Data Mining & Analysis on the Circulation of Online Publications:

Developing Regression Models and Methods to Predict the Popularity of Digital News & Media Articles

Akhila Reddy Bokka • bokka.a@northeastern.edu • (233) 280-9396
Shamhith Kamasani • kamasani.s@northeastern.edu • (857) 313-5653
Shreyas Risbud • risbud.s@northeastern.edu • (781) 664-7633

IE 7275 – Data Mining in Engineering

Professor Soumya Janardhanan

Northeastern University

Tuesday, December 12th, 2023

Abstract:

This project explores the implementation of data mining techniques on the circulation of online news articles and stories published by the digital media organization Mashable. For any given Mashable article included in this particular data set, the benchmark used to gauge its inherent popularity lies in the total number of recorded shares that it receives. During this study, five distinct regression models are developed to predict the natural logarithm of shares that an article gets based on a selected subset of numerical and categorical features. These models are assessed against each other with the aim of determining if regression is a viable machine learning method for forecasting the popularity of stories that are published online. Before passing the original data set into the data mining pipeline, the context surrounding the dissemination of information through digital media is discussed, and the concept of regression as a means of predicting user-driven sharing and reposting is introduced. The data set of Mashable links utilized for this problem statement is presented, along with the preliminary preprocessing that is conducted on it. Next, the distributions of the natural logarithm of shares are visualized with respect to each remaining feature in order to see if there are any noticeable relationships between the target variable and certain predictors. After these introductory stages, multiple, lasso, and ridge linear regression models which predict the natural logarithm of shares are trained and evaluated on the testing and validation sets. For additional comparison and assessment, a multiple linear regression model with principal component analysis employed, and a regression tree, are trained. The performance metrics from these models are then computed and weighed against one another. Based on the results of this investigation, further considerations into the real-world applicability of regression models for estimating online news popularity are examined and deliberated upon.

Keywords: Correlation • Distribution • Lasso Linear Regression • Mashable • Mean Absolute Error (MAE) • Mean Absolute Percentage Error (MAPE) • Mean Squared Error (MSE) • Multiple Linear Regression • Natural Logarithm • Online Popularity • Prediction Error • Principal Component Analysis (PCA) • Regression Tree • Ridge Linear Regression • Root Mean Squared Error (RMSE) • Shares • Standardization • Testing Set • Training Set • Validation Set

Background:

The advent of the internet and the subsequent emergence of digital outlets have substantially expanded the infrastructure surrounding modern-day discourse and information retrieval. For centuries, print journalism taking the form of newspapers, books, letters, and magazines was the dominant system of large-scale communication. With the invention and growing prominence of radio and television, broadcast news quickly became an effective method of bringing critical knowledge and intelligence to the masses during the 20th century. While these traditional avenues for consuming information still exist today, it is irrefutable that they have been hampered by the creation of websites, search engines, and social networking applications (Shearer et al., 2021). Apart from allowing news organizations to extend the reach of their audience, online media gives users more opportunity in how they wish to get their information. Specifically, when a news report is released online, a user may choose between an article, social media post, video, podcast, or alternative means to learn about the details of that story.

In 2020, the Pew Research Center carried out a survey to ascertain the current habits among the American public regarding their information consumption. According to the results of these studies, approximately 52% of American adults get their news coverage via a digital medium, while 35%, 7%, and 5% are likely to get it from television, radio, and printed articles, respectively. Moreover, within the category of online outlets, news websites and applications earned the highest ranked preference over search engines, social media platforms, and podcasts (Shearer et al., 2021). Because of these distributions within the American populace, both businesses and broadcasting companies, which are subject to streams of advertising revenue, have been forced to reallocate their assets from nondigital to digital outlets. A recent fact sheet issued by the Pew Research Center shows a distinct uptrend in digital advertising taking place between 2011 and 2022. For instance, about 21% of cumulative advertising revenues originated from digital platforms in 2011, while nearly 72% of them came digitally in 2022.

Social media and online networking have been among the strongest contributors to the rising annual digital advertising revenues. A major finding from a 2019 study published by Our World in Data reports that of the 3.5 billion internet users around the globe, about two out of every three of them are present on social media (Ortiz-Ospina et al., 2019). One of the foremost reasons for why social media has captured a sizable portion of the digital market space is user accessibility. The developers behind these applications put a great deal of investment in designing their interfaces to ensure that users will continue to utilize and remain engaged with their products. Although there are dozens of well-known social networking platforms and applications in existence, they all possess subsets of foundational features. One of these common attributes includes the ability to follow, subscribe to, or connect with different users. Furthermore, users are able to toggle like or dislike buttons for certain material, express themselves in posts, comments, and chats, and redistribute content by sharing or reposting it.

Introduction:

The redistribution capabilities build into websites and applications have been some of the key drivers behind today's fast and widespread online dissemination. On average, sites that allow users to share posted information with others garner roughly seven times more web traffic (Richards et al., 2023) than those that do not. With expectations like these, it is clear as to why marketing on social media takes up a considerable segment of total digital advertising revenues. To put these cash inflows into perspective, based on an updated report summary from Statista, promotion on social media is currently the second largest form of digital advertising, and the annual U.S. revenues in this market are anticipated to increase by at least 70% by 2027. Nevertheless, it is important to note that while online sharing greatly contributes to these types of figures, organizations must publish material that their customers deem to be shareable. Because a bulk of online revenues are dependent on audience viewership or readership, it would be ideal for companies to make projections over the circulation of content before issuing it to the public.

Online popularity is a concept that is guided by a diverse assortment of features and characteristics. In the case of news editorials that are posted on the internet, their inherent virality tends to be affected by the specific topics that they address. More precisely, articles pertaining to economics, politics, technology, or science may have different measures of popularity than those concentrating on sports, entertainment, popular culture, or lifestyle. In addition, the amount of redistribution an article receives may be contingent on the day of the week, or even the exact time of day, it is released. Apart from these contextual attributes, the purely technical aspects of a given

article may impact its online shareability. These elements can include the article's word length, headlines, and layout, along with the presence and prevalence of attachments such as images, videos, and hyperlinks (Petrova et al., 2021). In conjunction to popularity, online authors, editors, and publishers are interested in subjectivity and polarity metrics as they provide vital metrics related to user sentiment towards their articles.

Using the "Online News Popularity" data set from the UCI Machine Learning Repository (Fernandes et al., 2015), which contains records covering a diverse range of features on stories written on the website Mashable, the question of employing these features to forecast the traffic and circulation of articles online will be explored. Due to the fact that the number of shares a digital article gets is a continuous numerical quantity with a high level of variation, predictive regression models will be constructed during the investigation into this problem statement. In total, five different regression models will be trained, tuned, and evaluated against one another. The chief three among these will focus on multiple linear regression, lasso linear regression, and ridge linear regression. For further comparative analysis, a linear regression model with principal component analysis applied, and a simple regression tree model, will be built. As a result of running the necessary data mining and machine learning techniques on this data set, the objective will be to determine which combination of features has the greatest effect on the estimated popularity of online articles, and also assess whether or not regression stands as a practical approach to making these estimates.

Original Data Set:

The UCI Machine Learning Repository's "Online News Popularity" data set consists of 39,644 online news and culture articles published between January 7th, 2013 and December 27th, 2014 on Mashable. Founded in 2005, Mashable is a digital media and editorial company that writes and posts current events articles which center around several distinct categories, such as lifestyle, entertainment, business, social media, technology, and world news. With the sample space of the raw data set spanning 39,644 rows, it also comprises a feature space that extends to 61 columns. Among these is the nominal categorical variable `url`, which trivially lists the Mashable URLs used to access the sampled articles. Even though there are no other explicit categorical variables in this data set, it should be noted that the subjects of the articles as well as the days of the week on which they are published to the Mashable site have already been one-hot encoded to binary dummy variables. Because this data set is concerned with online news popularity, the target variable `shares` stands for the total number of shares an article receives.

