PROJECT REPORT

ONLINE NEWS POPULARITY-MASHABLE.COM

SUBMITTED BY: AYUSHI SINGH

ABSTRACT AND KEYWORDS

ABSTRACT:

Mashable is a media based company launched by Pete Cashmore in 2005, it was a lifestyle blog originally for the digital generation and currently they publish article of five different data channel. They produce tens of thousands of article per year, all with varying degrees of virality. The key to Media Company's profitability is the virality of the content. Even though the content of article is good still few articles don't get good number of shares. Popularity prediction of online news aims to predict the future popularity of new article prior to the publication estimating the number of shares, likes and comments that particular article will get depending on various features .This report focuses on determining whether the article will be popular or not depending on various features . Data processing like outlier treatment feature selection are used to enhance the performance of the model. After applying models like Logistic Regression , KNN, Naïve Bayes, Decision Tree and Random Forest. We get the best accuracy from Random Forest and Logistic Regression using RFE.

KEYWORDS: Online news popularity, machine learning, data processing, accuracy

TABLE OF CONTENTS

ABBREVIATIONS USED

ABBREVIATION	EXPANSION
LR	LOGISTIC REGRESSION
KNN	K-NEAREST NEIGHBOUR
DT	DECISION TREE
RF	RANDOM FOREST
AUC	AREA UNDER THE CURVE
ROC CURVE	RECIEVER OPERATIONG CHARACYERISTIC CURVE
BAG-DT	BAGGING-DECISION TREE
BOOST-DT	BOOSTING-DECISION TREE
FPR	FALSE POSITIVE RATE
FNR	FALSE NEGATIVE RATE

EXECUTIVE SUMMARY

BACKGROUND AND NEED FOR STUDY: The consumption of online news accelerates day by day due to widespread adoption of smartphones and use of social media and rise of social networks. Online news content comprises of various key properties. For instance, it is easily produced and small in size, its lifespan is short and the cost is low. Such properties make news content more effective to be consumed on social sharing platforms. This type of content can capture the attention of significant amount of Internet users within a short period of time. As a consequence researchers focus on the analysis of online news content such as predicting the popularity of news articles, demonstrating the decrease in interest of consumers and understanding them.

SCOPE AND OBJECTIVE: The objective of this project is to do a research and develop a methodology by building models for online News Popularity Analysis. By analysing, the factors or features on which shares depend. When does the number of share increase, which type of blog does consumer like . Whenever the number of shares are more those patterns will be analysed. Depending on these patterns action could be taken on understanding how the article should be and thus understanding the customer.

APPROACH AND METHODOLOGY: The data is available on https://archive.ics.uci.edu . After processing the dataset, treating outlier using IQR and Zscore. With the help of exploratory Data Analysis we tried to understand how the independent features are related to dependent features. We used feature selection to select the important features which adds to the accuracy score of the classification model. The predictive models is to understand the features that influence the number of shares.

KEY LEARNINGS: The number of shares were more for Wednesday and the least number of shares were on Friday. While lifestyle data channel article got the maximum number of shares and social media got the least. Thus publishers can follow the trend of the shares and take actions accordingly.

RECOMMENDATIONS AND ACTIONABLE INSIGHTS: The recommendations for the project are developed by predicting the customers choice of article (which got more shares). As the articles which were published on Wednesday or Monday got more shares publishers should choose the day of publishing wisely . While most of the articles which were shared contained the word appl in it.

CHAPTER 1- PROJECT OVERVIEW

Online news sharing or sharing of blogs has been accelerating day by day as the advancement of technology, social media, use of smart phones. The prediction of the popularity of online news content has remarkable practical values in many fields. For example, by utilizing the advantages of popularity prediction, news organization can gain a better understanding of different types of online news consumption of users. As a result, the news organization can deliver more relevant and engaging content in a proactive manner as well as the organization can allocate resources more wisely to develop stories over those content. Prediction of news content is also beneficial for trend forecasting, understanding the collective human behavior, advertisers to propose more profitable monetization techniques, invest on those particular advertisement and readers to filter the huge amount of information quickly and efficiently.This is where retail analytics comes into play.

Mashable is a media based company which publishes blogs and articles in different fields or data channel. It was launched by Pete Cashmore in 2005, it was a lifestyle blog originally for the digital generation.

➢ Industry Review

Most of peoples now have the habit of reading and sharing news online, for instance, using social media like Twitter and Facebook. With the help of Internet, the online news can be instantly spread around the world. Thus, it is interesting and meaningful to use the machine learning techniques to predict the popularity of online news articles. Various works have been done in prediction of online news popularity. The popularity of, online articles is analysed based on the users' comments. Mashable is the website that publishes the details & ratings based on the customer comments.

• Current practices

In 2015, **55%** of people reported that print was their preferred method for reading a newspaper, down 4% from 2014. The methods people use to get their news from digital means was at 28%, as opposed to 20% of people attaining the news through print newspapers. These trends indicate an increase in digital consumption of newspapers, as opposed to print. Today, advertisement revenue for digital forms of newspapers is nearly 25%, while print is constituting the remaining 75%.

• Background Research

In 2013, the Reuters Institute commissioned a cross-country survey on news consumption, and gathered data related to online newspaper use that emphasize the lack of use of paid online newspaper services. The countries surveyed were France, German, Denmark, Spain, Italy, Japan, Brazil, the United States, and the United Kingdom. All samples within each country were nationally representative.

Half of the sample reportedly paid for a print newspaper in the past 7 days, and only one-twentieth of the sample paid for online news in the past 7 days. That only 5% of the sample had recently paid for online newspaper access is likely because most people access news that is free. People with portable devices, like tablets or smartphones, were significantly more likely to subscribe to digital news content.

Additionally, younger people 25 to 34-year-olds are more willing to pay for digital news than older people across all countries. This is in line with the Pew Research Center's finding in a survey of U.S. Americans that the Internet is a leading source of news for people less than 50.

➢ Literature Survey

• History

An early example of an "online only" newspaper or magazine was (PLATO) News Report, an online newspaper created by Bruce Parrello in 1974 on the PLATO system at the University of Illinois. Beginning in 1987, the Brazilian newspaper Jornaldodia ran on the state owned Embratel network, moving to the internet in the 1990s. By the late 1990s, hundreds of U.S. newspapers were publishing online versions, but did not yet offer much interactivity One example is Britain's Weekend City Press Review, which provided a weekly news summary online beginning in 1995.

• Examples

Newspapers with specialized audiences such as The Wall Street Journal and The Chronicle of Higher Education successfully charge subscription fees. Most newspapers have an online edition, including The Los Angeles Times, The Washington Post, USA Today, Mid-Day, and Times. The experimented with new media in 2005, offering a free twelve-part weekly podcast series by Ricky Gervais. Another UK daily to go online is The Daily Telegraph.

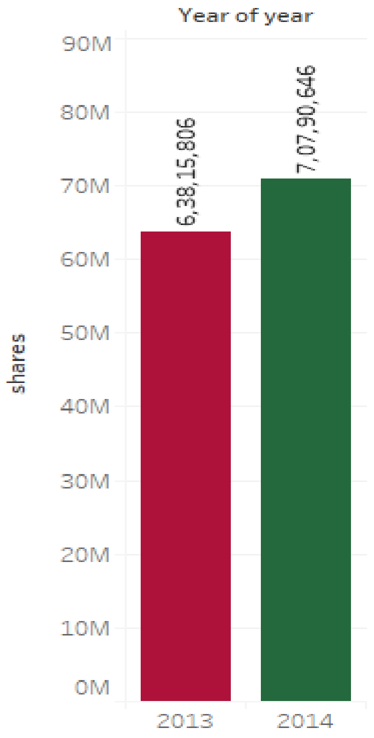
In Australia, most major newspapers offer an online version, with or without a paywalled subscription option. In Algeria, the number of daily visitors of news websites and online editions of newspapers surpasses the number of daily readers of print newspapers since the end of 2016.

