A

PROJECT REPORT

ON

'Predicting the Popularity of Online News'

SUBMITTED TO,

DEPARTMENT OF STATISTICS, SHIVAJI UNIVERSITY,

KOLHAPUR.

FOR THE PARTIAL FULFILLMENT OF THE DEGREE OF

MASTER OF SCIENCE

IN

STATISTICS

SUBMITTED BY,

Mr. JADHAV SAYAJI DATTATRAY

Under the guidance,

Mr. S. V. Rajguru.

DEPARTMENT OF STATISTICS,

SHIVAJI UNIVERSITY, KOLHAPUR.

2017-2018.

CIRTIFICATE

This is to certify that the project entitled '***Predicting the Popularity of Online News',*** being submitted by **Mr. Jadhav Sayaji Dattatray**, as partial fulfilment for the award of degree of M. Sc.in Statistics of Shivaji University Kolhapur, is a record of bonafide work carried out by him under my supervision and guidance.

To the best of my knowledge the matter presented in the project has not been submitted earlier.

Mr. S. V. Rajguru	**Dr. D. N. Kashid**
Project Guide.	**Head of the Department,**
	Department of statistics,
	Shivaji University, Kolhapur.

Place: Kolhapur
Date:

Acknowledgement

We express our thanks and deepest gratitude goes to Mr. Rajguru.S.V, who made valuable guidance and co-operation during our project work and we also specially thankful to Head of Department Dr. D.N.Kashid sir, for providing all valuable facilities and also thankful to Dr.D.T. Shirke sir, Dr. H.V. Kulkarni madam, Dr. Mahadik sir, Dr. Sakate sir, Mr. Pawar sir and also thankful to our Research students for giving us moral support.

The completion of this project could not have been possible without the participation and assistance of so many people whose names may not all be enumerated. Their contributions are sincerely appreciated and gratefully acknowledged.

Finally, we give sincerest thanks to all friends, relatives and others who shared their support morally, physically and financially.

Above all, to the great almighty, the author of knowledge and wisdom, for his countless love.

Yours Sincerely,

Mr. Jadhav Sayaji Dattatray
Department of Statistics,
Shivaji University, Kolhapur.

INDEX

❖ Introduction and description of the problem:

In the digital world, online news is primary source of information. Reading and sharing news have become the center of people's entertainment lives.

As we know, most people get information and knowledge from news and articles. In this era, people are also used to using through the internet to do everything. So, it is no doubt that online news and articles are playing a very important role in our daily life. We can get any news we want through internet quickly. Also, it is much easier to figure out which online news or articles we like through many internet outlets, such as shares, likes and comments.

As we can imagine, popular news can make the authors become famous, also it can help the social media company attract more people. So, they can make more profits. So, if an author can know what can make news or articles become popular, or one company can predict whether news or articles will be popular before them are published, they will definitely try their best to get the information.

In this project, based on the dataset including 39,643 news articles from website Mashable, we will try to find the best classification learning algorithm to accurately predict if a news article will become popular or not prior to publication.

❖ The Dataset and its variables:

The data set is downloaded from UCI Machine Learning repository and this summarizes the heterogeneous set of features about the article published by Mashable in period of two year, from 7-jan-2013 to 7-jan-2015.

Mashable.Inc is a digital media website founded in 2005 for news and blogging. As of November 2015, it has over 6,000,000 twitter follower and over 3,200,000 fans on Facebook.

The dataset contains 61 features (60 attribute & 1 target variable) and 39797 Observations.

I grouped them in the following categories:

- **Dependent Variable:** Number of shares of an article.
- **Document related:** Internal and external links, contains videos/images, number of words, topic modelling with LDA, sentiment analysis
- **Channel Type:** Entertainment, Lifestyle, Business, Social media, Technology.
- **Publish date**: Weekdays, days passed since the publishing date

The other details of the all the variable are attached to the appendix.

With these categories, we have 60 features to visualize and analyse the relationship with the number of shares (dependent variable). We can create two ML problems with this dataset:

1. Regression: Predict the number of shares.
2. Classification: Will the article be popular?

We define ‘popularity’ by establishing a minimum threshold to the number of shares. For this example: if the number of shares is above the median, it is **Popular** otherwise **Unpopular**. Using the median as threshold helps with the balance of the dependent variable for the classification project.

❖ Objective:

- I started analyze the data from exploratory data analysis. For this I established some objectives for EDA:
    - Understand which are the main features that motivate a user to share an article.
    - To predict channel wise and day wise popularity of articles.
    - Analyze the correlation of the features and its interactions

- The main goal of the project is to predict the number of shares of the articles in social media.

To predict the number of shares (Popularity) I decided to use Machine learning algorithms:

1. Regression: Predict the number of shares
2. Classification: will the article be popular?

   For classification I define minimum threshold to the number of shares. That is, I used the median: if the number of shares is above the median, it is popular.

Tools and Techniques used for Analyses: -

Statistical Technique:

- Graphical representation
- Tabular representation.
- Machine learning:
    - Linear regression.
    - Logistic regression
    - Decision tree classification using CART & C4.5.
    - K-nearest neighbor classification.
    - Naïve Bayes algorithm
    - Random forest

Statistical software's:

- R.
- Microsoft Excel.

❖ Data Analysis:

Before getting into any sophisticated analysis, the first step is to do a data cleaning and EDA. Since both categorical and continuous variables are included in the data set, variable description, appropriate tables and summary statistics are provided.

➢ DATA PREPARATION

Dataset has been thoroughly checked using various summarizing techniques for identifying

1. Missing data
2. Inconsistent data
3. Inappropriate data

- While exploring the data summary of the input variables few irrelevant values in 'n_tokens_content' attribute is found. The value for 'n_tokens_content' is 0 for few records. With the business understanding the number of words in the news content can never be 0. So, I removed these rows and considered as missing data.

- Since the column 'url' and 'is_weekend 'not useful in any kind of analysis or prediction, we have removed these columns from dataset.

- I check that the data has no missing value.

- Number of shares of a particular article is directly proportional to its popularity. So, in order to predict that the article is popular or not, I have set the threshold value to 1400(median of shares) i.e., if the number of shares is greater than threshold, then the value is replaced with "1" (popular), otherwise "0" (not popular).

- The data is split in two parts as a training data and testing data. 70% of the data is used as training data and remaining data used for test the model.

