Discovering Factors Leading to Popular Online News

Yuxin Zhang[1], Jizhou Kang[1]
Department of Statistical Science, UC Santa Cruz[1]

Abstract

The past few decades have witnessed a fast expansion of digital journalism. People nowadays read, create, and share articles on the Internet. How can journalists optimize their contents to receive more views and shares from audiences? Our project aims to answer this question by exploring a dataset collected from Mashable news. The dataset consists of features and number of shares for each sample online news. We use simple linear regression models and logistic regression models to get robust estimation on features' effect on the popularity of online news. A careful model selection is performed using AIC as criteria. The result shows that online news with strong title, positive attitude, more visuals and simple format tend to be more popular, and the popularity differs by the channels and published day.

KEY WORDS: Linear Regression, Logistic Regression, Model Selection, Online News.

1. Introduction

In the era of the Internet, online news have become people's primary source of information. We believe that analyzing important features could help companies better understand audiences' preferences and catch up with the current trends. In this project, our goal is to discover features that lead to popular online news. We used a widely used dataset with over 39000 articles from Mashable website and performed both simple linear regression and logistic regression

Some researchers have already done similar studies using this dataset. *K. Fernandes et.al [1]*, the creators of this dataset, performed an Intelligent Decision Support System (IDSS). This system first predicts if an article becomes popular based on the classification model built from the set of features. Then it optimizes a subset of features that can be easily changed by searching for an enhancement of the predicted popularity probability. The model with best prediction power is random forest, with a 67% accuracy. Moreover, *Md. Uddin et.al [2]* used the gradient boosting machine to predict shares of articles. Their result shows a 1.8% improvement of gradient boosting machine over the random forest. Similarly, *H. Ren et.al [3]* tried ten classification methods on predicting whether an article is popular or not based on carefully filtered features, resulting in the best prediction accuracy of 70% achieved by random forest model.

Despite the relatively good prediction accuracy the previous studies achieved, they all failed to consider which feature contributes the most regarding the popularity of an article. Meanwhile, none of the research evaluates the impact of interactions between different features on the popularity of articles. With different goals in mind, we instead focus on building models that can indicate which features and interaction terms make online articles more popular. Another captivating point is that this dataset can be analyzed from both regression and classification perspectives. Therefore, to receive more robust results, it is worth the effort to consider both regression and classification models. In conclusion, our goal is to solve the statistically inverse problem of calibrating the coefficients corresponding to features, and the calibration is performed under both regression and classification settings.

We proceed our analysis by three steps: data preprocessing, exploratory data analysis (EDA) as well as modeling and evaluation. In data preprocessing step, we check missing values and outliers, explore suitable transformation for scaled variables and the correlation between variables. The primary goals for EDA are: first of all, showing patterns between individual features and response variables; Then, discovering meaningful interactions between features to be considered in our model and deciding which features should be dropped. Finally, model and model evaluation are the essential part of this project. In regression, we use multiple linear regression models to estimate the effect of features on log number of shares. In classification, we equally partition articles into popular and unpopular groups, then using a logistic regression model to estimate the impact of features on the odds of being popular. In both cases, we perform forward stepwise model selection based on AIC criteria. After select the best model, we deliver several analysis such as out-of-sample testing, assumption checking, and concluding the project based on model estimation result.

The rest of the paper is constructed as follows. We introduce the data preprocessing approach in Section 2. EDA is discussed in Section 3. The main results of models and model evaluation are presented in Section 4. In Section 5 we summarize our significant findings and in Section 6, we discuss the possible future directions of this project.

2. Data Preprocessing

2.1 Data Acquisition

The data we use is originally acquired by *K. Fernandes et.al [1]* for their paper. The authors retrieved an exten-

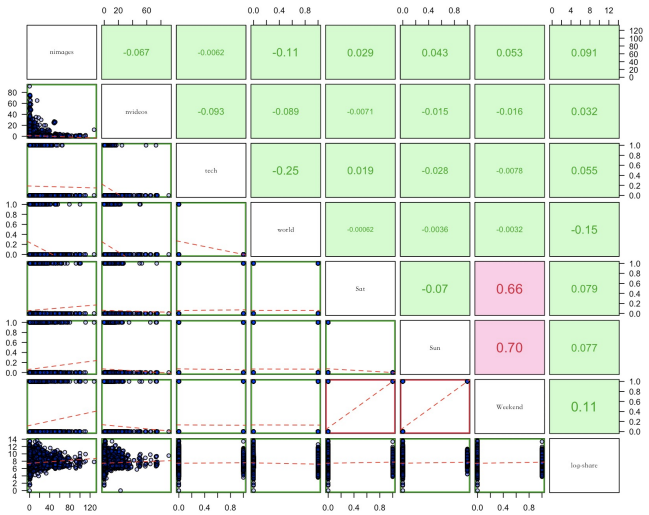
sive set (total 58) features from 39,644 articles published from January 7, 2013 to January 7, 2015 from Mashable, one of the largest news websites. The data is available at UCI Machine Learning repository.

The columns consist of features that describe different aspects of an article and that are considered possibly relevant to influence the number of shares. A full list of variables contained in this dataset is shown in Table 1. All features are divided into six groups: Words, Links, Digital Media, Time, Keywords, Natural Language Processing. In addition to directly acquiring features from the articles, the authors also extracted several natural language processing features by applying the Latent Dirichlet Allocation (LDA) algorithm and the Pattern web mining module. The adoption of these algorithm allows the computation of sentiment polarity, subjectivity scores, etc.