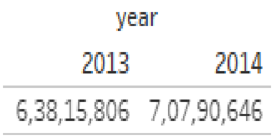
• Undergoing research

Hybrid newspapers are predominantly focused on online content, but also produce a print form. Trends in online newspapers indicate publications may switch to digital methods, especially online newspapers in the future. The New York Times is an example of this model of newspaper as it provides both a home delivery print subscription and a digital one as well. There are some newspapers which are predominantly online, but also provide limited hard copy publishing.

An example is annarbor.com, which replaced the Ann Arbor News in the summer of 2009. It is primarily an online newspaper, but publishes a hard copy twice a week. Other trends indicate that this business model is being adopted by many newspapers with the growth of digital media.

NEED FOR STUDY: As we can see, we have the data for the year 2013 and 2014, the number of shares increased from 6,38,15,806 to 7,07,90,646.





year	
2013	2014
6,38,15,806	7,07,90,646

Thus the media based company like mashable are gaining more number of shares with time.

PROBLEM STATEMENT AND PROJECT SCOPE: The dataset consist of all the features or variables of the articles.WE would use them to predict whether the article would be popular or not. The scientific goal is to develop a predictive model that can be used to determine the virality, how popular would the article be after getting published, of a given piece of content. Authors at Mashable or a similar media company could then take findings from the model and further optimize the virality of the content. These parameters could be simple things like word count of the title or on which day was it published on in which category it falls. Or more complex changes, such as overall sentiment of the piece.

DATA SOURCES: The data is available at archive.ics.uci.edu and source information is from Creators: Kelwin Fernandes , Pedro Vinagre and Pedro Sernadela. The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it.

DATA DESCRIPTION: The dataset online news popularity consists of 39644 entries and 61 columns. With url and timedelta as the non-predictive feature. And shares as the dependent variable. The minimum number of shares is 1 and maximum is 843300.While the median is 1400. The dataset has zero null values.

FEATURE DESCRIPTION:

The dataset consists 39644 entries and 61 columns. The dependent variable in the above dataset is Share i.e. number of shares of the article.

Feature Name	Feature Description	Type of Data
url	URL of the article	non-predictive
timedelta	Days between the article publication and the dataset acquisition	non-predictive
n_tokens_title	Number of words in the title	discrete
n_tokens_content	Number of words in the content	numerical
n_unique_tokens	Rate of unique words in the content	numerical
n_non_stop_words	Rate of non-stop words in the content	numerical
n_non_stop_unique_tokens	Rate of unique non-stop words in the content	numerical
num_hrefs	Number of links	numerical
num_self_hrefs	Number of links to other articles published by Mashable	numerical
num_imgs	Number of images	numerical
num_videos	Number of videos	numerical
average_token_length	Average length of the words in the content	numerical
num_keywords	Number of keywords in the metadata	discrete

data_channel_is_lifestyle	Is data channel 'Lifestyle'?	categorical
data_channel_is_entertainment	Is data channel 'Entertainment'?	categorical
data_channel_is_bus	Is data channel 'Business'?	categorical
data_channel_is_socmed	Is data channel 'Social Media'?	categorical
data_channel_is_tech	Is data channel 'Tech'?	categorical
data_channel_is_world	Is data channel 'World'?	categorical
kw_min_min	Worst keyword (min. shares)	numerical
kw_max_min	Worst keyword (max. shares)	numerical
kw_avg_min	Worst keyword (avg. shares)	numerical
kw_min_max	Best keyword (min. shares)	numerical
kw_max_max	Best keyword (max. shares)	numerical
kw_avg_max	Best keyword (avg. shares)	numerical
kw_min_avg	Avg. keyword (min. shares)	numerical
kw_max_avg	Avg. keyword (max. shares)	numerical
kw_avg_avg	Avg. keyword (avg. shares)	numerical
self_reference_min_shares	Min. shares of referenced articles in Mashable	numerical
self_reference_max_shares	Max. shares of referenced articles in Mashable	numerical
self_reference_avg_sharess	Avg. shares of referenced articles in Mashable	numerical
weekday_is_monday	Was the article published on a Monday?	categorical
weekday_is_tuesday	Was the article published on a Tuesday?	categorical
weekday_is_wednesday	Was the article published on a Wednesday?	categorical
weekday_is_thursday	Was the article published on a Thursday?	categorical

weekday_is_friday	Was the article published on a Friday?	categorical
weekday_is_saturday	Was the article published on a Saturday?	categorical
weekday_is_sunday	Was the article published on a Sunday?	categorical
is_weekend	Was the article published on the weekend?	categorical
LDA_00	Closeness to LDA topic 0	numerical
LDA_01	Closeness to LDA topic 1	numerical
LDA_02	Closeness to LDA topic 2	numerical
LDA_03	Closeness to LDA topic 3	numerical
LDA_04	Closeness to LDA topic 4	numerical
global_subjectivity	Text subjectivity	numerical
global_sentiment_polarity	Text sentiment polarity	numerical
global_rate_positive_words	Rate of positive words in the content	numerical
global_rate_negative_words	Rate of negative words in the content	numerical
rate_positive_words	Rate of positive words among non-neutral tokens	numerical
rate_negative_words	Rate of negative words among non-neutral tokens	numerical
avg_positive_polarity	Avg. polarity of positive words	numerical
min_positive_polarity	Min. polarity of positive words	numerical
max_positive_polarity	Max. polarity of positive words	numerical
avg_negative_polarity	Avg. polarity of negative words	numerical

min_negative_polarity	Min. polarity of negative words	numerical
max_negative_polarity	Max. polarity of negative words	numerical
title_subjectivity	Title subjectivity	numerical
title_sentiment_polarity	Title polarity	numerical
abs_title_subjectivity	Absolute subjectivity level	numerical
abs_title_sentiment_polarity	Absolute polarity level	numerical
shares	Number of shares (target)	numerical

CREATING COLUMN:

As we will perform classification on the above dataset so we need a labelled class. Thus we use the median value of shares i.e. 1440 as the threshold. The shares which are greater than 1440 will be considered as popular and are labelled 1 and which are below or equals to 1440 are considered to be not popular and are labelled as 0.

DATA PREPARATION AND CLEANUP:

After literature survey and understanding the data. The next step is to clean the data i.e. to check there are no missing values, or the data need to be label encoded or not, checkinf for outlier treatment, standardization of the data, oversampling is the data is unbalanced, apply transformation if the data is skewed to reduce the skewness.

Table 2.0

Label Encoding/Get Dummies	Outlier treatment	Standardisation	Dealing with unbalanced data
We don’t need to use label encoder as the categorical columns are already label encoded using get dummies.	The dataset consisted a lot of outliers in almost all the columns .We used boxplot to visualize all the columns. So we needed to treat those outliers. The method used to treat those outliers is IQR and zscore. Wherever the maximum value was not affecting the mean and standard deviation zscore method(3 sigma method) was used and in the rest IQR(interquartile range) method was used.	It is not necessary to standardize the dataset.	Upon creating the new columns class labels i.e. categorizing them into popular and not popular we checked for data imbalanace using count plot as clearly infere the dataset is balanced.