The numerical features, of which there are 46 in the initial data set, capture a wide array of measurements and values that relate to every listed article. For example, `n_tokens_title` denotes the number of words in the title, `n_tokens_content` the number of words in the article, `average_token_length` the average word length, and `num_keywords` the number of keywords in the article. In regard to article attachments, `num_hrefs` and `num_self_hrefs` indicate the numbers of referred links and Mashable articles, while `num_imgs` and `num_videos` signify the numbers of images and videos displayed. Moreover, the variables `title_subjectivity` with `title_sentiment_polarity`, `global_subjectivity` with `global_sentiment_polarity`, and `avg_positive_polarity` with `avg_negative_polarity` allow Mashable creators gauge user engagement and attitude towards their articles. Note that after going through the preprocessing stage of the data mining pipeline, these numerical features along with the binary dummy variables mentioned before are utilized to train the selected regression models.

Data Preprocessing:

The "Online News Popularity" data set is complete, which means that it has no missing records. This is advantageous for the data preparation phase, because it ensures that no imputation code needs to be written and executed for null data, and that no columns will altogether be eliminated due to an abundance of missing values. In spite of the completeness of the data set, there are a few processing steps that need to take place before passing it along to the modeling section. Regarding the `url` variable, every online Mashable article has a unique URL which contains its own distinguishing words or phrases. Since the purpose of this project is to determine how online news circulation varies with respect to features that are common to a large set of articles as opposed to a very small group of them, the entire `url` column is removed. The dummy variables `weekday_is_saturday` and `weekday_is_sunday` respectively denote whether or not an article is published on a Saturday or Sunday, and `is_weekend` designates is released over the weekend. Given that this latter variable is redundant and thereby collinear with the former two features, it is removed from the data set. Looking closer into the data, while there are no missing values, there are several observations which have been misrecorded and consequently dropped from the data set, such as those labeled as having no words and having unique word rates greater than one.

After visualizing the distributions for all numerical features, the ones that are highly positively skewed are dropped. These include nine keyword sharing columns as well as three self-reference sharing columns. Furthermore, it is evident that natural language processing has been done on this data set as a result of latent Dirichlet allocation (LDA) proportions being available for every article. These variables are removed from the data set, as this topic is beyond the scope of this investigation. Because the focus of this study is on regression, it is vital to avoid multicollinearity between the numerical predictors in order to reduce the variance of the calculated coefficients in the trained models. For this project, if two numerical features have a correlation coefficient with a magnitude of at least 0.50, then this indicates a moderate to strong linear relationship between them. In the computed correlation matrix for the remaining numerical features in the data set, there are 19 pairs of quantitative variables with correlation coefficients that fall within this threshold. With the removal of all the collinear columns that are in the data set, the final variable space spans 28 features. The 15 purely numerical features among these are scaled via standardization right before passing them into the regression models.

The target variable for the "Online News Popularity" data set is `shares`. As its name suggests, this quantity denotes the total number of shares that a Mashable article, editorial, or story receives from its readers. In the context of the problem statement being explored for this paper, the number of online shares that an article gets is used as the foundation for its popularity, in that articles with higher numbers of shares are understandably considered to be more popular than those with lower numbers of shares. The distribution of the `shares` column has a very high positive skew such that discounting all of the upper outliers would raise the symmetry displayed for the remaining values. Neither standardization nor normalization would have any effect on the shape of the raw `shares` data. However, taking the natural logarithm of every value in this column removes all outliers and makes this variable symmetric. Additionally, the natural logarithm of the number of shares adheres to the same basis as the initial target column, as an article with a higher natural logarithm of shares is fundamentally more popular than one with a lower natural logarithm of shares. As a result, `ln_shares` denotes the natural logarithm of `shares`, and is now the new target variable for the preprocessed data set.

Feature Exploration & Visualization:

Before dividing the fully preprocessed data into training, validation, and testing sets, fitting the five chosen regression models, and evaluating the performance of these models, some preliminary visualization of the dependent variable `ln_shares` with respect to the included independent predictor variables is required. Creating these visualizations helps in gaining an early impression over how `ln_shares` varies with each individual feature, and also whether or not there exists a potential discernible relationship between the two. For the one-hot encoded categorical features, side-by-side box plots showing the distribution of `ln_shares` according to each category are displayed. On the other hand, scatter plots are made to present the joint distributions between the target variable `ln_shares` and all of the continuous numerical features within the data set.

As discussed in a prior section of this paper, the two categorical variables which have already been converted to binary dummy variables relate to the main subject of a given Mashable article and the day of the week on which it is posted online. Looking at the former categorical feature first, its encoded dummy variables are designated as `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`. With 1 and 0 meaning "yes" and "no," respectively, these variables indicate whether or not an article is a lifestyle, entertainment, business, social media, technology, or world news piece. After grouping the values in the `ln_shares` column based on these article subjects, the following distributions are yielded:

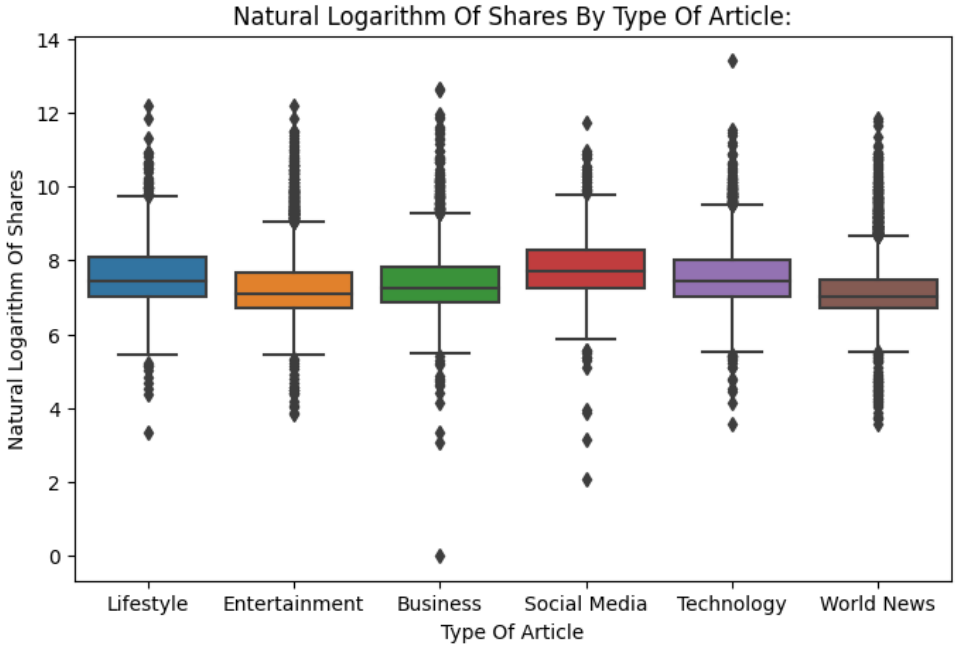

Figure 1: Side-by-side box plots for the distributions of the natural logarithm of shares by the type of article published

According to Figure 1 above, the distributions of the natural logarithm of shares for all the article types are roughly symmetric and similar to one another. While the business box plot appears to be spread over the widest range on account of its noticeable lower outlier, the interquartile ranges for every article subject look almost identical to each other. The median natural logarithm of shares

for lifestyle and social media articles are the highest, but the low inconsistency between these distributions does not immediately suggest which types of articles are more or less popular.