➢ Exploratory Data Analysis:

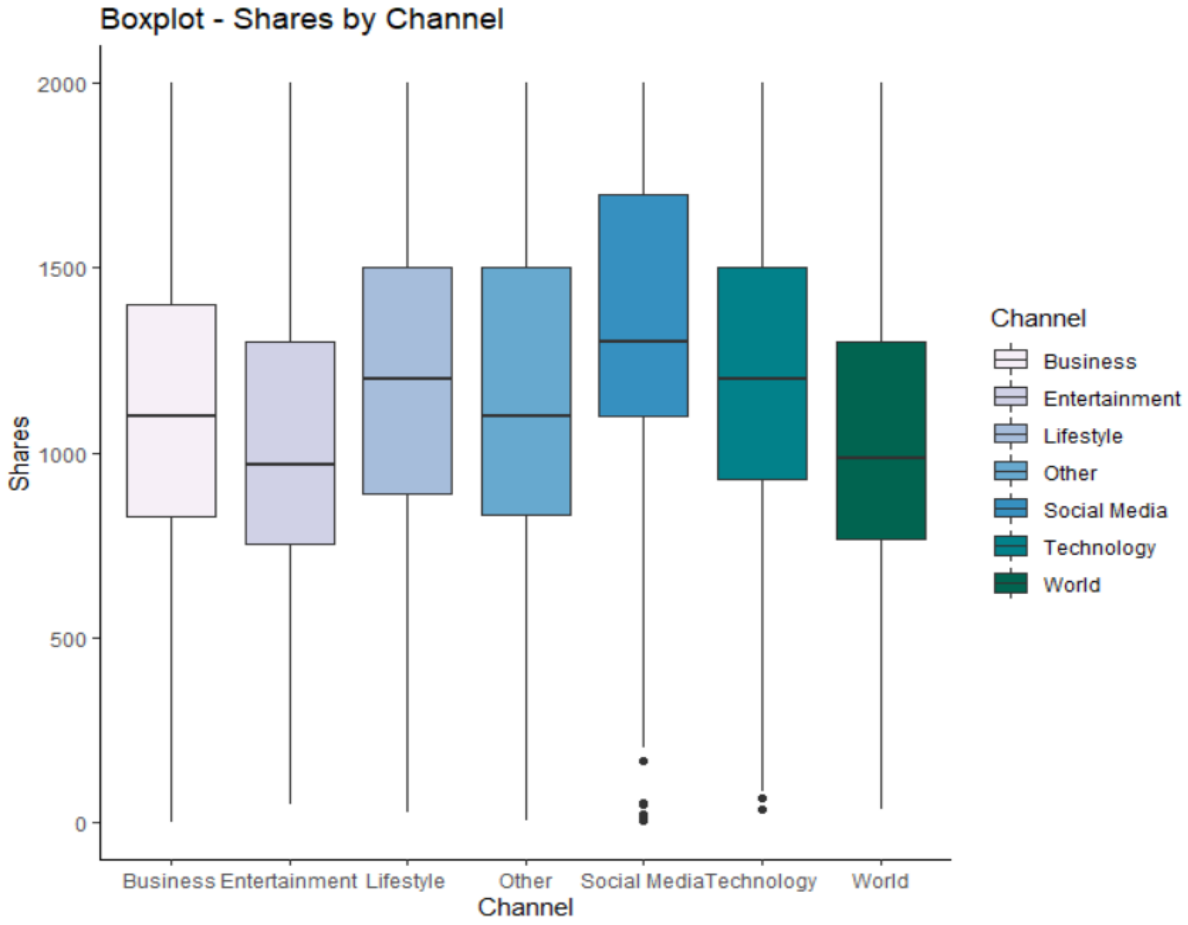
1) Channel wise popularity of Article

The dataset contains six categories for the articles written:

- Lifestyle
- Business
- Entertainment
- Technology
- Social Media
- World

Number of Shares by Channel of articles Boxplot:

The Limits have been reduced on the y-axis excluding everything above 2000 Shares

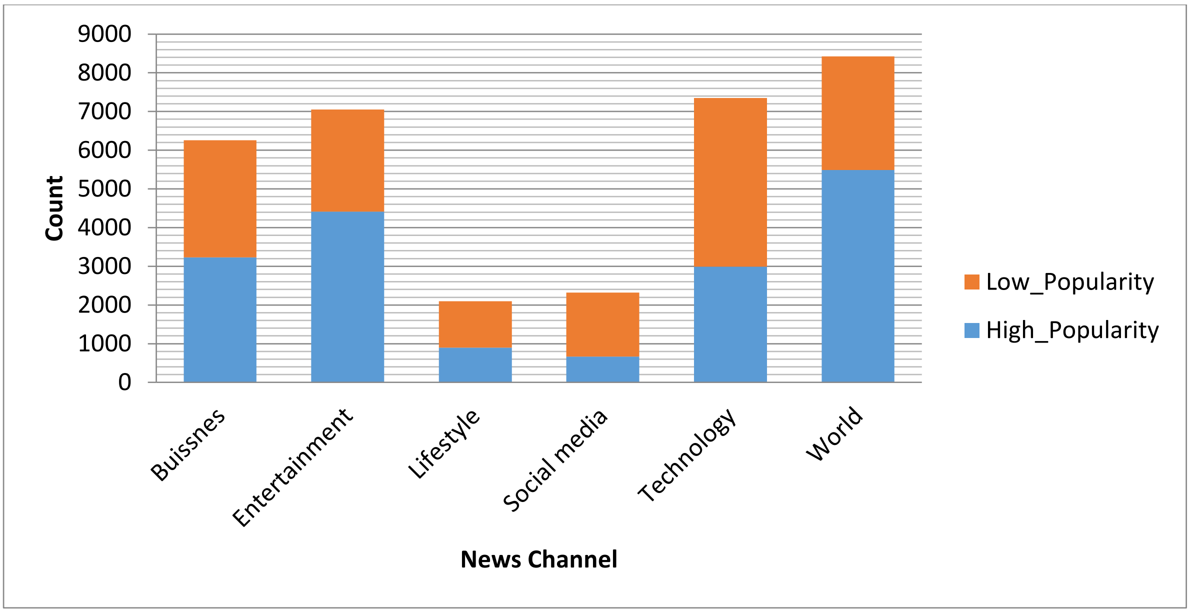


In the boxplot, the channel variable's effect is not as apparent.