2.2 Missing Value and Outlier

This data does not have any missing values. As for outliers, we detected an unusual observation that contains some extreme features. Features in terms of ratios, such as rate of unique words, corresponding to this article are greater than 500, while it should be between 0 and 1. Also, all features under the natural language processing group are 0. We further track down this outlier to its corresponding Mashable article. There's nothing suspicious. Therefore, we believe this outlier is due to data collection error. Since there is only 1 out of 39644 samples with this odd pattern, we decided to drop it simply.

Figure 1: Example of correlation checking plot. The upper triangular shows the correlation coefficients between corresponding variables. Color intensity are proportional to the correlation coefficients. The lower triangular shows scatter plots.



2.3 Data Exploration

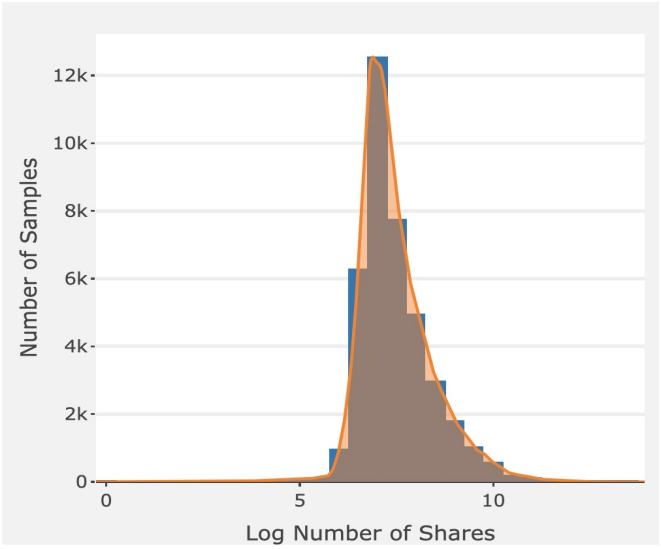
Since we consider the inverse problem, we care about multicollinearity issue because it could lead to biased estimation of coefficients. An ad hoc way of checking multicollinearity is from correlation between variables. Figure 1 gives an example of a plot that checks correlations between independent variables. We also add the response variable in this plot and use the scatter plot at the lower triangular part to decide whether a transformation is required on the variable.

3. Exploratory Data Analysis

3.1 Transforming Variables

The histogram of the response variable shows a highly skewed distribution, ranging from 1 to over 800 thousand. It could potentially cause problems when doing regression. Thus, we replace the original response variable by logarithm of the response variable in regression models. The histogram of log-transformed response variable is shown in Figure 2, from which we can see that a more symmetric pattern is achieved.

Figure 2: Histogram and kernel estimation of logarithm of response variable. The blue bar and orange curve show the histogram and the kernel density, respectively.



Furthermore, from the lower triangular section of Figure 1, we can identify the variables whose scatter plot with response variable are highly skewed. Finally, we find despite *number of words in title*, *average word length* and *number of keywords*, the rest of numeric variables require the log transformation defined by (1).

$$log.covariate = log(covariate + 1) \tag{1}$$

3.2 Categorical Variable Selection

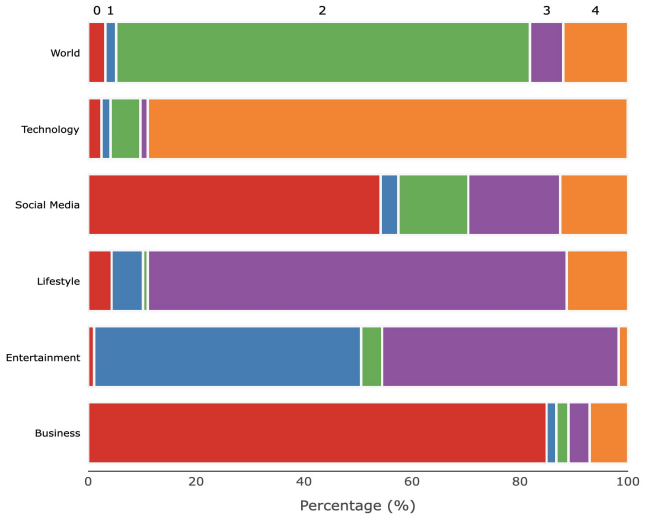
Given a large dataset like this one, we must consider the potential multicollinearity issues. Therefore, we must perform variable selection before fitting any models. We start with variables from the time category. There are

Table 1: List of all variables by category. Number inside the bracket indicate the number of columns in the dataset corresponding to that feature.

Feature	Type	Feature	Type
Words (6)		**Keywords (16)**	
Number of words in title	numeric (1)	Number of keywords	numeric (1)
Number of words in article	numeric (1)	Worst keyword (min./avg./max. shares)	numeric (3)
Average word length	numeric (1)	Avg. keyword (min./avg./max. shares)	numeric (3)
Rate of non-stop words	ratio (1)	Best keyword (min./avg./max. shares)	numeric (3)
Rate of unique words	ratio (1)	Article category (Mashable channel)	dummy (6)
Rate of unique non-stop words	ratio (1)	**Natural Language Processing (21)**	
Links (5)		Closeness to top 5 LDA topics	ratio (5)
Number of links	numeric (1)	Title subjectivity	ratio (1)
Number of Mashable links	numeric (1)	Text subjectivity score and difference	ratio (2)
Min. avg. and max. number of shares	numeric (3)	Title sentiment polarity	ratio (1)
Digital Media (2)		Rate of positive and negative words	ratio (2)
Number of images	numeric (1)	Pos. and Neg. words rate among non-neutral	ratio (2)
Number of videos	numeric (1)	Text polarity score and difference	ratio (2)
Time (8)		Polarity of pos. and neg. words	ratio (6)
Day of the week	dummy (7)	**Target**	
Published on a weekend?	bool (1)	Number of article shares	numeric (1)

two kinds of encoding format for publication time in this dataset, dummy variables for each day of a week and a bool variable for whether the article is published in the weekend. These two encodings generate features with linear dependency, so we should only keep one of them. We decide to keep the bool variable because from EDA, weekend and weekday display major differences in shares.