Statistical tools & techniques :

Various classification algorithms have been used to analyze the popularity of the articles. We have also used cross val score and grid search CV in model to increase the accuracy of the models and conduct hyper parameter tuning.
The various classification models are:

- Logistic Regression
- KNN
- Decision Tree
- Naïve Bayes
- Random Forest
- Bag-dt
- Boost-DT
- Bag-LR
- Boost-LR

Model performance measures used for evaluating models :

We will use the following model performance measures to check the model accuracy.

	Negative (Predicted)	**Positive (Predicted)**
Negative (Observed)	True Negative(TN)	False Positive(FP)
Positive (Observed)	False Negative(FN)	True Positive(TP)

Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The higher the accuracy the better the model

Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR). For a model sensitivity or recall should be more.

Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions. In the above case how many correct popular models were predicted as popular.

F1-Score

F1 is an overall measure of a model's accuracy that combines precision and recall .A good F1-Score means there are less number of false positive and false negatives and more number of true positive and true negative. So that we don't get any wrong predictions and al false alarms.

ROC CURVE:

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

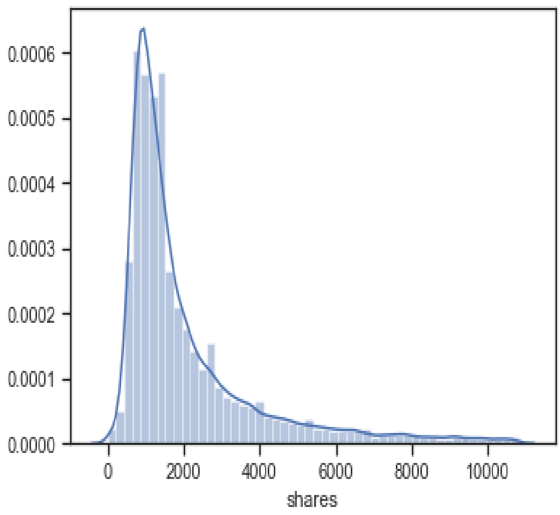
Level of significance

For all the hypothesis tests in the project, the level of significance is assumed as 5% unless specified otherwise.

Chapter 2 - Exploratory data analysis

The purpose for exploratory data analysis is to understand the data, their distribution, how they are related to shares. We need to find which features lead to increase in number of shares the most and how.

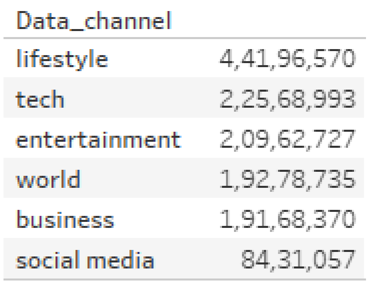
1. DISTRIBUTION OF SHARES



After removing the outliers from the shares column and keeping a threshold of 10800 as the maximum value for shares the distribution is :

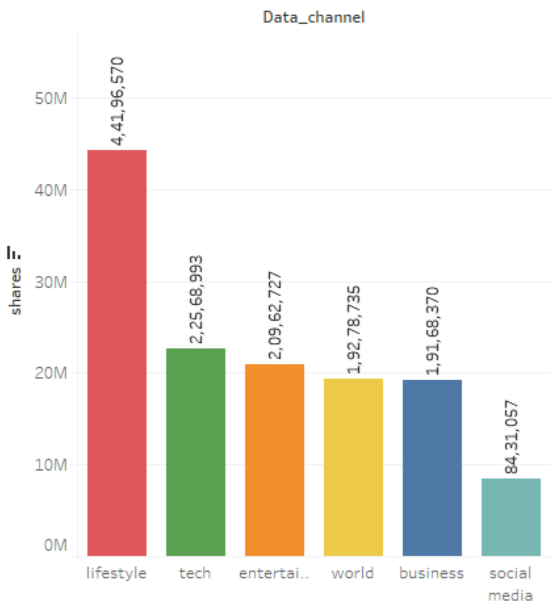
The maximum number of shares articles got were in the range 1000 to 3000.

SHARES WITH DATA CHANNEL:



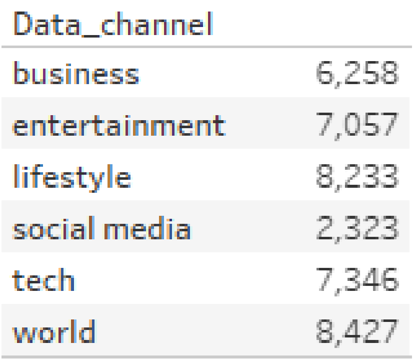
Data_channel	
lifestyle	4,41,96,570
tech	2,25,68,993
entertainment	2,09,62,727
world	1,92,78,735
business	1,91,68,370
social media	84,31,057

From the above table we can infer that lifestyle data channel got the maximum number of shares. While social media got the least number of shares.

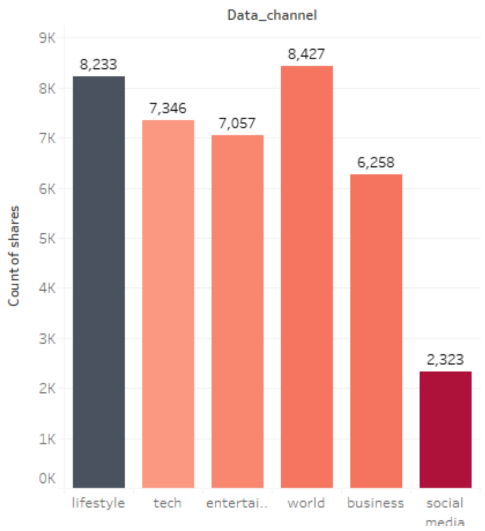


- Lifestyle has got a sum of shares of 4,4196,570
- From the count of data channel we see that most number of articles were of the data channel world while world got a very less number of shares.

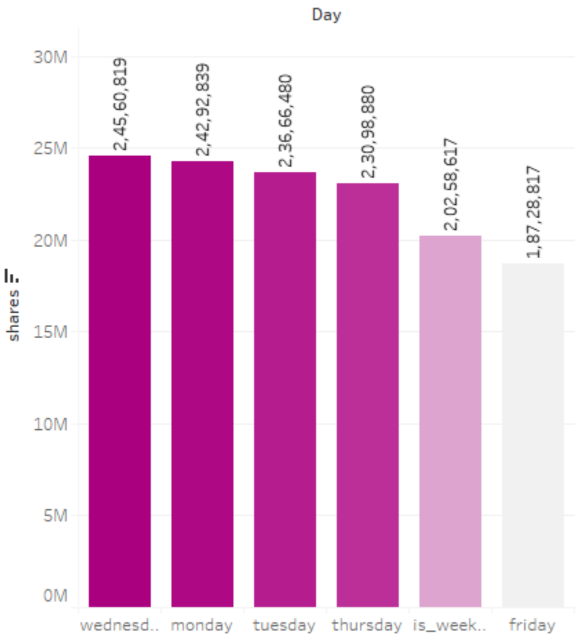
Table: count of article of that particular data channel



Data_channel	
business	6,258
entertainment	7,057
lifestyle	8,233
social media	2,323
tech	7,346
world	8,427