The below table represent channel wise popularity of articles:

Type of Data Channel	High Popularity	Low Popularity	Percentage of high Popularity
Business	3029	3229	48.40204538
Entertainment	2643	4413	37.45748299
Lifestyle	1200	899	57.17008099
Social media	1659	664	71.41627206
Technology	4359	2987	59.33841546
World	2936	5491	34.84039397

Bar Plot for data channel:

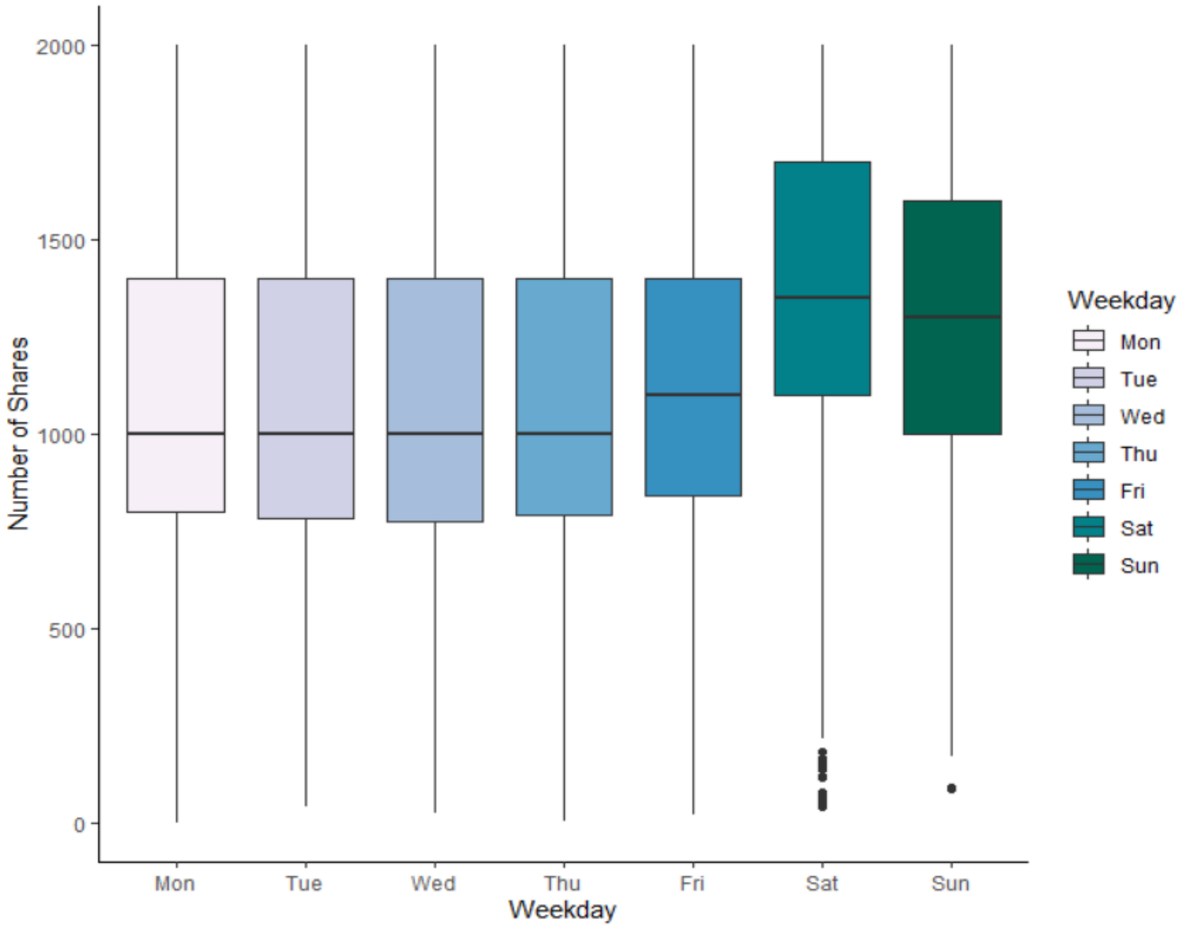


In the above bar plot, we can see how important the type of data channel is determining its popularity. Social media, Lifestyle and Technology are the most popular types of data channel. There are 49% chances that the news of category Business might be popular. News based on ‘Entertainment’ and ‘World’ are less popular (36% chances of being popular)

2) Day Wise Popularity of Article

Boxplot of Number of shares by weekdays:

The Limits have been reduced on the y-axis excluding everything above 2000 Shares

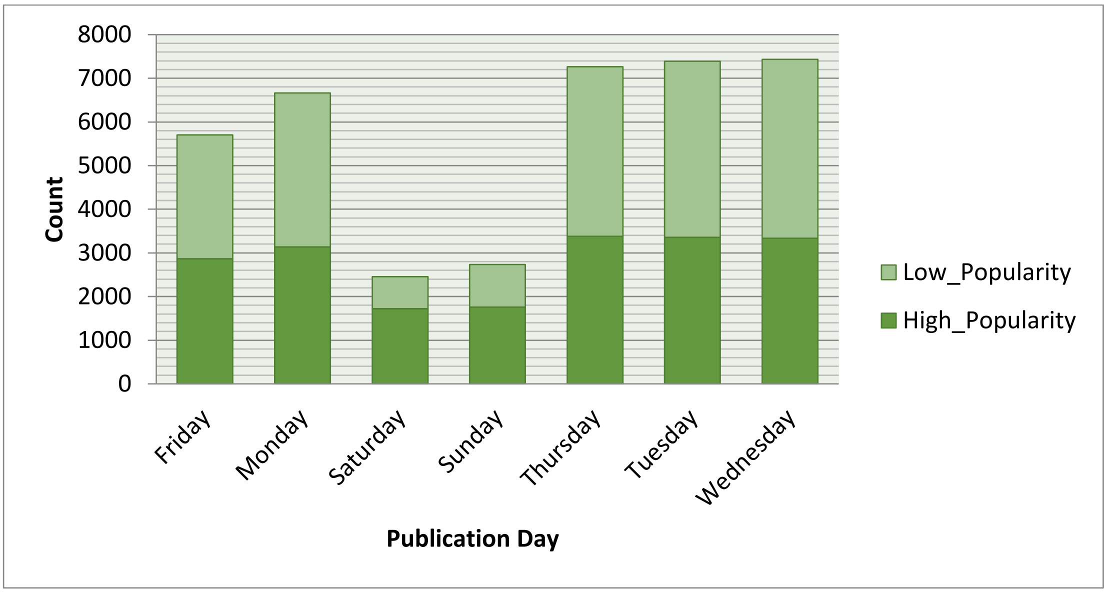


The above chart shows the effect of the weekends on the number of shares. It is clear that starting Friday, the probability of sharing increases significantly, Sunday being the highest.