There are naturally seven different binary dummy variables for the day of the week on which an article is released on Mashable, and they are identified as `weekday_is_sunday`, `weekday_is_monday`, `weekday_is_tuesday`, `weekday_is_wednesday`, `weekday_is_thursday`, `weekday_is_friday`, and `weekday_is_saturday`. Implementing the same code on these variables as on the preceding set of dummy features, the side-by-side box plots of the natural logarithm of shares by day of week for article publication are as follows:

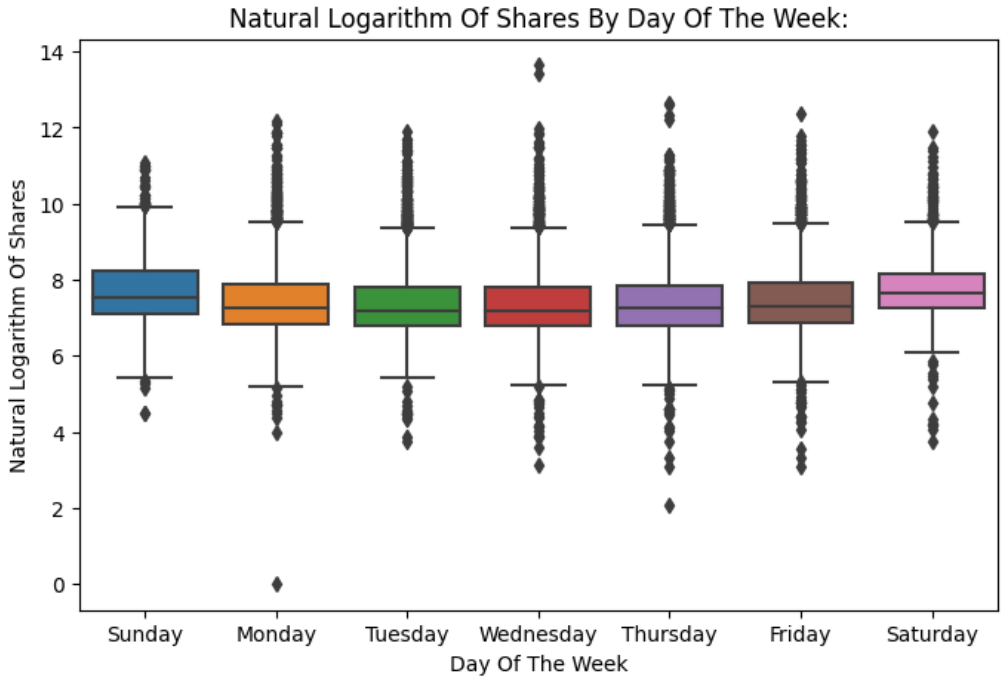

Figure 2: Side-by-side box plots for the distributions of the natural logarithm of shares by the day of the week on which the article is published

Much like the distributions provided in the first figure, the ones given in Figure 2 also all look symmetric and analogous to one another. The Sunday box plot has the shortest range, and Monday's box plot has the longest range as a result of is prominent lower outlier. This is likely the same outlier which increased the range of the business box plot seen in Figure 1. Nonetheless, the interquartile ranges for every day of the week are all nearly equivalent to each other. A noteworthy observation of these side-by-side box plots is that the median natural logarithm of shares for Monday, Tuesday, Wednesday, Thursday, and Friday look to be same, while the median values for Sunday and Saturday are slightly greater. A possible explanation for this discrepancy is that users are more likely to spend more time online over the weekend, and are resultantly more likely to share more of what they see on the Internet during this time.

Moving on to the numerical features of the preprocessed data set, the first one to inspect is `timedelta`, which stands for the number of days between an article's publication and the procurement of the original "Online News Popularity" data set. Because the collection of articles in the data set spans close to a two-year time frame, the minimum and maximum values in the `timedelta` column are 8 days and 731 days, respectively. In Figure 3 on the next page, there is evidently no correlation between the `ln_shares` target variable and this numerical predictor:

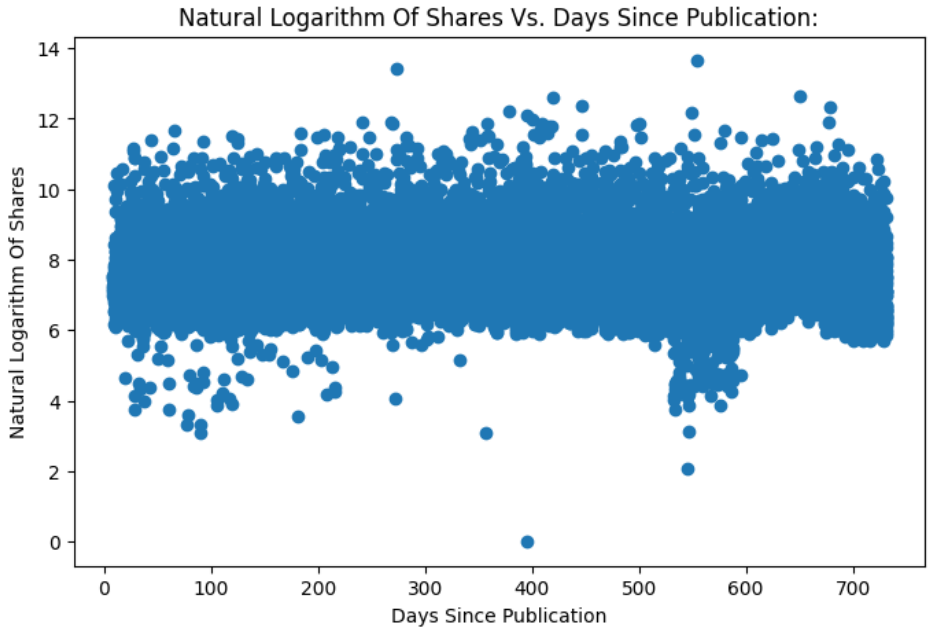

Figure 3: Scatter plot for the natural logarithm of shares with respect to the number of days since article publication

For the remaining numerical features, Figure 4 below shows the joint distributions of `ln_shares` with respect to the work and token data, and Figures 5 and 6 on the subsequent page plot this target variable against the article attachment features and the user engagement features, respectively. Similar to the scatter plot of the natural logarithm of shares versus the number of days since the publication of the articles, none of those displayed in the next three images demonstrate clear correlations or linear relationships, even though some of them, especially the ones in Figures 4 and 5 do share visual likenesses between one another:

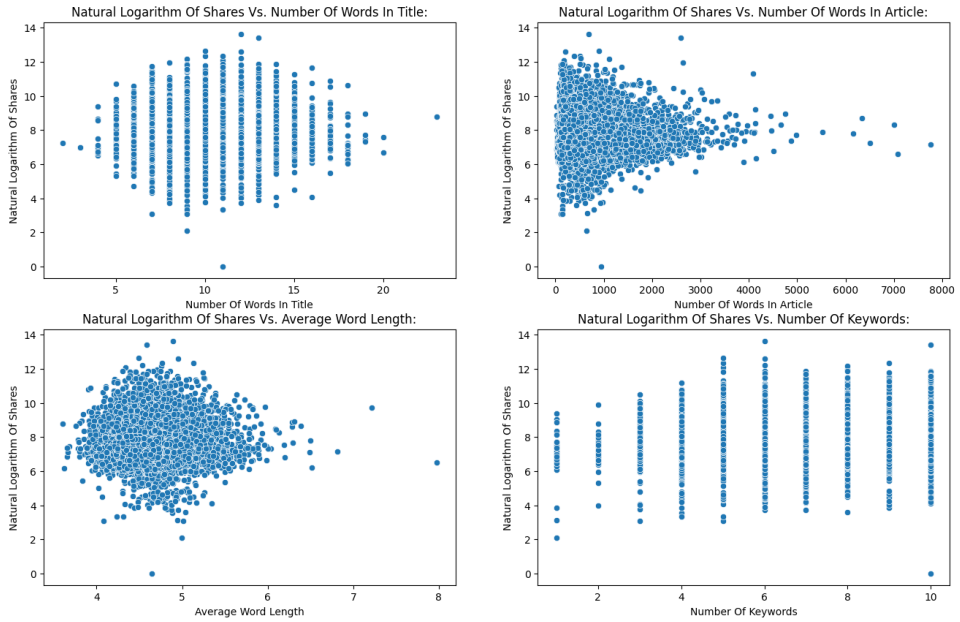

Figure 4: Scatter plots for the natural logarithm of shares with respect to number of words in the title, number of words in the article, average word length, and number of keywords