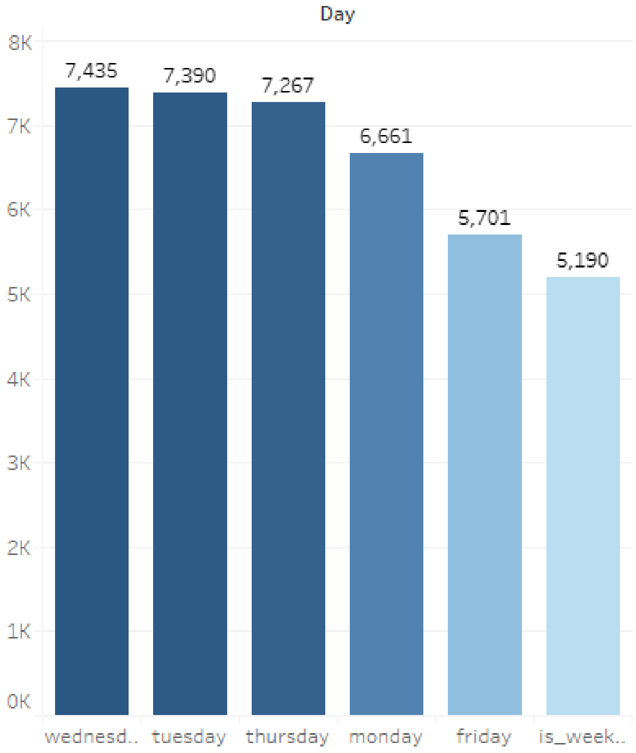


Weekday with shares:



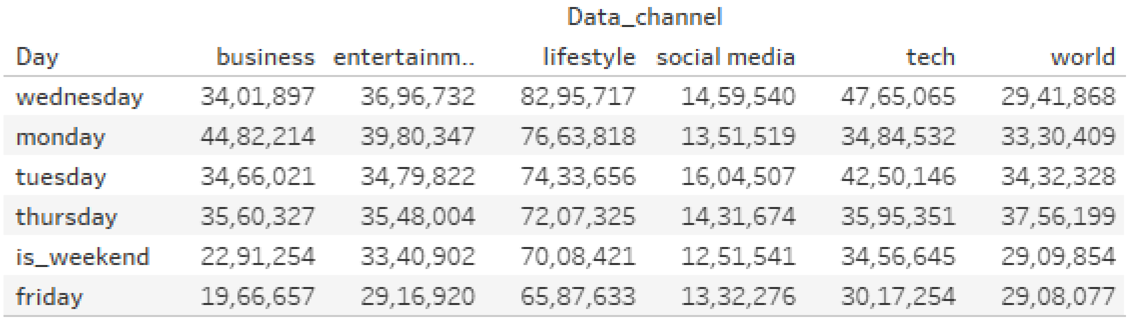
- From the plot we can infer that maximum number of shares were on Wednesday and least was on Friday.

Countplot of number of articles published



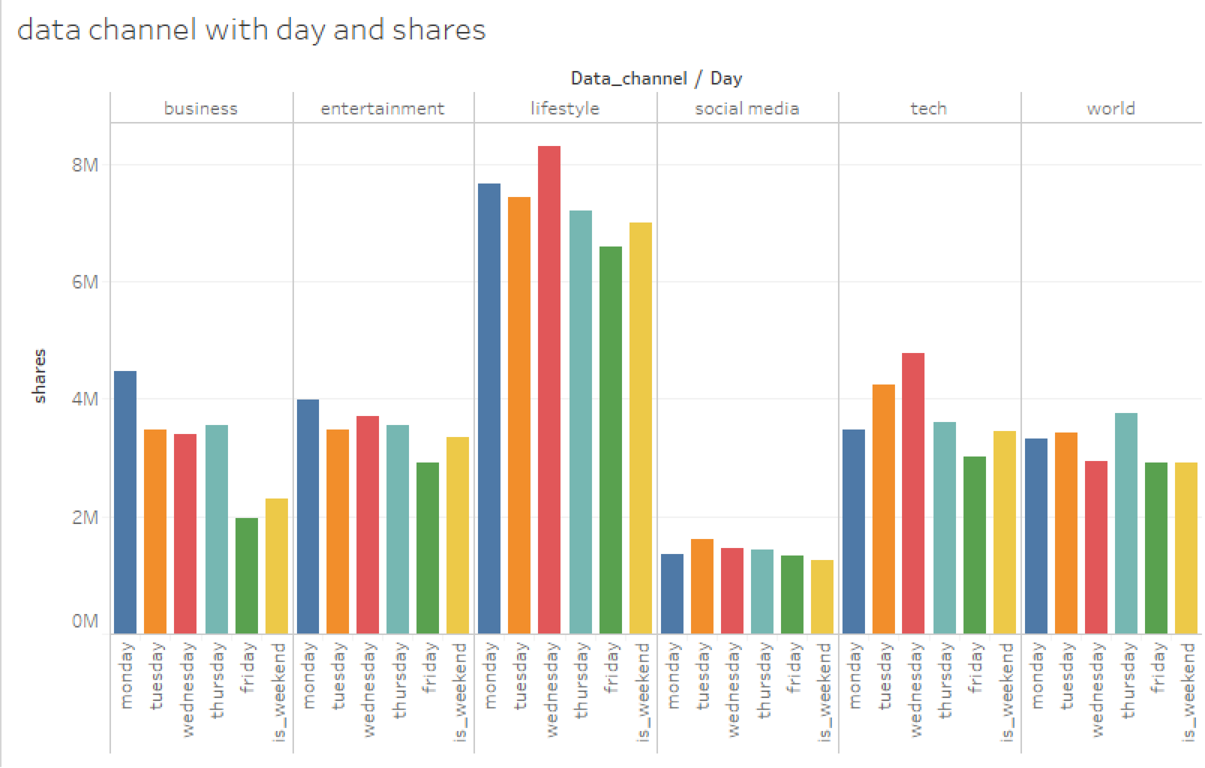
As the least number of articles published on weekend and Friday is very less compared to articles published on Wednesday.

While the number of articles published on Monday were less compared to Tuesday still the number of shares on Monday is more.