Figure 3: Percentage of closeness of articles *channels* to *LDA topics*. This plot can be interpreted as correlation between *channel* and *LDA variables*. For example, the bottom row shows that over 80% percent of articles in world channel are also closest to LDA0, which indicated strong correlation between them.

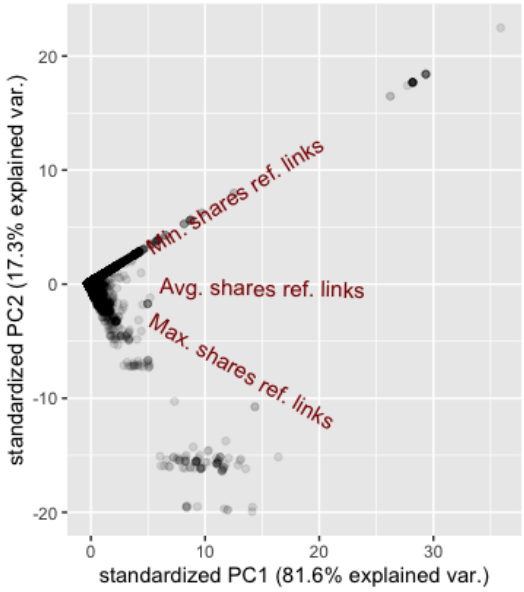


In addition, as shown in Figure 3, *LDA variables* generated by natural language processing are highly correlated with *channel*. Intuitively, *LDA variables* are hidden topics for each article generated by LDA algorithm. To make it more interpretable, we categorize each article using the closest LDA variable. By doing this, we convert five *LDA variables* into one nominal variable for the corresponding hidden topic for each article.

3.3 Numerical Variable Selection

The remaining features that may cause multicollinearity are the numerical variables. Numerical variables can be partitioned into groups, with each group containing highly correlated features that capture one specific property of the articles. For example, *minimum number of shares for referenced links*, *maximum number of shares for referenced links* and *average number of shares for referenced links* are descriptive statistics of referenced links. It is reasonable to perform PCA on them and keep only the first principal component. However, we should also take interpretablity of PCA result into consideration. Using the first principal component could lose some interpretability. Instead, we use the original features which have the direction parallel to the direction of the first PCA. This method is illustrated in Figure 4. Since we carefully group highly correlated features, we can always find an original feature which is parallel to the direction of the first principal component in that group. Applying the same method for all other groups, we finally finish variable selection with remaining variables listed in Table 2.

Figure 4: Principal component analysis for variables related to *number of shares of referenced links*. As the plot indicates, *avg. number of referenced link* is parallel to the direction of first principal component, which explained 81.6% of total variance. Further considering interpretability, we keep *avg. number of referenced link*.

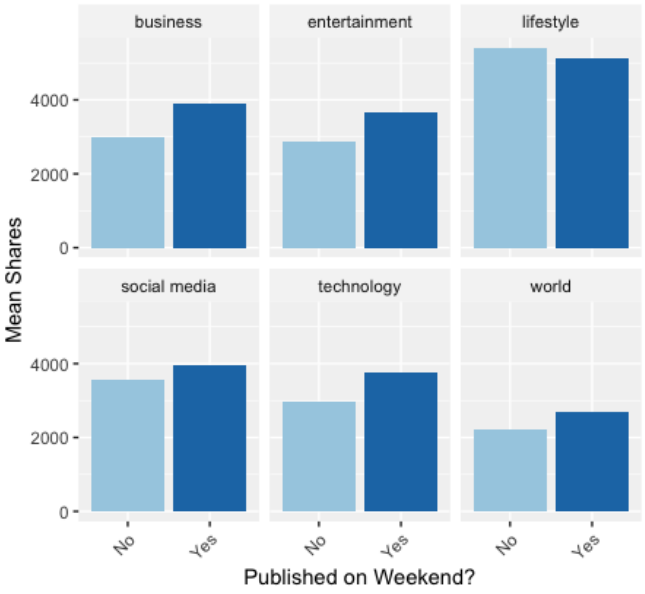


3.4 Discovering Interaction

We want to re-emphasize our goal here. We aim to understand which features could significantly impact popularity of online news, measured by the number of shares. Model interpretability is crucial when selecting variables. Thus, we only include interactions between categorical variables and interactions between categorical variables with another numerical variable. Inheriting from ANOVA and ANCOVA models, these kinds of interactions have certain meanings. Specifically, the interactions between categorical variables in linear regression models represent the difference in mean of shares for each category. The interactions between one numerical variable and categorical variables capture the impact of the numerical variable on the difference in mean of shares for each category. We do not consider interaction between numerical variables in our models. Though including them may increase the model fitting result or better predictions, doing so could destroy the simple linear relationship between the numerical covariates and response variable, and could make interpretation more difficult.