<u>The below table represent Day wise popularity of Article:</u>

Publication day	High Popularity	Low Popularity	Percentage of high Popularity
Friday	2865	2836	50.25434134
Monday	3140	3521	47.14006906
Saturday	1720	733	70.11822258
Sunday	1761	976	64.34051882
Thursday	3382	3885	46.53914958
Tuesday	3358	4031	45.44593314
Wednesday	3335	4100	44.85541358

Bar Plot for data channel:



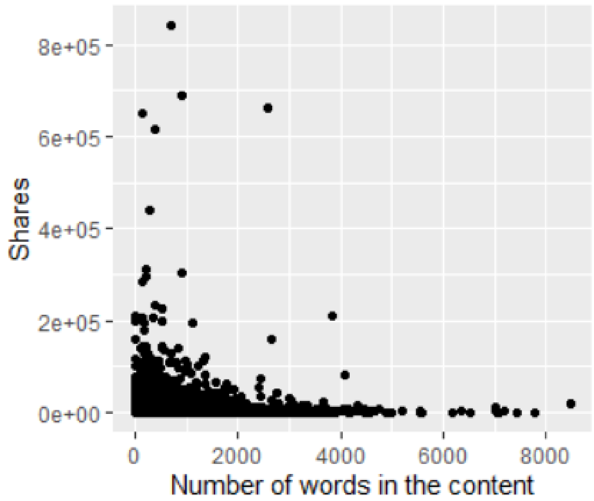
We see that news published on the weekends (i.e.: Sunday & Saturday) is more popular as compared to those published on the weekdays. For example, the majority of the news shared on Wednesday and Tuesday were unpopular.

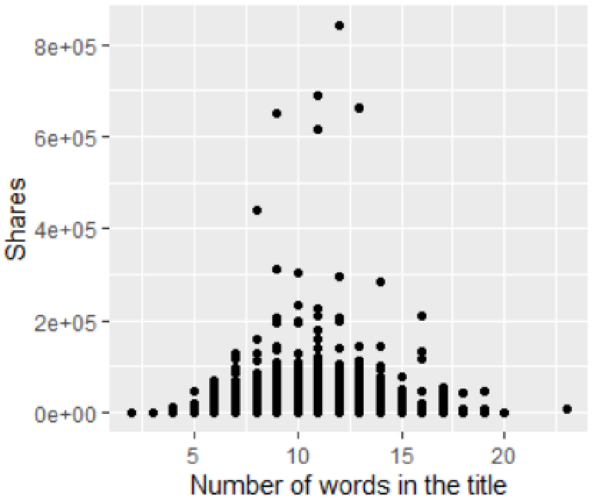
3) Which important characteristics of article’s affect the number of shares and popularity?

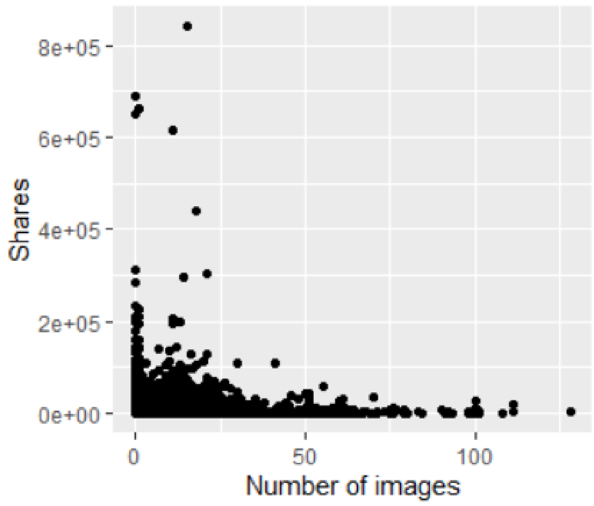
We have different quantity variables:

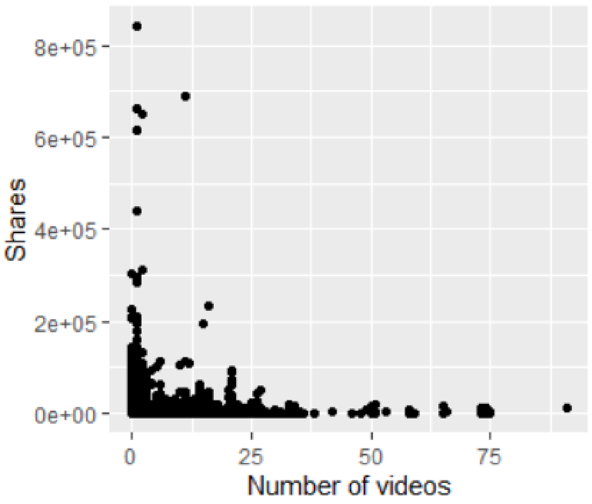
- Number of tokens in the article
- Number of tokens in the title
- Number of images
- Number of videos
- Total links
- Internal links
- Number of keywords in the metadata

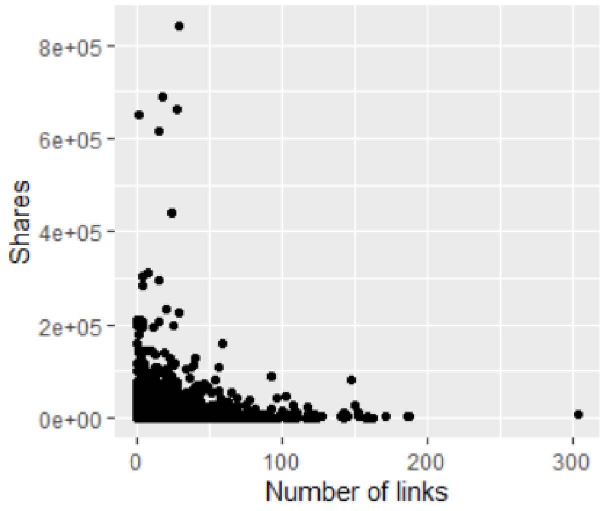
We compare the above variable with number of shares variable and check the relationship using scatter plot.