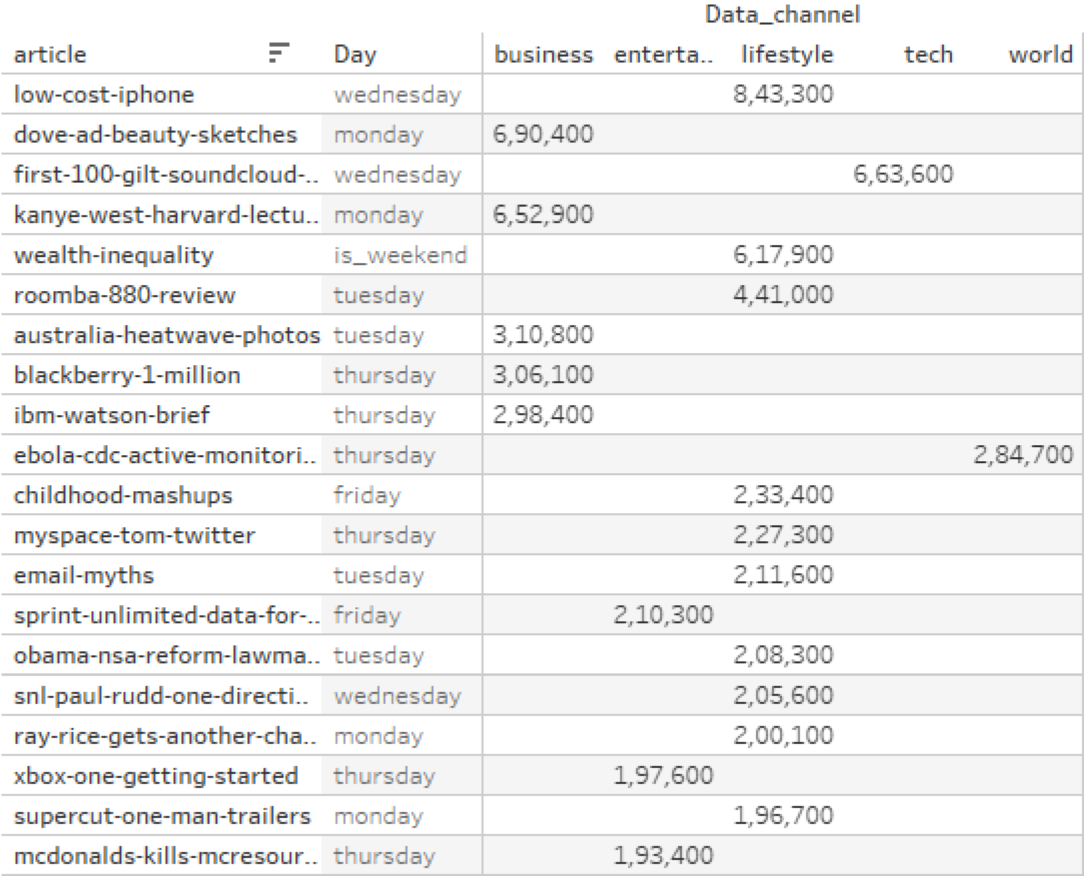
	Data_channel					
Day	business	entertainm..	lifestyle	social media	tech	world
wednesday	34,01,897	36,96,732	82,95,717	14,59,540	47,65,065	29,41,868
monday	44,82,214	39,80,347	76,63,818	13,51,519	34,84,532	33,30,409
tuesday	34,66,021	34,79,822	74,33,656	16,04,507	42,50,146	34,32,328
thursday	35,60,327	35,48,004	72,07,325	14,31,674	35,95,351	37,56,199
is_weekend	22,91,254	33,40,902	70,08,421	12,51,541	34,56,645	29,09,854
friday	19,66,657	29,16,920	65,87,633	13,32,276	30,17,254	29,08,077

Sum of shares received by articles on different day with different data channel.

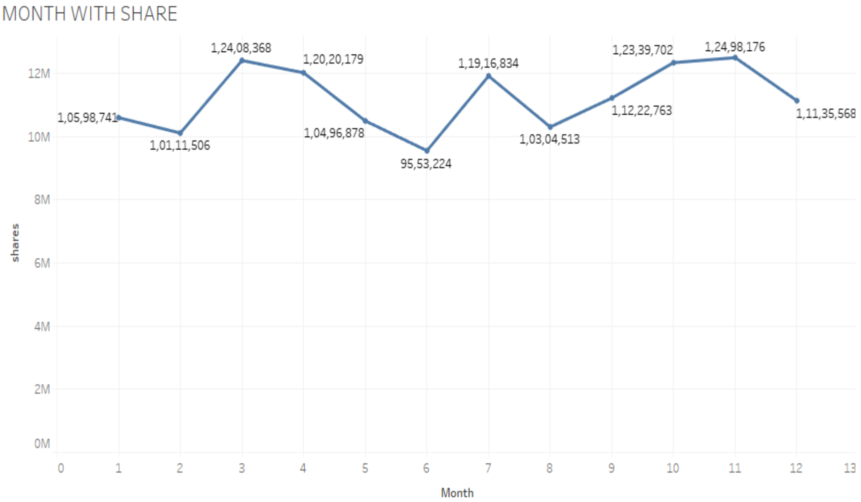


- From the above plots we understood the nature of user when the article gets maximum number of shares and also when the publishers tend to publish the article.
- Thus monday and lifestyle channel were the day and channel which got most number of shares.
- Friday and social media got the least number of shares.

TOP 20 ARTICLES WITH SHARES (DAY AND LIFESTYLE):

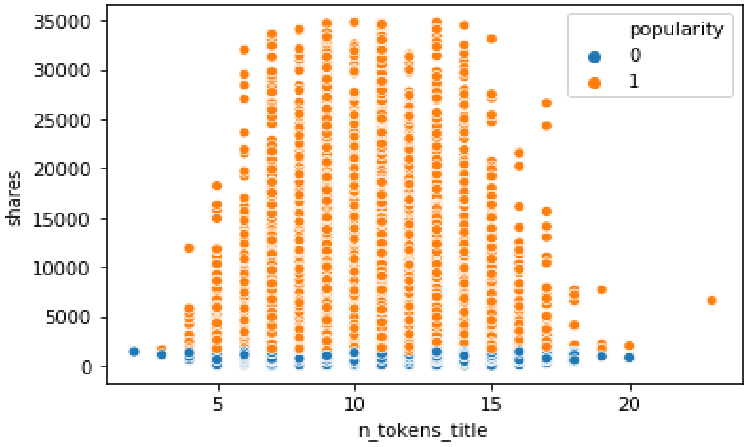


article	Day	Data_channel business	enterta..	lifestyle	tech	world
low-cost-iphone	wednesday			8,43,300		
dove-ad-beauty-sketches	monday	6,90,400				
first-100-gilt-soundcloud-..	wednesday				6,63,600	
kanye-west-harvard-lectu..	monday	6,52,900				
wealth-inequality	is_weekend			6,17,900		
roomba-880-review	tuesday			4,41,000		
australia-heatwave-photos	tuesday	3,10,800				
blackberry-1-million	thursday	3,06,100				
ibm-watson-brief	thursday	2,98,400				
ebola-cdc-active-monitori..	thursday					2,84,700
childhood-mashups	friday			2,33,400		
myspace-tom-twitter	thursday			2,27,300		
email-myths	tuesday			2,11,600		
sprint-unlimited-data-for-..	friday		2,10,300			
obama-nsa-reform-lawma..	tuesday			2,08,300		
snl-paul-rudd-one-directi..	wednesday			2,05,600		
ray-rice-gets-another-cha..	monday			2,00,100		
xbox-one-getting-started	thursday		1,97,600			
supercut-one-man-trailers	monday			1,96,700		
mcdonalds-kills-mcresour..	thursday		1,93,400			



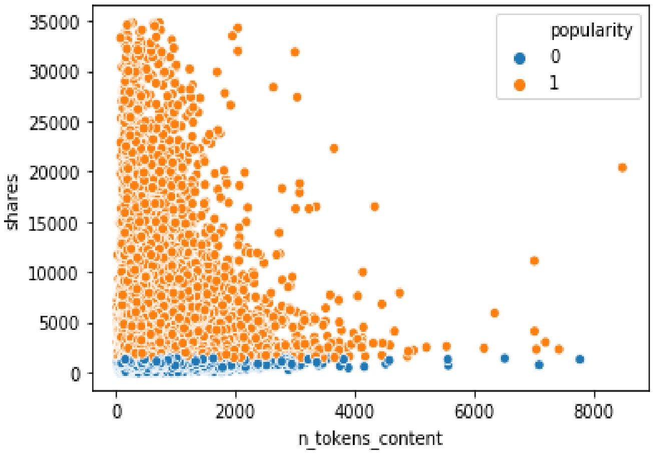
- ➢ From the line plot we can see that maximum number of shares the article received was in the 11th month.