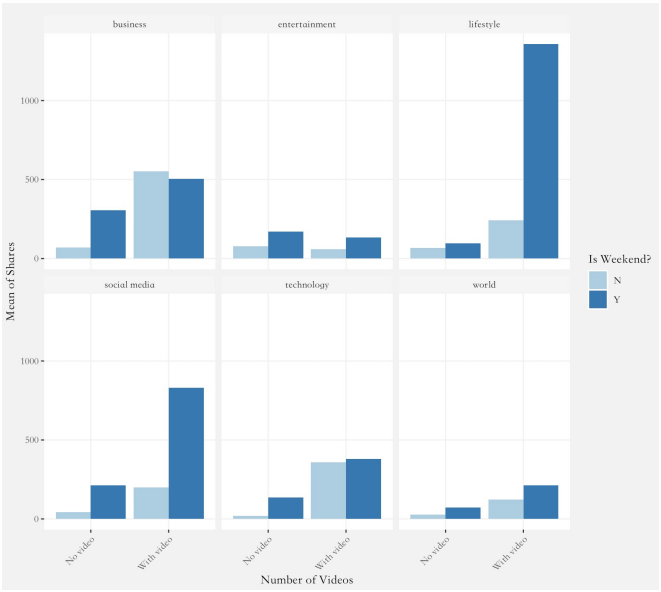
There are originally only two categorical variables in our dataset, *channel of article* and *published in weekend*. We compare shares among each group defined by published date and channel. To solve our curiosity, first we plot the weekend and channel against mean of shares. As indicated by Figure 5, different groups of articles have significantly different popularity.

Figure 5: Mean shares of articles categorized by both channel and weekend.



Meanwhile, understanding how numerical variables' impact on the popularity of articles differ among groups of articles are also worth concerning. Figure 6 shows an example of such case. In this example, number of videos shows significantly large impact in helping articles in business channel that are also published during weekdays to gain more popularity.

Figure 6: Mean shares of articles categorized by number of videos, channel and weekend. We use threshold value 1 to equally categorize articles into group of *With videos* and *No video* to show the impact of number of videos to the popularity of articles.



Based on the previous analysis, we use interactions that are first of all meaningful and then

Table 2: List of remaining variables after selection.

Feature	Log-transform?	Feature	Log-transform?
Words (4)		**Keywords (5)**	
Number of words in title	N	Article category	N
Number of words in article	Y	Number of keywords	N
Average word length	N	Avg. keyword (avg. shares)	Y
Rate of unique non-stop words	N	Best keyword (avg. shares)	Y
Links (3)		Worst keyword (avg. shares)	Y
Number of links	Y	**Natural Language Processing (7)**	
Number of Mashable links	Y	LDA (closeness to hidden topics)	N
Avg. number of shares	Y	Title subjectivity	N
Digital Media (2)		Title polarity	N
Number of images	Y	Avg. polarity of positive words	N
Number of videos	Y	Pos. words rate among non-neutral	N
Time (1)		Avg. polarity of negative words	N
Published on a weekend?	N	Neg. words rate among non-neutral	N

can indicate siginificant difference among groups, in our model. Such interactions are *channel:weekend*, *words in title:weekend*, *num.image:weekend:channel* and *num.video:weekend:channel*.

4. Model and Evaluation

4.1 Model Selection

We use forward stepwise selection method based on AIC criteria to find the best model under regression and classification setting. The algorithm is explicitly stated below.

Forward Stepwise Selection Algorithm

1. Let M_0 denote the baseline model.
2. For $k = 0, 1, \cdots, p-1$:
   (a) Consider all $p-k$ models that augment predictors in M_k with one additional predictor.
   (b) Choose the best among the $p-k$ models, and call it M_{k+1}. Here best is defined as having highest R^2 for linear regression model or smallest deviance for logistic regression model.
3. Select a single best model among $M_0, \cdots, M_p$ using AIC criteria. Best means the model with smallest AIC.

As suggested in *J. Faraway [4]*, AIC criteria is defined by 2. It works well as a model selection criteria when $\frac{n}{p}$ is large. In this problem, we have $n = 39633$ and $p \approx 50$, which satisfy the condition of using AIC.

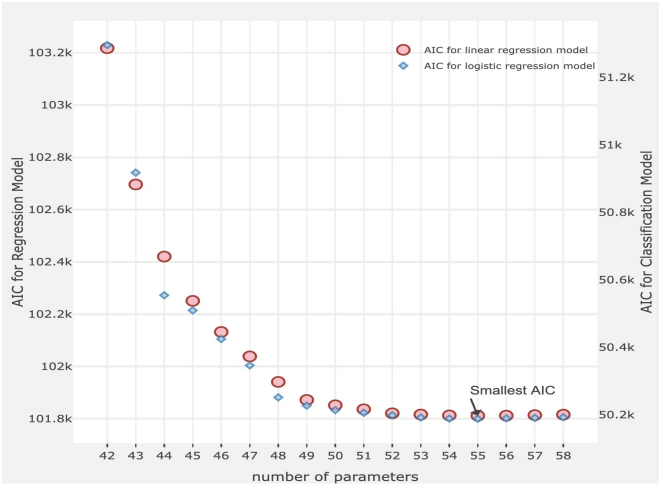
$$AIC = -2log\_likelihood + 2p \tag{2}$$

4.1.1 Baseline Model

For baseline model M_0, we fit the model with only categorical variables and interactions selected from previous analysis. Variables of lower order is also added to ensure the model is legal. The logic behind this choice of baseline model instead of a null model with only intercept is that we discover differences among these different groups and we are interested in estimating the differences. The R formula for baseline model of simple linear regression and logistic regression is given by (3).

$$\begin{aligned} &log.shares \sim title + channel + weekend + log.nimage \\ &+ log.nvideo + lda + channel * weekend + title * weekend \\ &+ log.nimage * weekend * channel \\ &+ log.nvideo * weekend * channel \end{aligned} \tag{3}$$

Figure 7: Stepwise AIC selection for selecting linear regression and logistic regression models. The red circle with respect to the left vertical axis shows AIC for linear regression models while the blue diamond with respect to the right vertical axis shows AIC for logistic regression model.