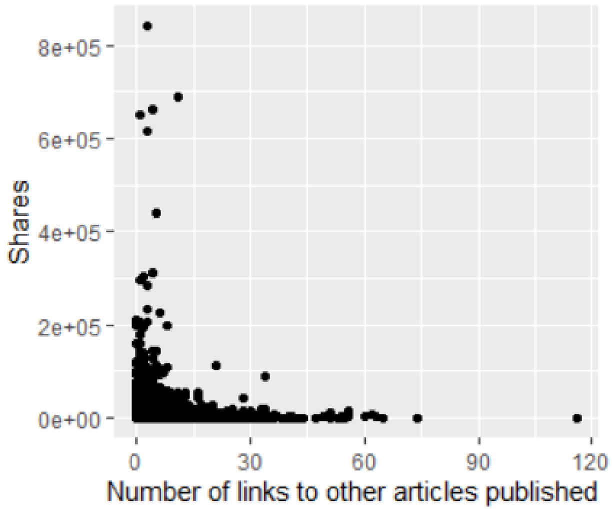


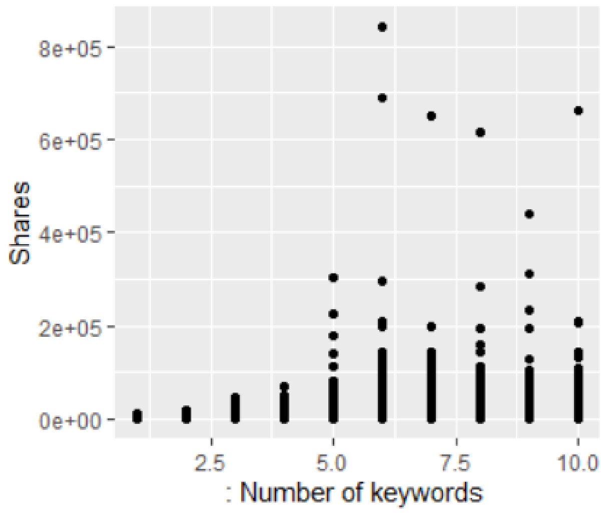












Below is the summary of the above plots:

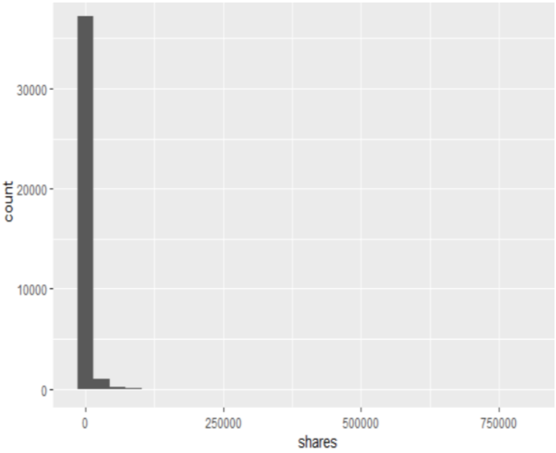
- The longer the articles, the higher chances of sharing the articles
- The longer the title or less numbers of words in the title, have the less chances of sharing the articles
- There is negative correlation between shares and Number of images and number of videos.
- If the articles have high number of links, then chances of sharing the articles (or popularity of article's) is also high
- Having a lot of internal links, may increase the chances of sharing
- The keywords in the articles are high (above 5), then chances of sharing the articles (or popularity of article's) is high.

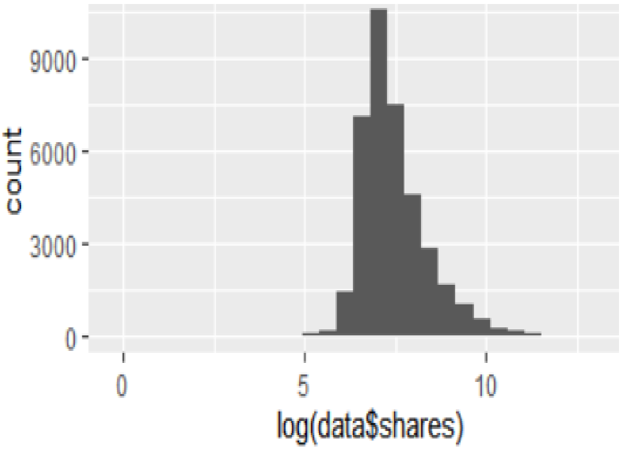
4) Summary Statistics of response (Number of shares):

Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
1	945	1400	3355	2700	843300

Skewness Analysis:

The target variable is highly (right) skewed. So, we apply the log transformation on the target variable and it will reduce the skewness. The graph below on left shows the distribution of 'shares' before applying log transformation and the right graph shows the distribution of shares after applying log transformation.





Feature selection:

Initially there are 61 attributes and we exclude the non-predictive variable 'url' and 'is_weekend' from data. Clearly, we can see that these two attributes do not possess any predictive power and cannot contribute to predicting the popularity of a new news article. So, these features have been eliminated from the data.

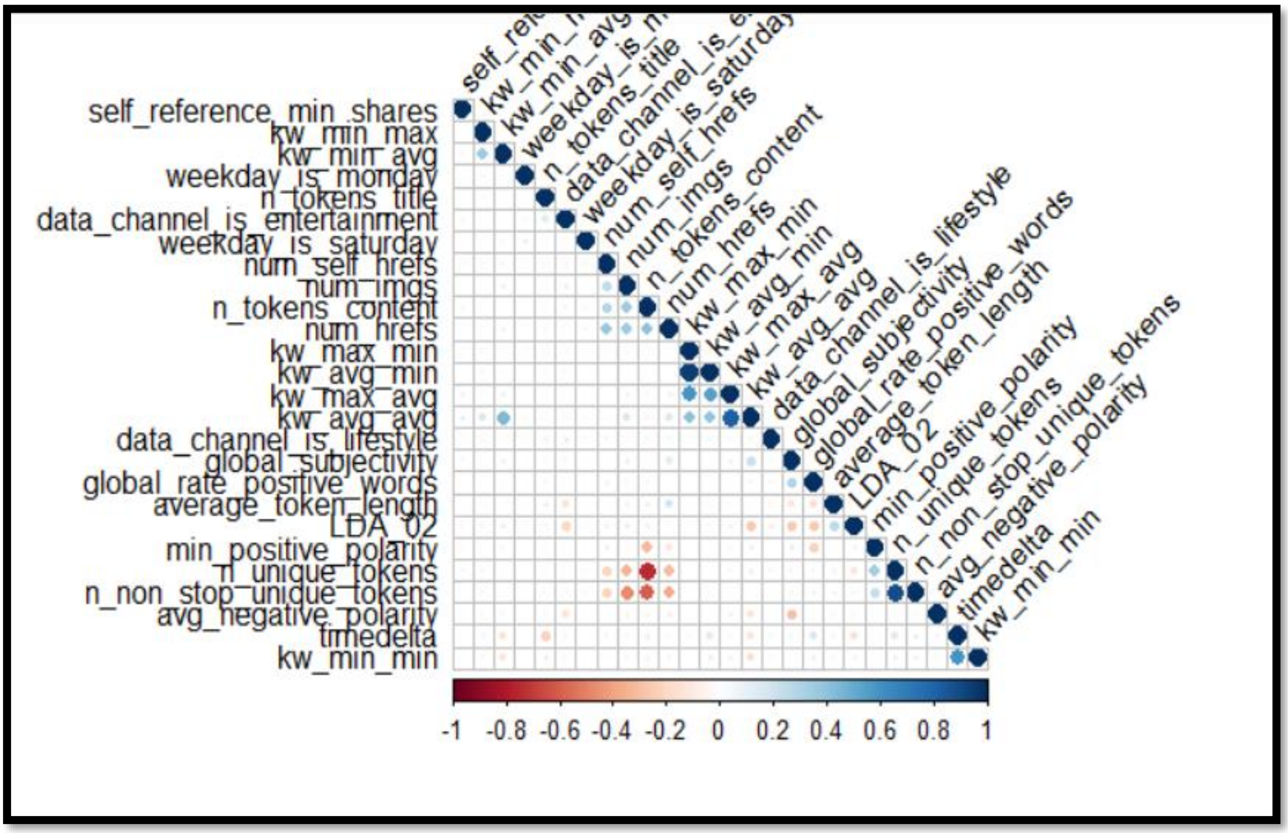
To select the most important features from the 59 features, I used stepwise regression variable section method. Stepwise selection is a hybrid of forward and backward selection. It starts with zero features and adds the one feature with the lowest significant p-value .Then, it goes through and finds the second feature with the lowest significant p-value. On the third iteration, it will look for the next feature with the lowest significant p-value, and it will also remove any features that were previously added that now have an insignificant p-value. This allows for the final model to have all of the features included be significant.