Relationship between n_tokens_title vs shares:



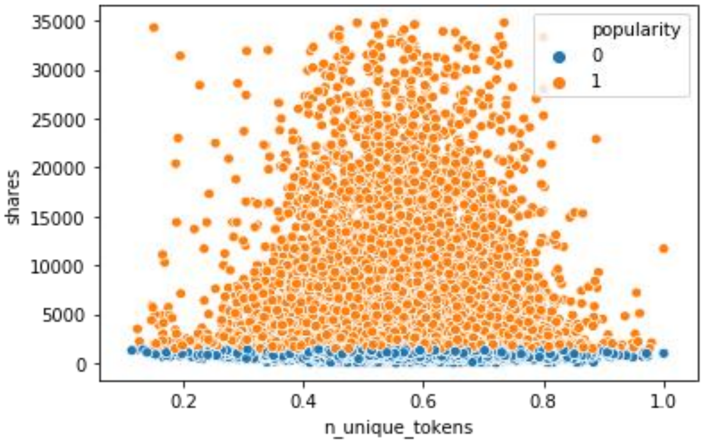
- ➢ The articles which are having 6 to 15 number of words in the title are having more popularity.
- ➢ So the publisher can keep this in notice and accordingly choose how many words to keep in title. User tends to like articles with title which are neither too short nor too long.

Relationship between n_tokens_content vs shares:



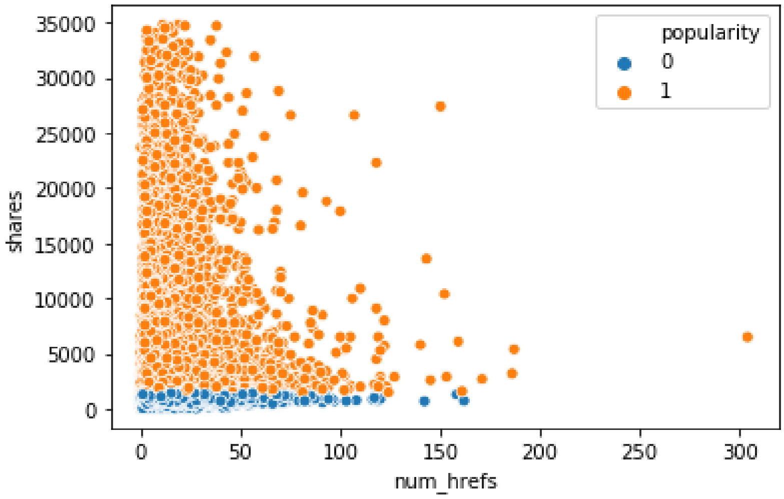
- ➢ There are some articles which are having so many numbers of words in it.
- ➢ There are more number of popular article/news which is having around 500-1500 words in it.
- ➢ So the lesser the number of words in the content more would be the share.

Relationship between number of unique tokens vs shares:



- ➢ Most of the popular article are having 0.4 to 0.8 rate of unique words in the content.
- ➢ It can be recommended to use more number of unique words in an article.
- ➢ Articles which got more number of shares had a rate between 0.4 to 0.7 of unique words used

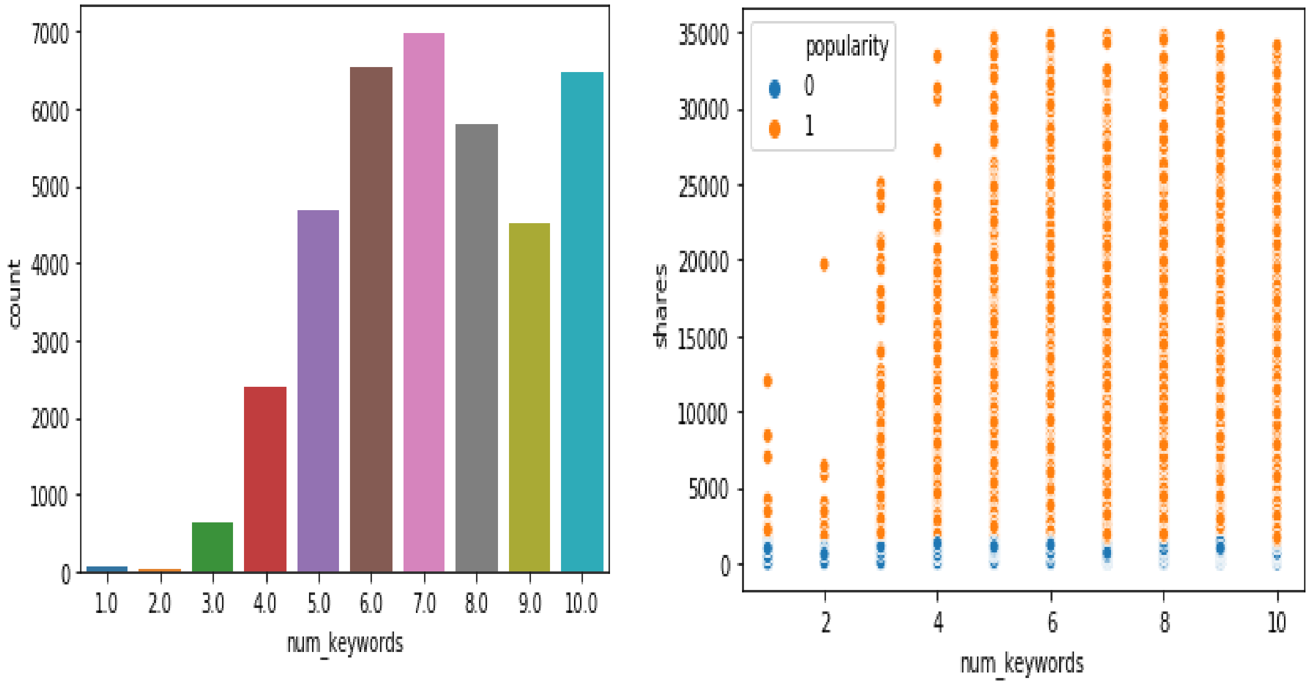
Relationship between number of links and shares:



- ➢ It can be seen from the plot that there are some articles which are having some extreme number of links in it.
- ➢ And most of the popular/good article is having less than 50 links in it. So, it is good for an article to have less number of links in it to gain popularity.

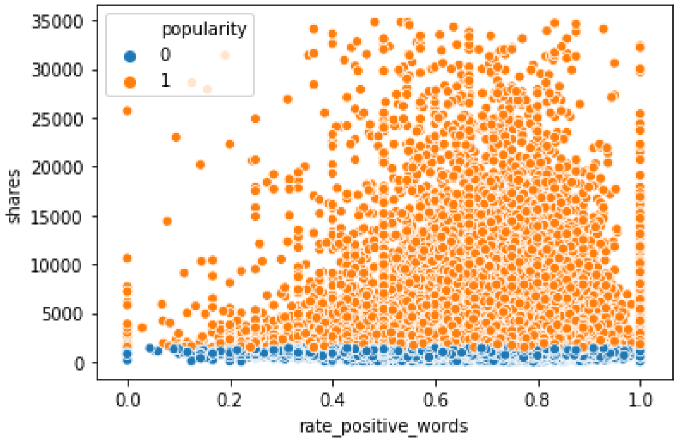
`

Count of Number of keywords in the metadata and relationship of num_keywords with shares:



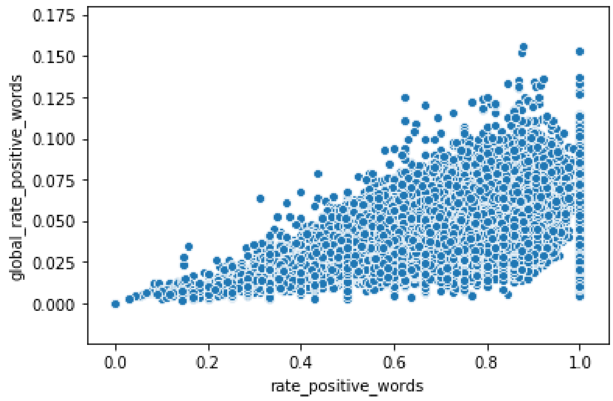
- ➢ From the above right graph, it can be inferred that the articles which are having 7 number of keywords in the metadata are published more.
- ➢ From the scatter plot, we can infer that the articles which are having more than 4 keywords in the metadata are having high chances of getting more shares and becoming popular.