4.1.2 *Best Model*

Figure 7 shows AIC score with respect to different steps. In both regression and classification settings, model with in total 55 parameters is selected based on AIC criteria. The full summary table of corresponding model is shown in appendix Table 11 and Table 12 for linear model and logistic model, respectively.

4.2 Linear Regression Result

4.2.1 *Model Comparison*

We define the full model to be the one with all possible features. Since the baseline model, our selected model and full model are nested, we can perform pairwise F-test to compare the goodness-of-fit for our selected model with baseline model and full model. The test result is in favor of our selected model.

A further model comparison is based on comparing the prediction power of each model. Both in-sample prediction and out-of-sample prediction done by 10 folds cross validation are performed, the corresponding mean squared error is reported in 3. In both in-sample test and out-of-sample test, full model has the smallest MSE, while our selected model does not differ a lot from the best model . Since our major concern is the inverse problem, we choose the model selected by AIC as the best model and perform further analysis based on that.

Table 3: Mean square error when using selected linear regression model to do in-sample and out-of-sample prediction.

Num. Parameters	**43**	**44**	**45**	**46**
in-sample MSE	0.779	0.774	0.770	0.768
Cross-validated MSE	0.779	0.774	0.771	0.768
Num. Parameters	**47**	**48**	**49**	**50**
in-sample MSE	0.766	0.764	0.762	0.762
Cross-validated MSE	0.766	0.764	0.763	0.763
Num. Parameters	**51**	**52**	**53**	**54**
in-sample MSE	0.762	0.761	0.761	0.761
Cross-validated MSE	0.762	0.762	0.761	0.761
Num. Parameters	**55**	**56**	**57**	**58**
in-sample MSE	0.761	0.761	0.761	0.761
Cross-validated MSE	0.761	0.761	0.761	0.761

4.2.2 *Parameter Estimation*

General Effects - Figure 8 list top ten features with largest effects to the response variables.

The regression coefficients $\hat{\beta}_i$ corresponding to feature i can be interpreted as when fixing other features, changing feature i by 1 unit will increase the $log.shares$ by $\hat{\beta}_i$ units.

Figure 8: Top 10 regression coefficients in best linear regression model with largest absolute value.

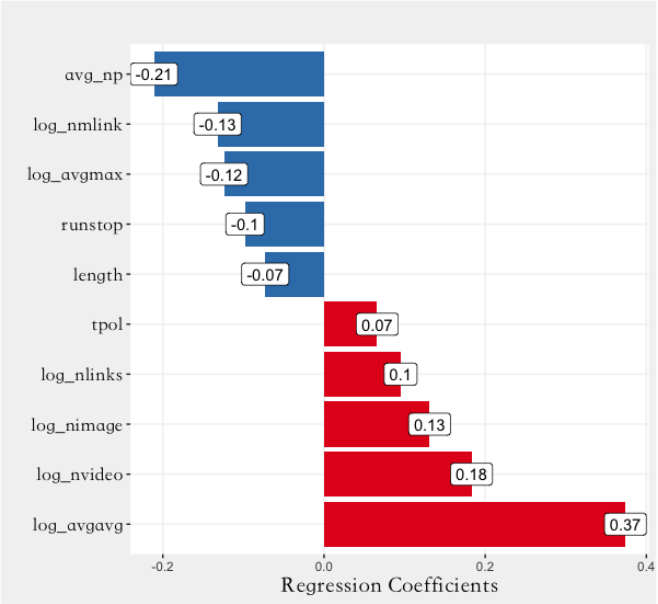


Table 4: Intercept in linear regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that the corresponding group is more popular.

Date	**Business**	**Entertainment**	**Lifestyle**
Weekend	6.324(4)	6.221(5)	6.188(6)
Weekday	5.891(11)	5.815(12)	6.123(8)
Date	**Social Media**	**Technology**	**World**
Weekend	6.549(1)	6.518(2)	6.097(9)
Weekday	6.412(3)	6.160(7)	5.902(10)

Table 5: Coefficients of log number of image in linear regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that adding number of image will have larger positive impact in making the corresponding group popular.

Date	**Business**	**Entertainment**	**Lifestyle**
Weekend	0.070(3)	-0.037(10)	0.025(8)
Weekday	0.131(1)	0.035(7)	0.046(6)
Date	**Social Media**	**Technology**	**World**
Weekend	-0.111(12)	-0.062(11)	0.068(4)
Weekday	-0.026(9)	0.051(5)	0.087(2)

Table 6: Coefficients of log number of video in linear regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that adding number of video will have larger positive impact in making the corresponding group popular.