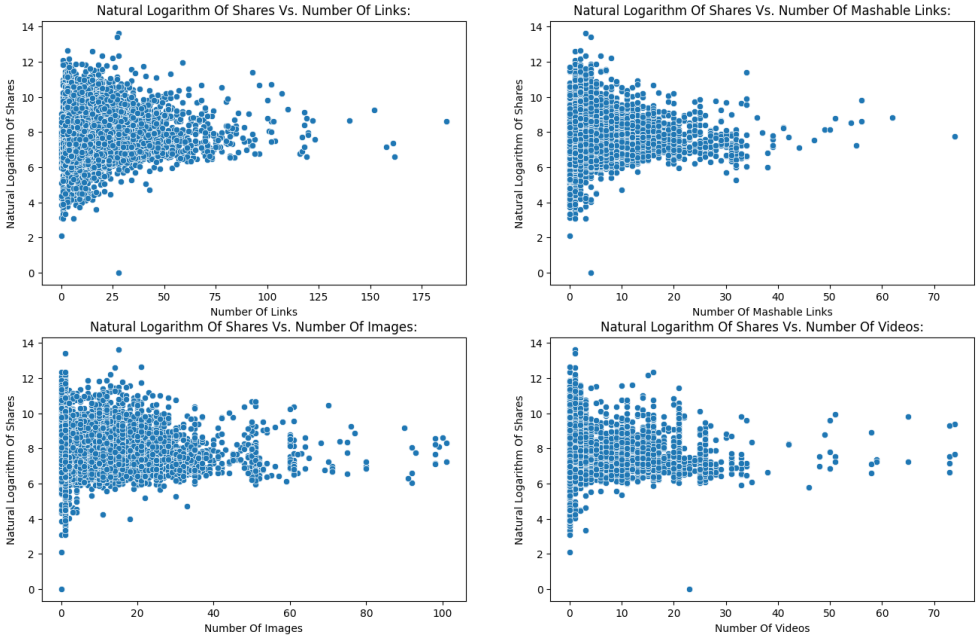

Figure 5: Scatter plots for the natural logarithm of shares with respect to number of links, number of Mashable links, number of images, and number of videos

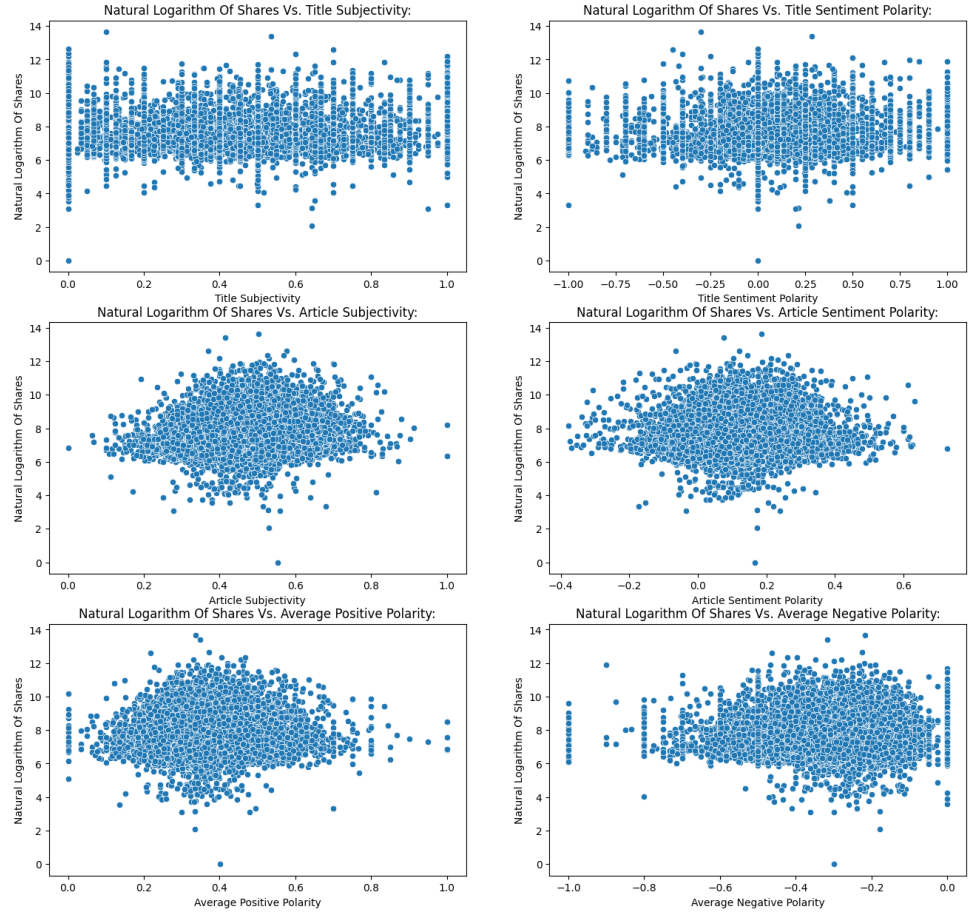

Figure 6: Scatter plots for the natural logarithm of shares with respect to title subjectivity and sentiment polarity, article subjectivity and sentiment polarity, and average positive and negative polarity

Multiple, Lasso, & Ridge Linear Regression:

With all the features explored and the natural logarithm of shares visualized with respect to each one, the available data can now be utilized to develop the regression models for estimating the popularity of Mashable's online news articles. The preprocessed data is separated such that 60% of it is used for training, 20% is used for validation, and 20% is used for testing. The 15 continuous numerical features in the training set are all scaled through standardization, and the same code is then applied to the validation and testing sets. Multiple linear regression is a statistical modeling technique in which a set of independent numerical variables are employed to predict the value of a dependent variable. If $\hat{y}$ denotes the predicted value of the target variable, x_i is a numerical feature input and β_i is its coefficient, and there are n numerical features in total, then the standard multiple linear regression equality is expressed as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{1}$$

where $i \in \mathbb{Z}$ such that $1 \leq i \leq n$. Here, β_0 symbolizes the bias that exists in the multiple linear regression model, and is also known as the y-intercept. Moreover, the ε term is the irreducible error observed in the model. If the second summation term is expanded, then the simple multiple linear regression equation can also be defined by the following:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \varepsilon \tag{2}$$

While lasso and ridge both take the fundamental form of multiple linear regression equalities, these techniques bring in regularization and penalty components to their models. With feature selection embedded in the lasso and ridge linear regression, the parameters with lower importance are penalized, and practically excluded from the models.

As is known, the subject of this study is centered around predicting the popularity of articles published online based on the natural logarithm of shares that it gets. The feature space for this problem has a dimensionality $n = 28$ such that 15 are continuous numerical input parameters while the remaining 13 are binary dummy variables. For independent variables p_i and their estimated coefficients γ_i, if $\hat{s}$ denotes the projected number of online shares, then:

$$\ln(\hat{s}) = \gamma_0 + \sum_{i=1}^{28} \gamma_i p_i + \delta \tag{3}$$

where $i \in \mathbb{Z}$ such that $1 \leq i \leq 28$, γ_0 is the bias, and δ is the irreducible error. Just like in Equation 2 above, the formula below provides the expansion of Equation 3 to put the natural logarithm of the number of shares in terms of the simple multiple linear regression model:

$$\ln(\hat{s}) = \gamma_0 + \gamma_1 p_1 + \gamma_2 p_2 + \gamma_3 p_3 + \cdots + \gamma_{28} p_{28} + \delta \tag{4}$$

Because the natural logarithm of shares is proposed to be adhering to a linear regression model, a critical postulate of this statement is that the number of shares varies exponentially as follows:

$$\hat{s} = e^{\gamma_0 + \gamma_1 p_1 + \gamma_2 p_2 + \gamma_3 p_3 + \ldots + \gamma_{28} p_{28} + \delta} \quad (5)$$

For the multiple, lasso, and ridge linear regression models, repeated k-fold cross-validation with the number of folds set to 10 and the number of repeats set to 3 is implemented. Beginning with simple linear regression, the built-in model with recursive feature elimination (RFE) is fit to the scaled training set data. Using grid-search cross-validation for hyperparameter tuning with the root mean squared error (RMSE) as the scoring parameter, the following chart displays this metric for the training and testing results against the number of selected features:

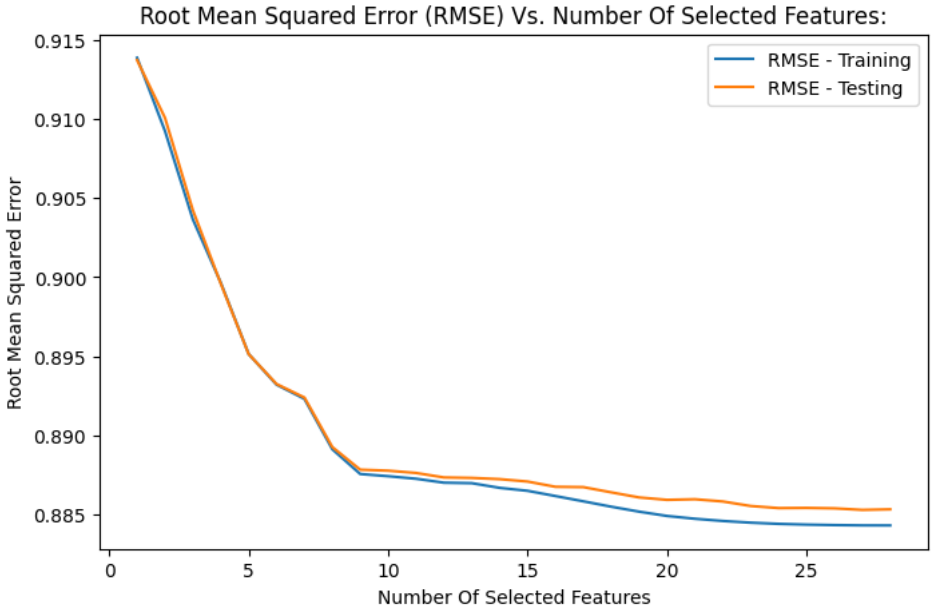

Figure 7: Root mean squared errors (RMSE) of the training and testing sets versus the number of selected features for the multiple linear regression model

According to the graph in Figure 7 above, incorporating all 28 numerical and categorical in the multiple linear regression model minimizes the root mean squared error for both the training and testing results, and is helpful for comparatively narrowing down which ones hold the greatest weight. After retraining the multiple linear regression model with all 28 features selected and passing the validation and testing sets into it, the resultant performance metrics are returned:

Table 1: Performance metrics for the multiple linear regression model

Error Metric	Validation Set	Testing Set
Root Mean Squared Error (RMSE)	0.899280	0.873873
Mean Squared Error (MSE)	0.808704	0.763655
Mean Absolute Error (MAE)	0.660506	0.654707
Mean Absolute Percentage Error (MAPE)	0.086735	0.086347

Note that for all of the regression methods discussed in this paper, the root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are taken as the performance metrics. The differences between the actual observations and the predicted values from the validation and testing sets are also calculated and visualized in histograms like the two in Figure 8 below:

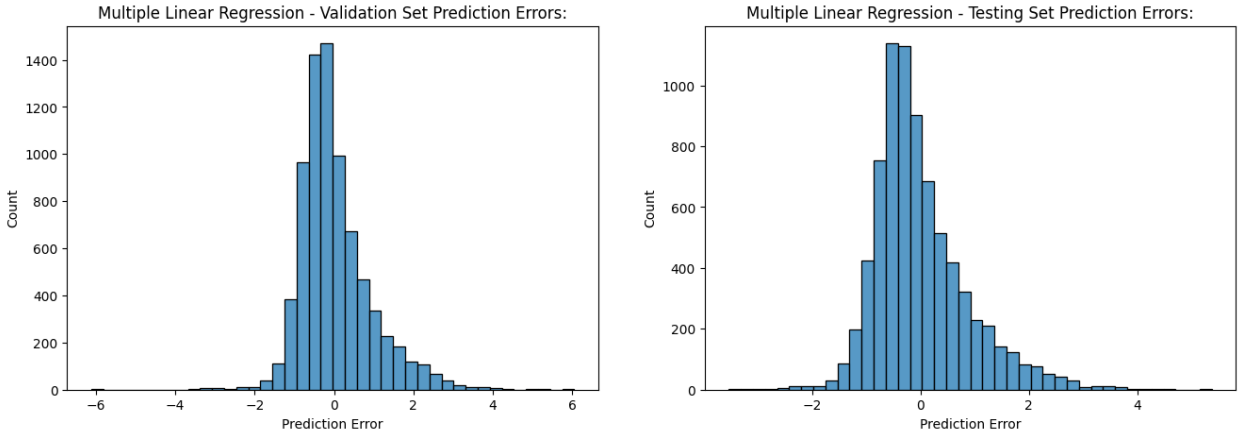

Figure 8: Prediction errors for the multiple linear regression model on the validation and testing sets

It is clear that the distributions of the prediction errors on the multiple linear regression model for both the validation and testing sets are unimodal, approximately centered around zero, and largely symmetric with that for the testing set having slightly more of a positive skew. Tables 2 and 3 rank the five most significant numerical features and dummy variables in the multiple linear regression model in regard to the magnitudes of their weights:

Table 2: Five most important numerical features for the multiple linear regression model

Rank	Numerical Feature	Weight
1	num_hrefs	0.0661
2	global_subjectivity	0.0491
3	num_imgs	0.0347
4	num_self_hrefs	– 0.0334
5	num_keywords	0.0323

Table 3: Five most important dummy variables for the multiple linear regression model

Rank	Dummy Variable	Weight
1	data_channel_is_world	– 0.4964
2	data_channel_is_ entertainment	– 0.4470
3	data_channel_is_bus	– 0.2542
4	weekday_is_saturday	0.2128
5	weekday_is_sunday	0.1823

Moving forward, these will be the five most important numerical features and dummy variables for the lasso and ridge linear regression models as well.

Using the built-in lasso cross-validation function with α values that are defined to be in the set {0.0001, 0.001, 0.01, 0.1, 1, 10}, the predictive performance metrics for the lasso linear regression model are presented in Table 4 on the next page:

Table 4: Performance metrics for the lasso linear regression model

Error Metric	Validation Set	Testing Set
Root Mean Squared Error (RMSE)	0.899308	0.873858
Mean Squared Error (MSE)	0.808755	0.763628
Mean Absolute Error (MAE)	0.660538	0.654730
Mean Absolute Percentage Error (MAPE)	0.086740	0.086351

The prediction errors of the lasso linear regression model on the validation and testing sets have virtually identical distributions to those for the multiple linear regression model:

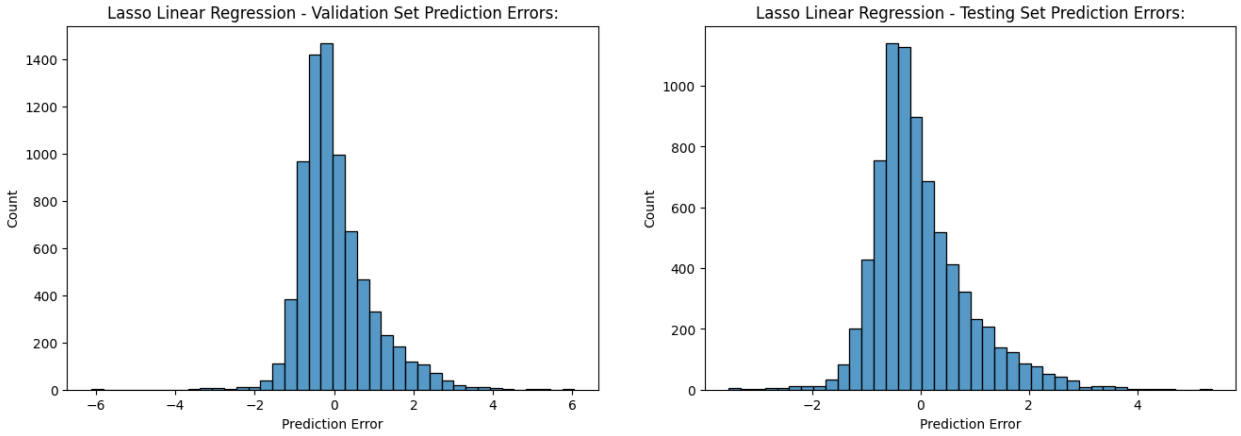

Figure 9: Prediction errors for the lasso linear regression model on the validation and testing sets

In addition, the `data_channel_is_bus` and `weekday_is_saturday` rankings are switched:

Table 5: Five most important numerical features for the lasso linear regression model