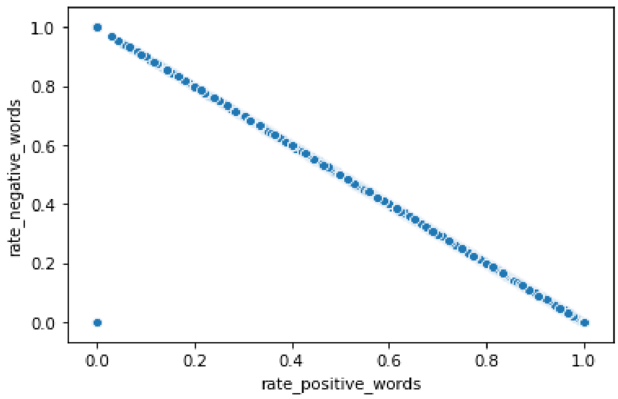
Relationship between Rate of positive words and shares:

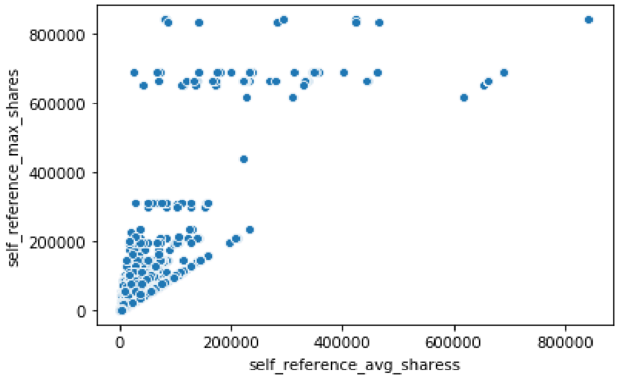


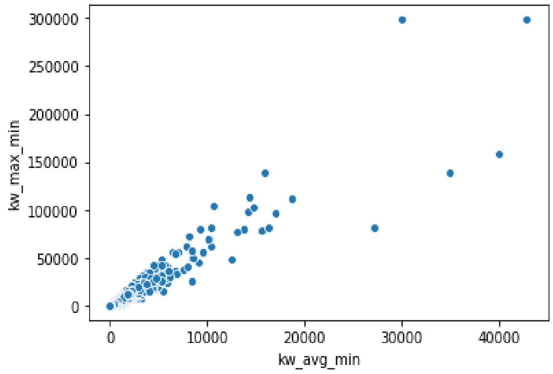
- From the above plot, the articles which are having rate of positive words greater than 0.5 are getting more popular.
- It can be inferred that, people tends to read and share the article which is having more positive content.

Check for Multicollinearity:









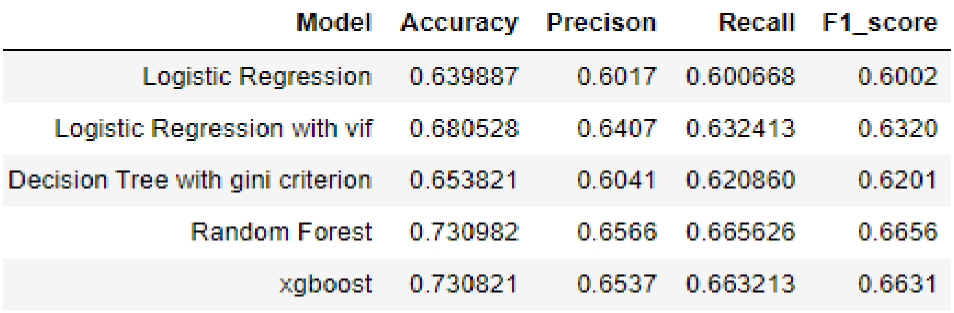
From the above plots, we can infer that there exists:

- Linear relationship between rate_positive_words and rate_negative_words. A strong negative correlation exists between them.
- Features global_rate_positive_words and rate_positive_words ,self_reference_max_shares and self_reference_avg_shares, kw_avg_min and kw_max_min are having strong positive correlation with each other.
- Therefore, it can be inferred that the multicollinearity exists among the features.

CHAPTER 3- Feature Selection & Model Building

Feature Selection is used to filter out features which are not important and use those features which have utmost importance for the model. In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features. However, using backward feature selection we got 39 features for our classification model whereas with vif method we got 54 features. Besides, considering the real-time usage of the proposed system, achieving better or similar classification performance with less number of features will improve the scalability of the system since less number of features will be kept track during the session

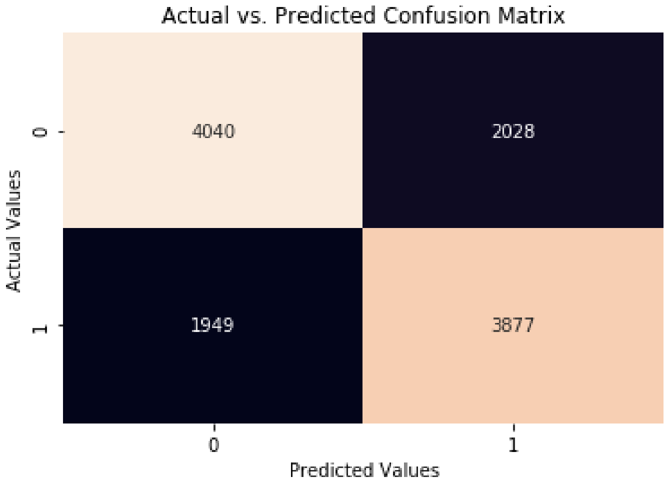
CLASSIFICATION RESULT:



Model	Accuracy	Precison	Recall	F1_score
Logistic Regression	0.639887	0.6017	0.600668	0.6002
Logistic Regression with vif	0.680528	0.6407	0.632413	0.6320
Decision Tree with gini criterion	0.653821	0.6041	0.620860	0.6201
Random Forest	0.730982	0.6566	0.665626	0.6656
xgboost	0.730821	0.6537	0.663213	0.6631

When comparing results we got from LR,LR with vif, RF, xgboost the highest accuracy achieved was from random forest using randomizedSearchCV with an accuracy of 73.1%.Using vif feature selection method accuracy increased for Logistic Regression. XGBoost gave the roc accuracy near to random forest but the precision and recall scores are more for RF. As a result we consider random forest to be the best model for predicting online shares to an article.