Date	Business	Entertainment	Lifestyle
Weekend	0.031(9)	0.004(10)	0.051(6)
Weekday	0.183(2)	0.040(7)	0.034(8)
Date	**Social Media**	**Technology**	**World**
Weekend	-0.015(11)	0.131(3)	0.108(5)
Weekday	-0.060(12)	0.207(1)	0.111(4)

Effects differed by group - As indicated in Table 4, Table 5 and Table 6, articles published during weekend and in channel *social media* tend to have more shares. Articles in less popular category are more desperate in need digital media to help generating more shares.

4.2.3 Assumption Checking

A Q-Q normal plot and a residual versus fitted value plot for best linear regression model are made to check whether the model satisfies the normality and homoscedasticity assumption on error terms. The plots display some points that differ from the referenced line. As seen in data preprocessing step, because the relationship between features and response variable is quite complicated, it should not be surprising that a linear model may not capture the relationship well enough. We further analyze the outliers that potentially cause the deviation from the normality assumption. The article that has the most number of shares (i.e 84407) publishes the leaked details about the new line of cheap iPhone 5 in 2013. The article includes several leaked images of unreleased iPhone 5c. The release date of iPhone 5c (plastic shells for low-cost iPhone 5) is September, 2013; However this article was published on July, 2013. Apple has known for building buzz before any new product release. This article might excite many Apple loyal fans to share this information with the rest of Apple community. The article that has the lowest number of share(i.e 1 share) is categorized as travel channel. This article includes two images and few outside links that direct audiences to an online game store. It seems that the product this article tries to promote is not a popular choice for most online shoppers.

4.3 Logistic Regression Result

4.3.1 Model Comparison

Again, for logistic regression, since we use stepwise selection method, we can still perform likelihood ratio test to compare the goodness-of-fit for our selected model with baseline model and full model. The test we use is Rao score test. The result is in favor of our selected model.

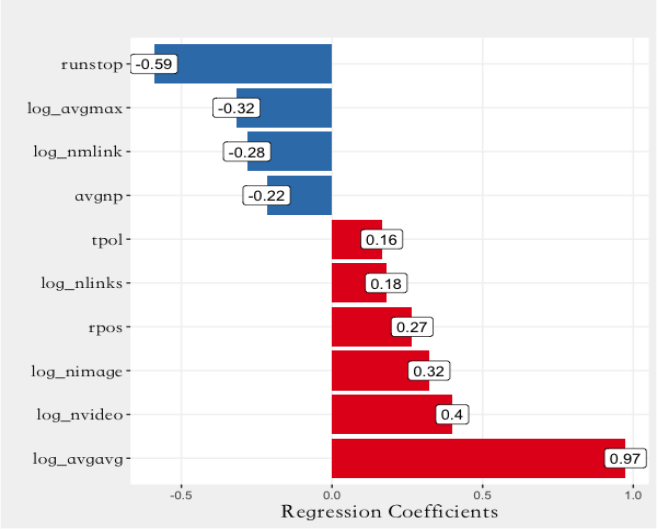
A further model comparison is based on comparing the prediction power of each model. Both in-sample prediction and out-of-sample prediction done by 10 folds cross validation are performed, the corresponding prediction accuracy is reported in 7. In both in-sample test and out-of-sample test, the model with highest accuracy only differs from our selected model by one covariates. Again, since our major concern is the inverse problem, we choose the model selected by AIC as the best model and perform further analysis based on that.

Table 7: Prediction accuracy when using selected logistic regression model to do in sample and out of sample prediction.

Num. Parameters	**43**	**44**	**45**	**46**
in-sample MSE	0.624	0.633	0.635	0.637
Cross-validated MSE	0.591	0.601	0.603	0.605
Num. Parameters	**47**	**48**	**49**	**50**
in-sample MSE	0.637	0.640	0.642	0.643
Cross-validated MSE	0.608	0.611	0.612	0.612
Num. Parameters	**51**	**52**	**53**	**54**
in-sample MSE	0.642	0.642	0.643	0.643
Cross-validated MSE	0.611	0.611	0.612	0.612
Num. Parameters	**55**	**56**	**57**	**58**
in-sample MSE	0.643	0.643	0.642	0.642
Cross-validated MSE	0.612	0.613	0.612	0.612

4.3.2 Parameter Estimation

Figure 9: Top 10 logistic regression coefficients in best classification model with largest absolute value.



General Effects - Figure 9 list top ten features with largest effects to the response variables.

Table 8: Intercept in logistic regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that the corresponding group is more popular.

Date	Business	Entertainment	Lifestyle
Weekend	-3.046(4)	-3.379(5)	-3.574(7)
Weekday	-4.131(10)	-4.283(12)	-3.684(9)
Date	**Social Media**	**Technology**	**World**
Weekend	-1.924(1)	-2.480(2)	-3.668(8)
Weekday	-2.731(3)	-3.388(6)	-4.143(11)

Table 9: Coefficients of log number of image in logistic regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that adding number of image will have larger positive impact in increasing the log odds of being popular.

Date	Business	Entertainment	Lifestyle
Weekend	0.287(2)	-0.158(11)	0.041(8)
Weekday	0.322(1)	0.047(7)	0.059(6)
Date	**Social Media**	**Technology**	**World**
Weekend	-0.438(12)	0.005(9)	0.152(4)
Weekday	-0.074(10)	0.078(5)	0.191(3)

Table 10: Coefficients of log number of video in logistic regression model corresponding to each group of sample articles categorized by published date and channel. The number inside the bracket is the rank among group, from largest to smallest. Smaller number indicates that adding number of video will have larger positive impact in increasing the log odds of being popular.