In such a way, I found the following important features.

Best features found using feature selection method:

Content and reference related attributes	n_tokens_content, n_tokens_title, timedelta, n_unique_tokens, n_non_stop_unique_tokens, num_imgs, num_hrefs, num_self_hrefs, average_token_length, self_reference_min_shares, weekday_is_Monday, weekday_is_saturday
keyword related attributes	kw_min_min, kw_max_min, kw_avg_min, kw_min_max, kw_min_avg, kw_max_avg, kw_avg_avg
LDA and category attributes	LDA_02, data_channel_is_lifestyle, data_channel_is_entertainment
NLP attributes (Sentiment analysis and Subjectivity)	global_subjectivity, global_rate_positive_words, min_positive_polarity, avg_negative_polarity

The correlation plot of the above selected variable:



❖ DATA MODELING

We prepared data by removing all the outliers and irrelevant data completely. To predict the popularity of online news we use both regression and classification algorithm. First linear regression model was applied to predict popularity and understand the problem better. Then we use classification algorithm to predict or classify whether the news is popular or not.

➢ Regression:

1) Linear Regression:

We have split the 70 % data as training data and remaining 30% data used as testing data. A linear regression model is applied to training set. And fitted model is tested on testing dataset.

We get the following regression output

Residual:

Min	1Q	Median	3Q	Max
-5.042	-0.5499	-0.1605	0.4007	5.6106

Coefficient	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.293	0.1346	54.188	< 2e-16	***
timedelta	8.87E-05	3.45E-05	2.573	0.01009	*
n_tokens_title	0.00564	0.0027	2.089	0.03674	*
n_tokens_content	4.78E-05	1.95E-05	2.457	0.01401	*
n_unique_tokens	0.09464	0.1647	0.575	0.56557	
n_non_stop_unique_tokens	-0.3214	0.1434	-2.241	0.02505	*
num_imgs	0.000806	0.00082	0.983	0.32578	
num_hrefs	0.004863	0.000601	8.085	6.5E-16	***
num_self_hrefs	-0.00336	0.001581	-2.127	0.0334	*
average_token_length	-0.1155	0.0221	-5.226	1.74E-07	***
data_channel_is_lifestyle1	-0.06567	0.02491	-2.637	0.00838	**
data_channel_is_entertainment1	-0.2896	0.01529	-18.94	< 2e-16	***
kw_min_min	0.000818	9.95E-05	8.221	< 2e-16	***
kw_max_min	7.18E-06	5.08E-06	1.414	0.15726	
kw_avg_min	-5.4E-05	3.05E-05	-1.786	0.07406	.
kw_min_max	-6.6E-07	1.03E-07	-6.429	1.31E-10	***
kw_min_avg	-4E-05	6.75E-06	-5.936	2.96E-09	***
kw_max_avg	-3.6E-05	2.15E-06	-16.64	< 2e-16	***
kw_avg_avg	0.000291	1.12E-05	25.925	< 2e-16	***
self_reference_min_shares	2.31E-06	2.76E-07	8.393	< 2e-16	***

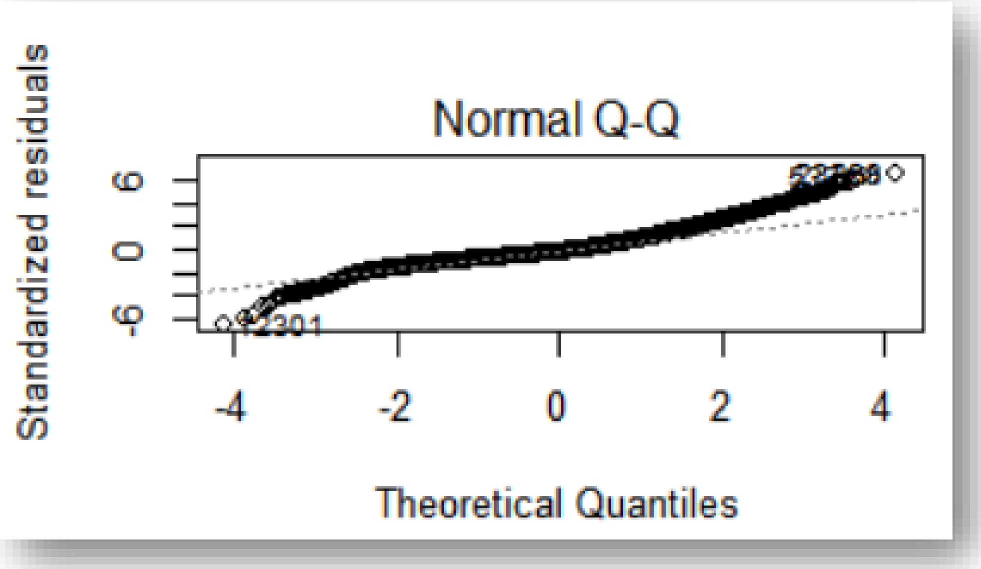
weekday_is_monday1	0.02627	0.01467	1.791	0.07336	.
weekday_is_saturday1	0.2756	0.02287	12.052	< 2e-16	***
LDA_02	-0.3152	0.02351	-13.41	< 2e-16	***
global_subjectivity	0.4434	0.07278	6.093	1.13E-09	***
global_rate_positive_words	-0.371	0.3886	-0.955	0.3397	
min_positive_polarity	-0.4273	0.09029	-4.732	2.23E-06	***
avg_negative_polarity	-0.01549	0.04809	-0.322	0.74741	