Rank	Numerical Feature	Weight
1	`num_hrefs`	0.0663
2	`global_subjectivity`	0.0493
3	`num_imgs`	0.0349
4	`num_self_hrefs`	– 0.0333
5	`num_keywords`	0.0322

Table 6: Five most important dummy variables for the lasso linear regression model

Rank	Dummy Variable	Weight
1	`data_channel_is_world`	– 0.4922
2	`data_channel_is_entertainment`	– 0.4431
3	`weekday_is_saturday`	0.2617
4	`data_channel_is_bus`	– 0.2500
5	`weekday_is_sunday`	0.2314

Similar to the code that is used for the lasso linear regression model, a ridge cross-validation function with $\alpha \in \{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ is run in order to generate the best-performing ridge linear regression model. The root mean squared errors (RMSE), mean squared errors (MSE), mean absolute errors (MAE), and mean absolute percentage errors (MAPE) for the validation and testing sets on this ridge linear regression model are listed in Table 7 below:

Table 7: Performance metrics for the ridge linear regression model

Error Metric	Validation Set	Testing Set
Root Mean Squared Error (RMSE)	0.899325	0.873843
Mean Squared Error (MSE)	0.808785	0.763602
Mean Absolute Error (MAE)	0.660554	0.654725
Mean Absolute Percentage Error (MAPE)	0.086350	0.086350

These performance for the ridge linear regression model resemble those computed for both the multiple linear regression model and the lasso linear regression, albeit with some very minor and negligible disparities. The same observation holds true for the prediction errors of the ridge linear regression model after feeding it with the separate data from the validation set and the testing set. Specifically, Figure 10 underneath shows validation and testing set prediction error distributions that are nearly indistinguishable from the earlier histograms in Figures 8 and 9 for the multiple linear regression and lasso linear regression models, respectively:

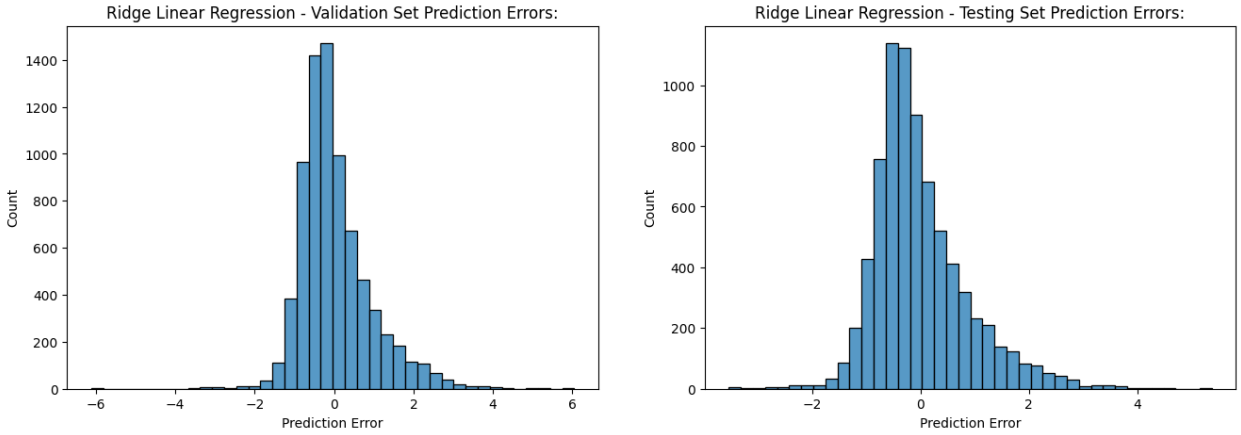

Figure 10: Prediction errors for the ridge linear regression model on the validation and testing sets

Expectedly, the five most important numerical features are `num_hrefs`, `global_subjectivity`, `num_imgs`, `num_self_hrefs`, and `num_keywords`, and the five most important dummy variables for this third model are `data_channel_is_world`, `data_channel_is_entertainment`, `data_channel_is_bus`, `weekday_is_saturday`, and `weekday_is_sunday`:

Table 8: Five most important numerical features for the ridge linear regression model

Rank	Numerical Feature	Weight
1	num_hrefs	0.0664
2	global_subjectivity	0.0495
3	num_imgs	0.0352
4	num_self_hrefs	– 0.0334
5	num_keywords	0.0323

Table 9: Five most important dummy variables for the ridge linear regression model

Rank	Dummy Variable	Weight
1	data_channel_is_world	– 0.4894
2	data_channel_is_ entertainment	– 0.4408
3	data_channel_is_bus	– 0.2478
4	weekday_is_saturday	0.2116
5	weekday_is_sunday	0.1816

Figure 11 below displays how the coefficients of the predictors for the simple multiple, lasso, and ridge linear regression models all compare with one another. For the numerical features the weights are all essentially the same. The article subject dummy variables show identical coefficients for multiple and lasso linear regression, with slightly increased coefficients in ridge linear regression. Meanwhile, the day of the week dummy variables present equivalent weights for multiple linear regression and ridge linear regression, but the coefficients of this latter set of dummy variables for lasso linear regression are markedly greater than those for the former models:

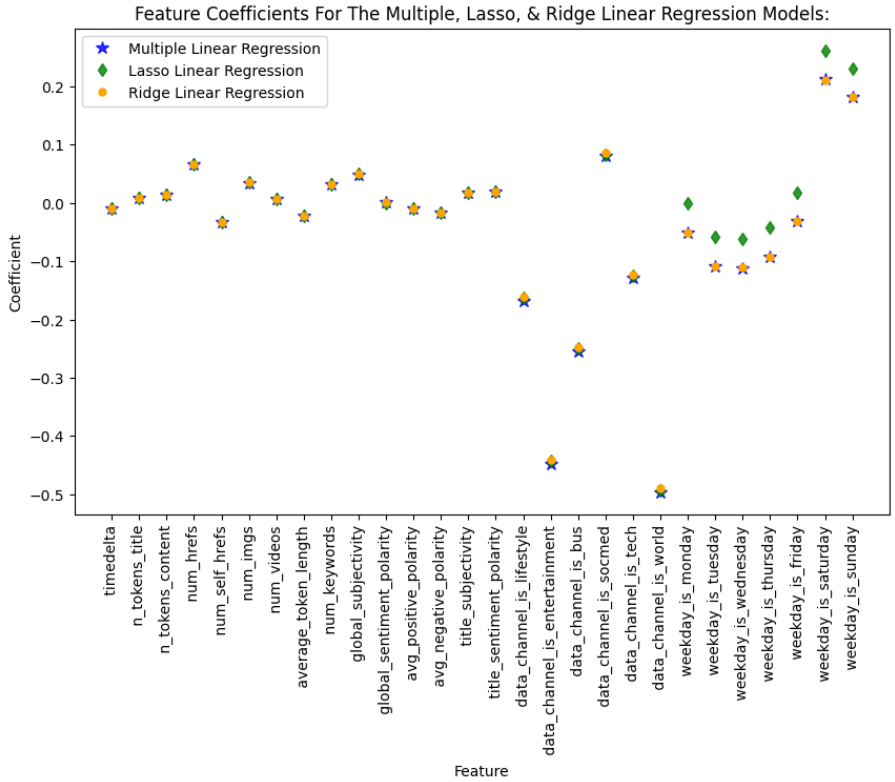

Figure 11: Coefficients for the numerical features and dummy variables in the multiple, lasso, and ridge linear regression models

Principal Component Analysis Linear Regression:

In the data preprocessing stage of this paper, the dimensionality of the "Online News Popularity" data set is substantially reduced from 61 columns to 29 columns. If principal component analysis (PCA) is applied to the preprocessed data set, then the column space can be reduced even further. Based on the following chart in Figure 12, the root mean squared error (RMSE) is at a minimum when the number of principal components is equal to 28.