Date	Business	Entertainment	Lifestyle
Weekend	0.158(6)	0.042(8)	0.244(3)
Weekday	0.399(1)	0.052(7)	0.031(9)
Date	**Social Media**	**Technology**	**World**
Weekend	0.346(10)	-0.135(3)	0.172(5)
Weekday	-0.134(11)	0.390(2)	0.242(4)

The regression coefficients $\hat{\beta}_i$ corresponding to feature i can be interpreted as when fixing other features, changing feature i by 1 unit will increase the log odds of articles being popular by $\hat{\beta}_i$ units. Or alternatively, it will increase the odds of being popular by $\hat{\beta}_i$ times.

Effects differed by group - We obtain similar result as in regression model, which is shown in Table 8, Table 9 and Table 10Articles published during weekend and in channel *social media* tend to have more shares. Articles in less popular category are more desperate in need digital media to help generating more shares. Furthermore, articles with more technical staff are more desperate in need of digital media.

Assumption checking is also performed for logistic regression model. The result is basically the same. Therefore it is skipped in this report.

5. Summary of Findings

With the exploration and modeling, we could conclude our major findings. There are both expected and unexpected findings, and we believe understanding them can help online journalists optimize their contents.

Individual Effects - Including more digital media such as images and videos could make articles more popular. From both the linear and logistic regression models, increase in images and videos leads to increase in number of shares. The coefficients are significant at 99% level. When it comes to communication information online, a lot of studies show that our brains process visuals faster and can transmit more information if the information is communicated visually. *D. R. Vogel[5]*. Moreover, strong or even biased titles could lead to more popularity. Increase in score of polarity and subjectivity leads to increase in number of shares. Strong titles like "The Impeachment is Doomed" may speak more effectively to people, leading more impulsive shares. We also find that adding positive attitude has positive impact on popularity. In contrast, as rate of negative words increases, the number of shares decrease. Moreover, simpler words could increase number of shares based on the significant negative coefficients of word length. Articles or news provide a faster way for people to understand what happen in the world. Using complex words might lose readers' interest.

Interaction Effects - Regardless if the publication is on the weekday or the weekend, social media related articles generally are more popular than other types from both models. The interaction term between weekend and social media has the largest coefficients. This outcome is expected given the increase engagement and fast-paced change of social media in our life. However, during the weekend, technology articles are a lot more popular than on the weekday. Tech articles are more complicated and tailor to people with tech background. On the weekend, readers can spend more time carefully reading those articles. In addition, we also find that for articles that do not belong to social media or technology, they are more sensitive to adding features such as digital media.

Regression versus Classification - Though linear regres-

sion and classification method give similar outcomes and interpretation, linear regression has less robust result and worse goodness-of-fit. Intuitively, considering whether an article is popular is also more sensible and practical than predicting the exact number of shares.

6. Future Work

Using the linear model to analyze this dataset is a naive attempt. The reason for us to consider linear models in this problem is that linear models are easy to interpret. Its result can be used to make simple suggestions to online journalists. Despite this benefit, linear model is far from a good tool in this problem. As suggested by *T. Hastie [6]*, tree-based methods are much more suitable in capturing complicated relationship between features and response variable. More specifically, random forest is even better in the classification model when features are correlated. Therefore, tree-based models should be more suitable for this problem.

Moreover, we believe the future direction of this project should be based on the specific goal researchers want to achieve. If the goal is to understand the relationship between features of online news and their popularity, a better approach might be as following. First of all, we could build a classification model with the same features. The model could define the probability of popular for each article based on its features. Then we should apply the stochastic hill climbing algorithm. This algorithm starts from the initial solution (a specified article), and tries to find a point within the neighborhood of the initial solution but has a higher probability for the popular class. The drawbacks of this approach are firstly, using these features, we might not be able to get a classifier with high accuracy. However, the best classification accuracy obtained by previous research is around 70%, which may suggest that the features themselves may not be sufficient enough to predict the popularity of online news. Secondly, the stochastic hill climbing algorithm may not work well when dealing with a high dimensional searching problem.

However, if the goal is to predict the popularity of online news, a better modeling approach is through deep learning method. Neural network is known to be successful in predicting problem like this project, while the way it works is a black-box. In conclusion, a neural network model may generate much better prediction result, but researchers should also keep in mind that such models could cause difficult interpretations.

REFERENCES

[1] Kelwin Fernandes, Pedro Vinagre and Paulo Cortez (2015), "A proactive intelligent decision support system for predicting the popularity of online news," *Progress in Artificial Intelligence*, **EPIA 2015**, 535-546.

[2] Md. Taufeeq Uddin, Mohammad Shahadat Hossain, Muhammed J. A. Patwary, and Tanveer Ahsan (2016), "Predicting the popularity of online news using gradient boosting machine," *2016 ICISET*, **IEEE Bangladesh**, 1-5.

[3] He Ren and Quan Yang (2015), "Predicting and evaluating the popularity of online news," *Stanford Machine Learning Repository*, 1-5.

[4] Julian J. Faraway (2016), "Linear Model with R (Second Edition)," *CRC Press*, 154-159.

[5] D. R. Vogel, G.W. Dickson, and J.A.Lehman (1986), "Persuasion and the Role of Visual Presentation Support: The UM/3M Study", *Management Information Systems Research Center, University of Minnesota*.

[6] Trevor Hastie, Robert Tibshirani and Jerome Friedman, "The Element of Statistical Learning (Second Edition)," *Springer Series in Statistics*. 305-310.

A. Summary Tables of Best Model

A.1 Best Linear Regression Model Summary

Table 11: Summary table of best linear regression model selected by smallest AIC criteria.

