K Nearest Neighbors Regression
Linear Regression
Model Comparison
Determine Significant Variables & Evaluate

Fig 2. Methodology Flow Diagram

STATISTICAL RESULTS

Due to its low Mean Squared Error and Mean Absolute Value, the Random Forest has better results in comparison to the Linear Regression and K-Nearest Neighbors models.



Note: KNN is not included in this graphical representation due to its' extremely high MSE & MAE. KNN is not feasible because of the sheer size of the data and dimensions. With predicting important variables there can be a high variation in the results.

Fig 3. Model Comparison Results

EXPECTED BUSINESS IMPACT

As our world becomes more and more digitized it is imperative for online news platforms to remain competitive by focusing on factors which have the greatest impact on the overall shares of an article. As an article is shared more and more times it will result in greater revenue for the company. If we assume that each share results in one view and that a company yields roughly .0001 cents per view, it becomes crucial for the company to gain an understanding of the factors that drive shares. If a company were to focus on the top four share-driving features, they should steer their interests and resources on the number of times an article is referenced, the number of days between publications, world-centric articles, entertainment-centric articles, and articles shared on the weekend. Focusing on these factors will have the greatest impact on the overall shares an article receives and will in turn lead to more views and revenue for the company.

Variables	Importance to Shares*
Number of Times article was referenced	24%
Time between publication and data acquired	7%
Within category World	6%
Posted on the Weekend	6%
Within category Entertainment	5%

Fig 4. Ranked Important Variables

*The percentage that the prediction error increases when the variable is removed

CONCLUSIONS

The most important variables in identifying an online article’s shares are the number of times an article is referenced, the number of days between publications, world-centric articles, entertainment-centric articles, and articles shared on the weekends. After analyzing the online-news popularity data, certain relationships between variables, and the response variable, “shares”, became apparent. After our analysis of the variables, the most significant features for the models were chosen and those correlated with one another were dropped.

After developing models with K-Nearest Neighbors Regression, Random Forest Regressor, and Linear Regression with only the significant features, the results show that Random Forest Regressor achieved the lowest expected error results. This makes Random Forest the best model for determining the driving factors for an online-news platform’s article “shares”. These shares help companies identify which articles people enjoy reading the most.

Study Limitations:

- No proper data dictionary - poorly defined variables
- Derived data - it produced a model that was not easily interpretable

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Xinyu Wang, Theo Ginting, and our graduate student mentor's Xuan Mai Gguyen and Tianyi Yang for their guidance and support on this project.