Confusion matrix of random forest:



True positives: 3877 (Articles which were popular and were predicted as popular)

True Negative: 4040 (Articles which were predicted as not popular and were actually popular)

False Positive: 2028 (Articles which were predicted as popular but were not popular)

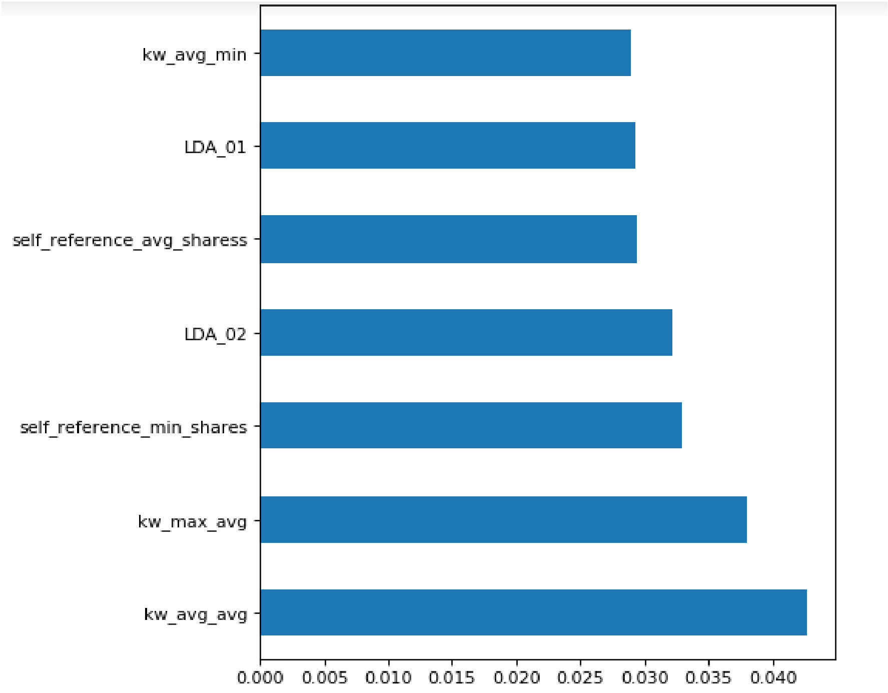
False Negative: 1949 (Articles which were predicted as not popular but were actually popular.

Precision=TP/ (TP+FP) =3877/ (3877+2028) =0.6566

Recall=TP/ (TP+FN) =3877/ (3877+1949) =0.6654(sensitivity)

Specificity=TN / (TN + FP)= 4040 / (4040 +2028) =0.66578

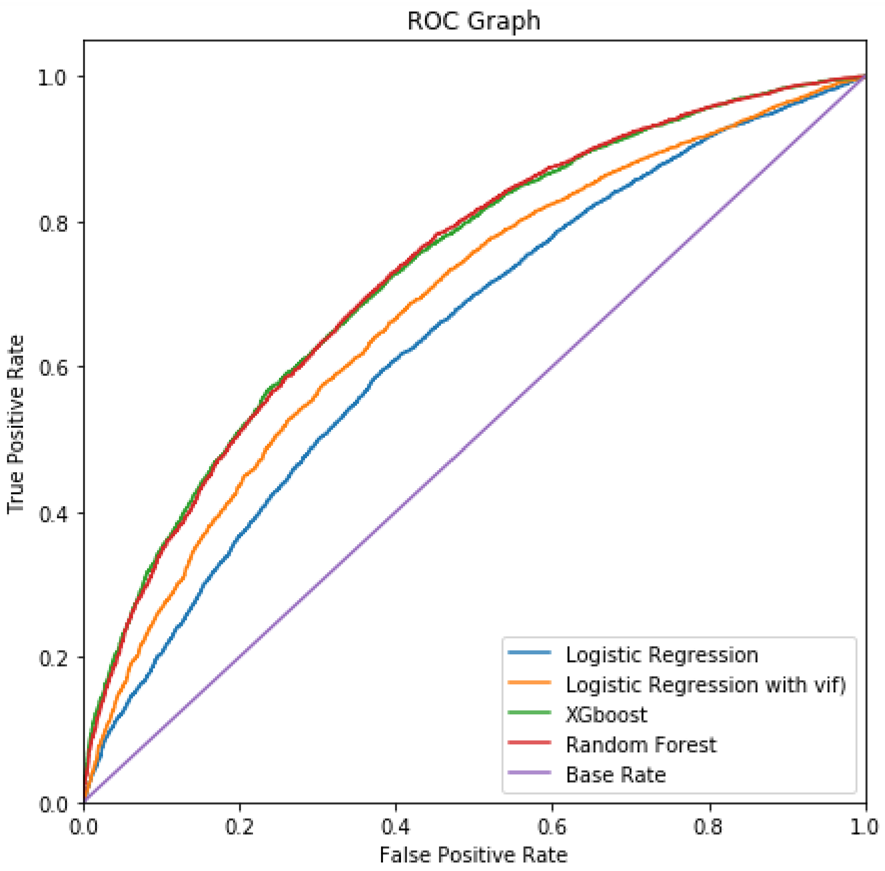
FEATURE IMPORTANCE USING RANDOM FOREST :



Based on the random forest algorithm the best features we got were kw_avg_min, LDA_01, self_reference_avg_sharess, LDA_02 ,self_reference_min_shares , kw_max_avg , kw_avg_avg .On which most of the features are the number of avg keywords with minimum or maximum shares.

Attribute	pval	0(not popular)	1(popular)
kw_avg_min	9.7121e-75	288.903	336.453
self_reference_avg_sharess	1.7029e-211	4890.604	7952.958
self_reference_min_shares	3.3586e-201	2991.107	5033.188
kw_max_avg	3.1187e-278	5207.127	6119.258
kw_avg_avg	0.0	2925.0296	3352.291

Roc curve :



Comparison to benchmark

Compared to our base model which was Logistic Regression we got an accuracy of 63.9%.But on further working with different models accuracy increased to 73.1% with random Forest on using RandomizedSearchCV. And we got the most important features which are contributing to the prediction.

LIMITATIONS

- As one of the feature of the dataset is URL on which text mining can be performed which would improve the accuracy. But as we don’t have knowledge of text mining we couldn’t use that.
- According to our dataset we needed more features relevant to shares prediction which would contribute to accuracy of the classification models.
- Most of the dependent and the independent features had no linear relationship so it was hard to come to proper inferences.
- LDA_00, LDA_01, LDA_02, LDA_03, LDA_04 topics are undefined topical categories. Since they do not specify which topic they refer to, they cannot be used to focus on a particular topic.

CONCLUSIONS

In this project, we are working on retail analytics, using a dataset on Mashable a media company which publishes blogs and articles. The main objective is to predict whether the article will be popular or not depending on various features like the number of words the article consists or the number of images, videos or links were shared.

While working on visualising the features we found that the there were not much correlation between the dependent and independent feature but when each columns were checked against shares we got some insights like when the title had neither less nor more number of words in the title, that particular article got more number of shares. Another observation was when the words in content were in the range 6-15 the articles were popular.

On applying various classification models along with different feature selection we found that random forest got the best accuracy out of all the models with an accuracy of 73.1%. As the data was balanced so we didn’t use any sampling technique. We used IQR and zscore method to remove outliers.

For our model we are considering recall rather than precision as we want to focus on articles where the number of shares were more .Thus focusing on the articles which predicted as popular and were actually popular. So accordingly we will focus on those articles to see what are the factors which will increase the popularity of the articles.