	Dependent variable:
	log_shares
title	0.003
channelentertainment	-0.076^{**}
channellifestyle	0.232^{***}
channelsocial media	0.521^{***}
channeltechnology	0.269^{***}
channelworld	0.011
weekendY	0.433^{***}
log_nimage	0.131^{***}
log_nvideo	0.183^{***}
ldaLDA1	-0.080^{***}
ldaLDA2	-0.214^{***}
ldaLDA3	-0.037^{*}
ldaLDA4	-0.106^{***}
log_avgavg	0.374^{***}
log_avgmax	-0.124^{***}
log_avgref	0.040^{***}
log_nmlink	-0.132^{***}
runstop	-0.098^{*}
log_nlinks	0.095^{***}
length	-0.073^{***}
tsub	0.049^{***}
avgnp	-0.211^{***}
tpol	0.066^{***}
nkey	0.006^{**}
rneg	-0.070^{**}
log_article	-0.012^{*}
channelentertainment:weekendY	-0.027
channellifestyle:weekendY	-0.368^{***}
channelsocial media:weekendY	-0.296^{**}
channeltechnology:weekendY	-0.075
channelworld:weekendY	-0.238^{***}
title:weekendY	0.008
weekendY:log_nimage	-0.061
channelentertainment:log_nimage	-0.096^{***}
channellifestyle:log_nimage	-0.085^{***}
channelsocial media:log_nimage	-0.157^{***}
channeltechnology:log_nimage	-0.080^{***}
channelworld:log_nimage	-0.044^{*}
weekendY:log_nvideo	-0.152^{**}
channelentertainment:log_nvideo	-0.143^{***}
channellifestyle:log_nvideo	-0.149^{***}
channelsocial media:log_nvideo	-0.243^{***}
channeltechnology:log_nvideo	0.024
channelworld:log_nvideo	-0.072^{**}
channelentertainment:weekendY:log_nimage	-0.011
channellifestyle:weekendY:log_nimage	0.040
channelsocial media:weekendY:log_nimage	-0.024
channeltechnology:weekendY:log_nimage	-0.052
channelworld:weekendY:log_nimage	0.042
channelentertainment:weekendY:log_nvideo	0.116
channellifestyle:weekendY:log_nvideo	0.169^{**}
channelsocial media:weekendY:log_nvideo	0.197^{*}
channeltechnology:weekendY:log_nvideo	0.076
channelworld:weekendY:log_nvideo	0.149
Constant	5.891^{***}
Observations	39,643
R^2	0.119
Adjusted R^2	0.118
Residual Std. Error	0.873 (df = 39588)
F Statistic	98.944^{***} (df = 54; 39588)
Note:	*p<0.1; **p<0.05; ***p<0.01

A.2 Best Logistic Regression Model Summary

Table 12: Summary table of best logistic regression model selected by smallest AIC criteria.

	Dependent variable:
	label
title	-0.009
channelentertainment	-0.152^{*}
channellifestyle	0.447^{***}
channelsocial media	1.400^{***}
channeltechnology	0.743^{***}
channelworld	-0.012
weekendY	1.085^{***}
log_nimage	0.322^{***}
log_nvideo	0.399^{***}
ldaLDA1	-0.347^{***}
ldaLDA2	-0.544^{***}
ldaLDA3	-0.273^{***}
ldaLDA4	-0.262^{***}
log_avgavg	0.974^{***}
log_avgmax	-0.317^{***}
log_avgref	0.068^{***}
runstop	-0.591^{***}
log_nmlink	-0.282^{***}
log_nlinks	0.180^{***}
tpol	0.165^{***}
length	-0.125^{***}
rpos	0.266^{***}
tsub	0.099^{***}
nkey	0.016^{**}
avgnp	-0.217^{**}
log_avgmin	0.019^{*}
channelentertainment:weekendY	-0.181
channellifestyle:weekendY	-0.975^{***}
channelsocial media:weekendY	-0.278
channeltechnology:weekendY	-0.175
channelworld:weekendY	-0.610^{**}
title:weekendY	0.031^{*}
weekendY:log_nimage	-0.035
channelentertainment:log_nimage	-0.275^{***}
channellifestyle:log_nimage	-0.263^{***}
channelsocial media:log_nimage	-0.396^{***}
channeltechnology:log_nimage	-0.244^{***}
channelworld:log_nimage	-0.131^{**}
weekendY:log_nvideo	-0.241
channelentertainment:log_nvideo	-0.347^{***}
channellifestyle:log_nvideo	-0.368^{***}
channelsocial media:log_nvideo	-0.533^{***}
channeltechnology:log_nvideo	-0.009
channelworld:log_nvideo	-0.157^{*}
channelentertainment:weekendY:log_nimage	-0.170
channellifestyle:weekendY:log_nimage	0.017
channelsocial media:weekendY:log_nimage	-0.329
channeltechnology:weekendY:log_nimage	-0.038
channelworld:weekendY:log_nimage	-0.004
channelentertainment:weekendY:log_nvideo	0.231
channellifestyle:weekendY:log_nvideo	0.454
channelsocial media:weekendY:log_nvideo	0.346
channeltechnology:weekendY:log_nvideo	-0.284
channelworld:weekendY:log_nvideo	0.171
Constant	-4.131^{***}
Observations	39,643
Log Likelihood	$-25,039.060$
Akaike Inf. Crit.	50,188.120
Note:	*p<0.1; **p<0.05; ***p<0.01