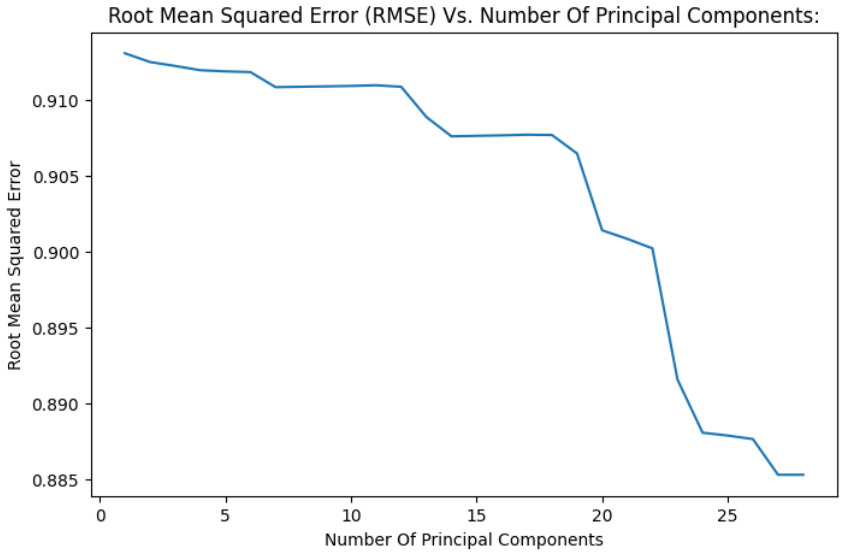

Figure 12: Root mean squared error (RMSE) versus the number of principal components for the principal component analysis regression model

However, as illustrated in the supplementary code and the graph below, the increase in the cumulative explained variance after 14 principal components is only marginal and gradual:

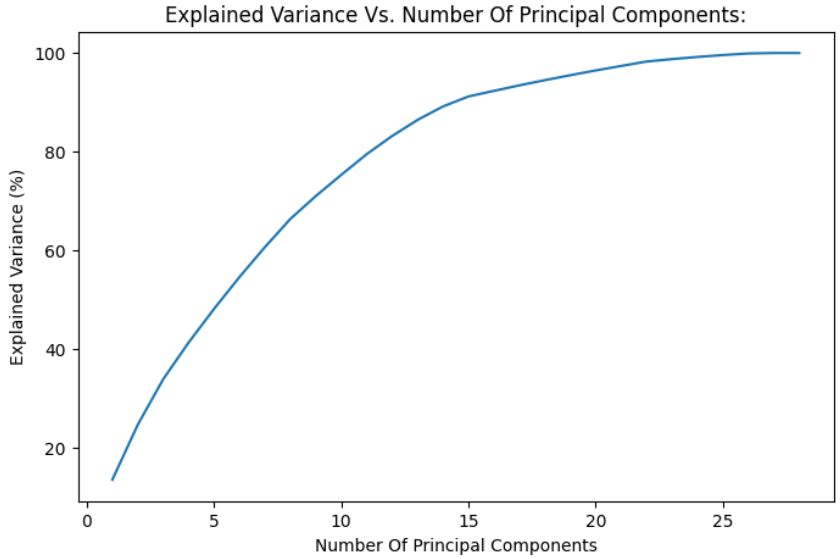

Figure 13: Explained variance versus number of principal components for the principal component analysis linear regression model

As determined in the attached code, the total explained variance for 14 principal components is about 89.19%. A built-in principal component analysis function with the number of principal components set to 14 is fit on the training set in which only the numerical features are scaled according to standardization. A multiple linear regression is then trained on the PCA-adjusted data set, and subsequently evaluated on the validation and testing sets:

Table 10: Performance metrics for the principal component analysis linear regression model

Error Metric	Validation Set	Testing Set
Root Mean Squared Error (RMSE)	0.921202	0.891678
Mean Squared Error (MSE)	0.848614	0.795089
Mean Absolute Error (MAE)	0.684914	0.676192
Mean Absolute Percentage Error (MAPE)	0.090172	0.089411

Table 10 above tabulates the root mean squared errors (RMSE), mean squared errors (MSE), mean absolute errors (MAE), and the mean absolute percentage errors (MAPE) of the principal component analysis linear regression model on the preprocessed validation and testing sets. Looking at these performance metrics for the principal component analysis linear regression model, they are greater than those found for the earlier multiple linear regression, lasso linear regression, and ridge linear regression models. Figure 14 below provides histograms of the prediction errors of this model for the validation set on the left and the testing set on the right:

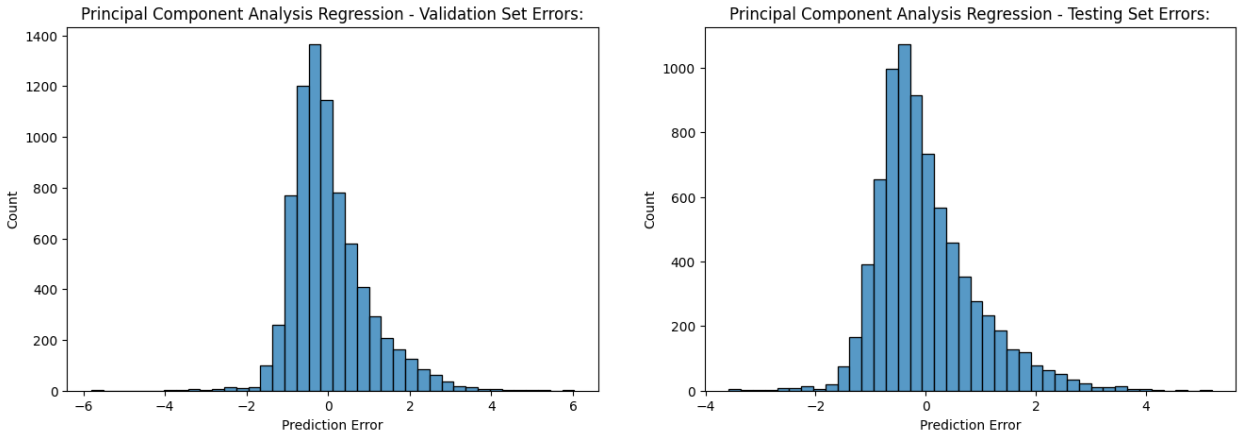

Figure 14: Prediction errors for the principal component analysis (PCA) linear regression model on the validation and testing sets

In spite of the slightly higher error metrics for principal component analysis linear regression, the prediction errors for the principal component analysis linear regression model on the validation and testing sets practically have the same unimodal, slightly positively skewed distributions centered around zero as those shown in the histograms for the multiple, lasso, and ridge linear regression models in the earlier Figures 8, 9, and 10.

Regression Tree Model:

For further comparison, a regression tree model is built to predict the natural logarithm of the number of shares. The maximum depth of the regression tree is set to four levels for simple and computationally inexpensive results, with the mean squared error as the scoring metric:

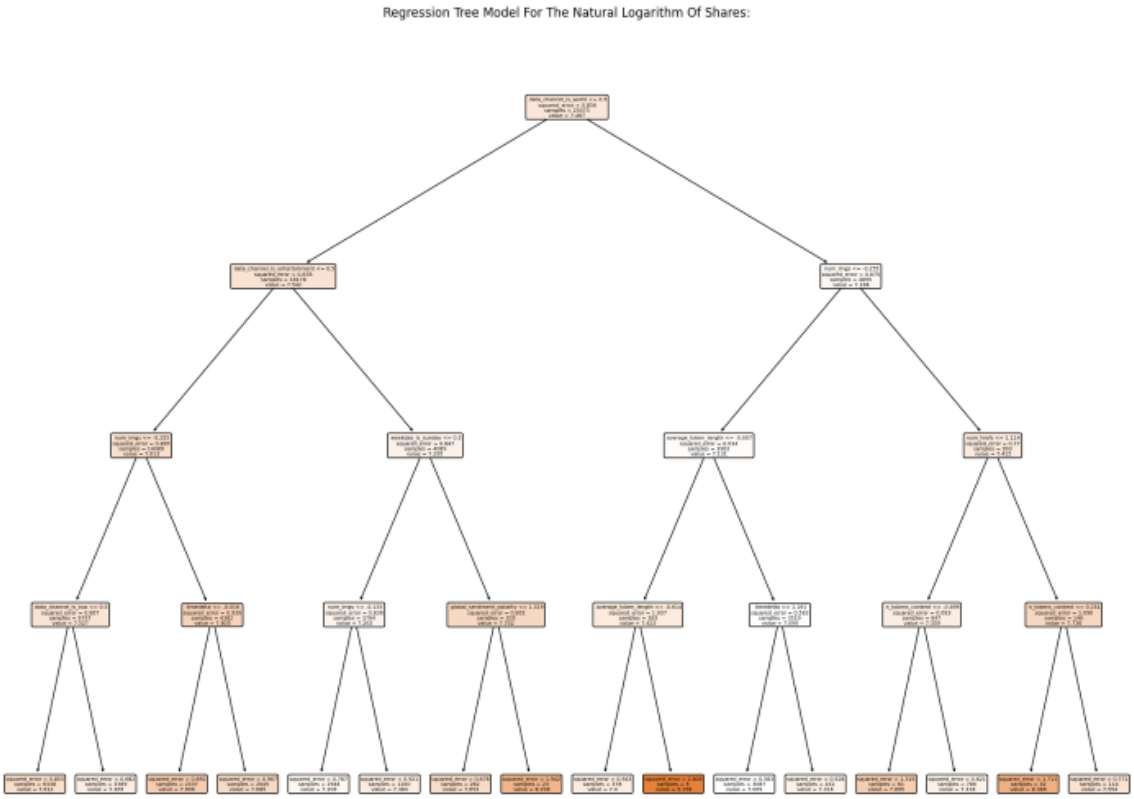

Figure 15: Regression tree model for predicting the natural logarithm of shares

The performance metrics of this model on the validation set and the testing set are as follows:

Table 11: Performance metrics for the regression tree model

Error Metric	Validation Set	Testing Set
Root Mean Squared Error (RMSE)	0.908574	0.885044
Mean Squared Error (MSE)	0.825506	0.783303
Mean Absolute Error (MAE)	0.668887	0.665067
Mean Absolute Percentage Error (MAPE)	0.087855	0.087807

Figure 16 has the distributions of this model's prediction errors on the validation and testing sets:

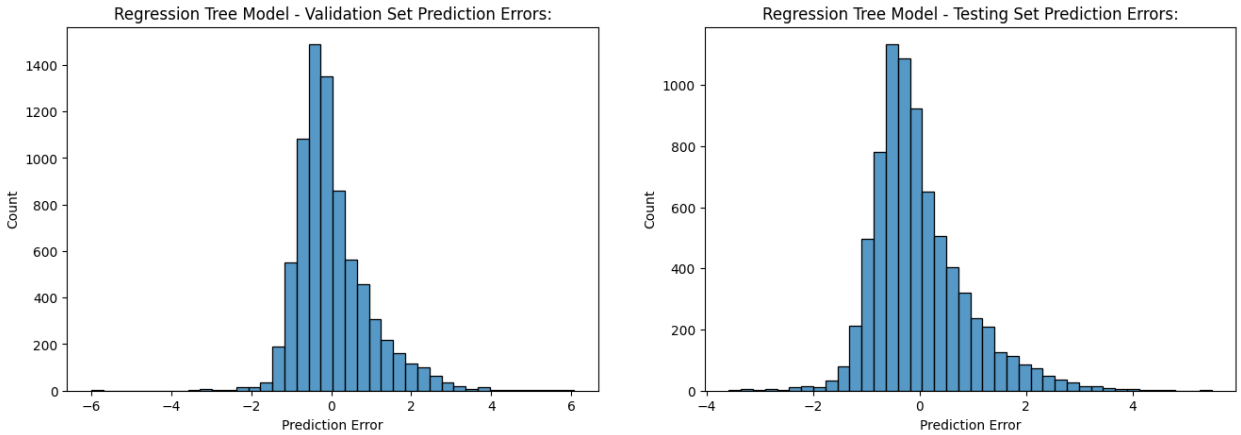

Figure 16: Prediction errors for the regression tree model on the validation and testing sets

Conclusion:

The preceding paper examined the "Online News Popularity" data set (Fernandes et al., 2015) of 39,644 editorials published between January 7th, 2013 and December 27th, 2014 on the website Mashable with the aim of determining if regression is a feasible machine learning method for predicting the circulation of articles posted online. This original collection from the UCI Machine Learning Repository first entered the preprocessing step of the data mining pipeline, and was ultimately able to be reduced to a data set with 38,456 rows and 29 columns consisting of 15 numerical features, 13 dummy variables. The target variable of this data set was changed from the total number of shares that a given Mashable article receives to the natural logarithm of these shares. With the numerical features scaled by means of standardization, this preprocessed data was split such that 60% of it made up the training set, 20% the validation set, and the remaining 20% the testing set. The data from the training set was utilized to fit five separate regression models, in which the main three of interest were multiple, lasso, and ridge linear regression models. For additional comparison a linear regression model with principal component analysis (PCA) applied was built. Also, while not fully practical, a simple, computationally inexpensive regression tree model was made. These models were then evaluated on the validation and testing sets, with their performance measures calculated and displayed.

For the multiple, lasso, and ridge linear regression models, the root mean squared errors (RMSE), mean squared errors (MSE), mean absolute errors (MAE), and mean absolute percentage errors (MAPE) are all virtually equivalent. The three sets histograms showing the distributions of the prediction errors for the models on the validation and testing data are also essentially the exact same images. Furthermore, all three of these models have the numbers of links, Mashable links, images, and keywords along with the global subjectivity as the most important numerical features, as well as whether or not the article is a world news, entertainment, or business story along with if it is released on a Saturday or Sunday as the most important dummy variables. For the principal component analysis linear regression model with 14 set as the number of principal components, the error metrics are evidently greater on both the validation and testing sets. On the other hand, the error metrics from the regression tree model with a maximum depth of four levels are less than those from the principal component analysis linear regression model, but still greater than those from the first three models. Even so, the prediction error distributions for these two models are almost the same as the first three.

While there is consistency in the performance results across all of the regression models explored in this study, there is still much bias in them. The multiple, lasso, and ridge linear regression models each have combined bias and irreducible error values that are equivalent to approximately 7.766, 7.712, and 7.761, respectively. In addition, the principal component analysis linear regression model only manages to bring the combined bias and irreducible error to about 7.467. With this in mind, there are definitely a number of other ways for which the found data could have been prepared and preprocessed before passing it into the regression models in order to significantly minimize the bias that would emerge from these techniques. Perhaps using different feature scaling or even keeping the number of shares as the target variable as opposed to its natural logarithm would produce better, unbiased models. It is even possible that the popularity of online publications varies nonlinearly by polynomial or exponential functions. But despite these shortcomings, if an organization does want to make accurate quantifiable projections of how well their online material will circulate, the real-world solutions to this problem will ultimately lie in models that are fundamentally constructed around regression. Moreover, the question of whether or not certain material is in fact deemed "popular" is an issue of classification which can be tied back into these actual regression models.

References & Data Sources:

Esteban Ortiz-Ospina (2019) - "The rise of social media" Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/rise-of-social-media [Online Resource].

Fernandes, Kelwin; Vinagre, Pedro; Cortez, Paulo; and Sernadela, Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V.

Petrova, A. (2021). The Anatomy of Top Performing Articles: Successful vs. Invisible Content [Semrush Study]. Semrush. [Research Article]. Available at https://www.semrush.com/blog/anatomy-of-top-performing-articles/.

Pew Research Center. (2023). Digital News Fact Sheet. [Research Article]. Available at https://www.pewresearch.org/journalism/fact-sheet/digital-news/.

Richards, D. (2023). Shareable Content: What Is It? & Why Is It So Important?. [Online Post]. Available at https://tuckerhall.com/shareable-content-important/#:~:text=Shareable%20content%20can%20lead%20to,Social%20validation.

Shearer, E. (2021). More than eight-in-ten Americans get news from digital devices. Pew Research Center. [Research Article]. Available at https://www.pewresearch.org/short-reads/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/.

Statista. (2023). Social Media Advertising: market data & analysis. [Research Article]. Available at https://www.statista.com/study/36294/digital-advertising-report-social-media-advertising/#:~:text=The%20Social%20Media%20Advertising%20market,What's%20included%3F