Residual standard error: 0.8715 on 25614 degrees of freedom
Multiple R-squared: 0.114, **Adjusted R-squared: 0.1131**
F-statistic: 126.7 on 26 and 25614 DF p-value: < 2.2e-16

The RMSE and Adjusted R-squared for testing data is

Root Mean square error	Adjusted R-squared:
0.8754	**0.1131**

So linear regression model is built to predict the exact value of the target variable. Since the correlation was very mere value the linear regression model couldn’t fit the target value. The linear model produced the mean R-Square value of 11.31%, which is very low, which undeniably explains the variance in the data. Also, the residual quantile plot doesn't have a normal distribution as shown below.



➢ Classification:

Since the linear regression model will not be able to provide accurate solutions. Now we use classification algorithm to predict the popularity of articles using some machine learning algorithm. We split the 70 % data as training data and remaining 30% data used as testing data. All the news articles having shares greater than 1400 are considered as Popular news and those having shares lesser than 1400 are considered as Unpopular news.

Here response variable is,

Shares = 1 ; if News are popular(Shares>1400)
= 0 ; if News are unpopular(Shares<1400)

2) Logistic Regression:

The confusion matrix is:

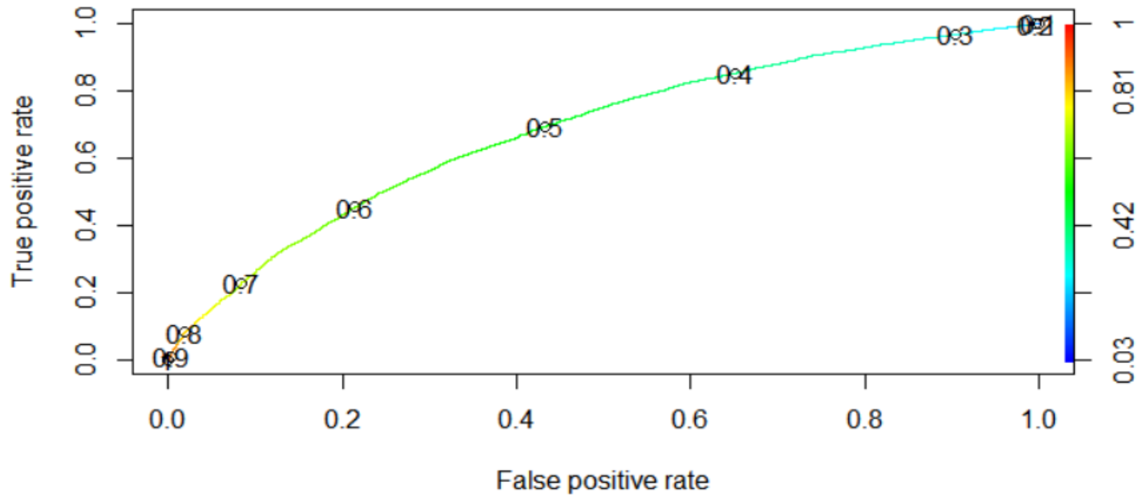
	Predicted value		
Actual value	**0 (Unpopular)**	**1 (Popular)**	**Total**
0 (Unpopular)	**3395**	**2583**	**5978**
1 (Popular)	**2105**	**4740**	**6845**
Total	**5500**	**7323**	**12823**

Model performance measure:

Accuracy	**(3395+4740)/(3395+2583+2105+4740)**	**63.44%**
Sensitivity	**3395/(3395+2583)**	**56.79%**
Specificity	**4740/(4740+2105)**	**69.24%**

From the above table, it’s evident that, the logistic regression model predicts the two category results with a balanced accuracy of 63.44%. Accuracy curve for different cut-offs is as below, so to maximize the accuracy the threshold limit is chosen as 0.5. With this threshold Confusion matrix and the ROC curve(area=0.6799) have been studied.

ROC Curve:



From this "Receiver Operating Characteristic" curve we get an area under the curve is 0.6799, it indicates that performance of model is good.

3) K-nearest neighbors (KNN):

The KNN or k-nearest neighbors algorithm is a type of instance-based learning, where new data are classified based on stored, labelled instances.

The confusion matrix:

	Actual Value		
Predicted Value	**0**	**1**	**Total**
0	**3101**	**1340**	**4441**
1	**2877**	**5505**	**8382**
Total	**5978**	**6845**	**12823**

Model performance measure:

Accuracy	**67.11%**
Sensitivity	**51.87%**
Specificity	**80.42%**

3) Decision Tree: Using Classification and Regression Tree (CART) Algorithm

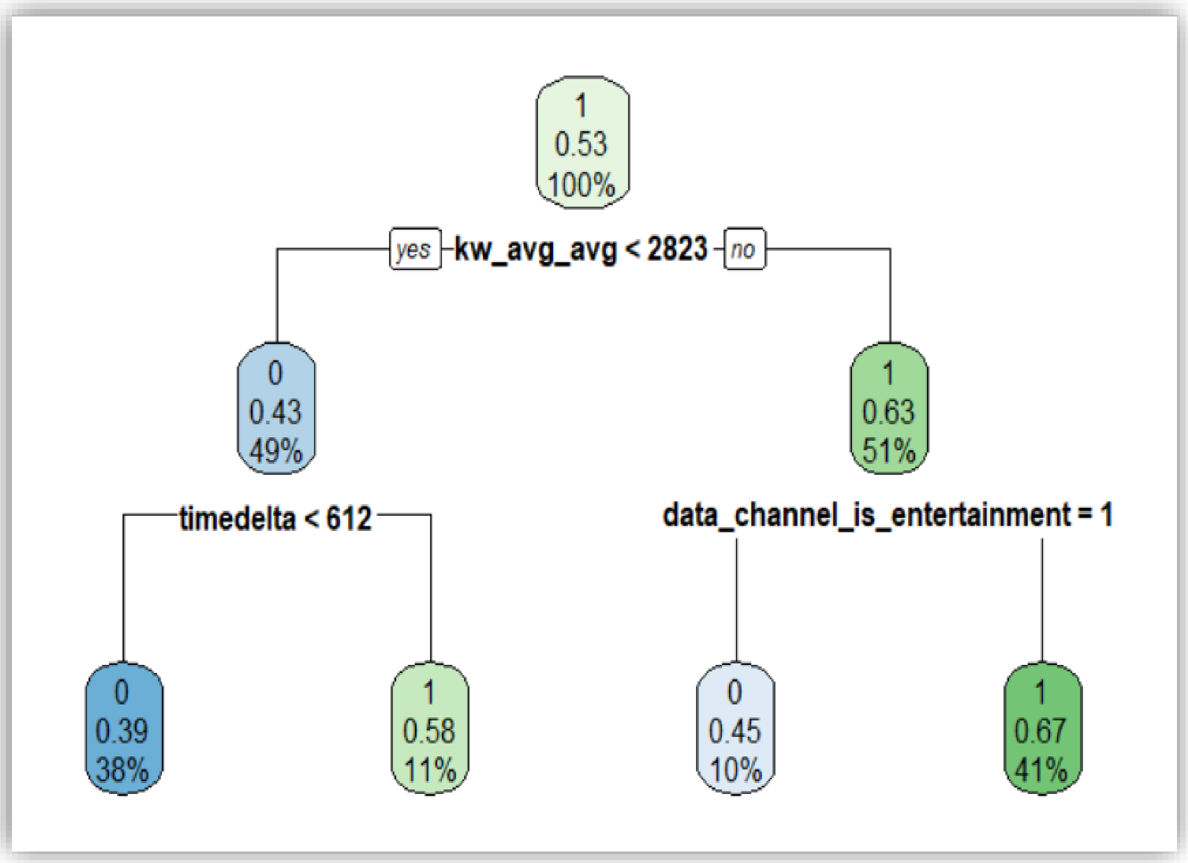
The confusion matrix is:

	Actual Value		
Predicted Value	**0**	**1**	**Total**
0	**3551**	**2537**	**6088**
1	**2427**	**4308**	**6735**
Total	**5978**	**6845**	**12823**

Model performance measure:

Accuracy	**61.29**
Sensitivity	**59.40%**
Specificity	**62.94%**
Balanced Accuracy	**61.17%**

Decision Tree:



4) Decision Tree: Using C5.0 Algorithm

C5.0 is decision trees and rule-based models for Predictions.

The confusion matrix is:

	Actual Value		
Predicted Value	**0**	**1**	**Total**
0	**3459**	**1862**	**5321**
1	**2519**	**4983**	**7502**
Total	**5978**	**6845**	**12823**

Model performance measure:

Accuracy	**65.83%**
Sensitivity	**57.86%**
Specificity	**72.80%**
Balanced Accuracy	**65.33%**

5) Naïve Bayes Algorithm:

Naive Bayes was used as baseline for feature selection since it assumes all the features are conditionally independent. I perform Naïve Bayes classifier again for selected 26 features.

The result is below:

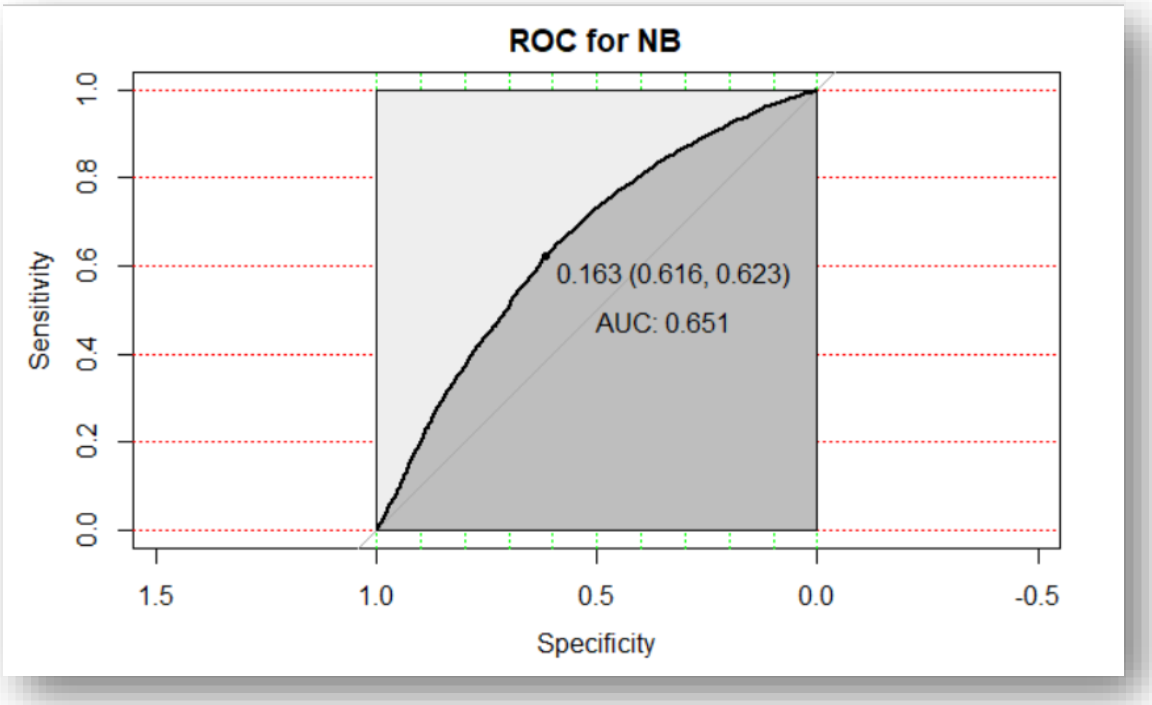
The confusion matrix is:

	Actual Value		
Predicted Value	**0**	**1**	**Total**
0	**5115**	**4886**	**10001**
1	**863**	**1959**	**2822**
Total	**5978**	**6845**	**12823**

Model performance measure:

Accuracy	**55.17%**
Sensitivity	**85.56%**
Specificity	**28.62%**
Balanced Accuracy	**57.09%**

ROC Curve



6) Random Forests:

We used random forests to generate more advanced model of decision trees which implemented the ensemble methods

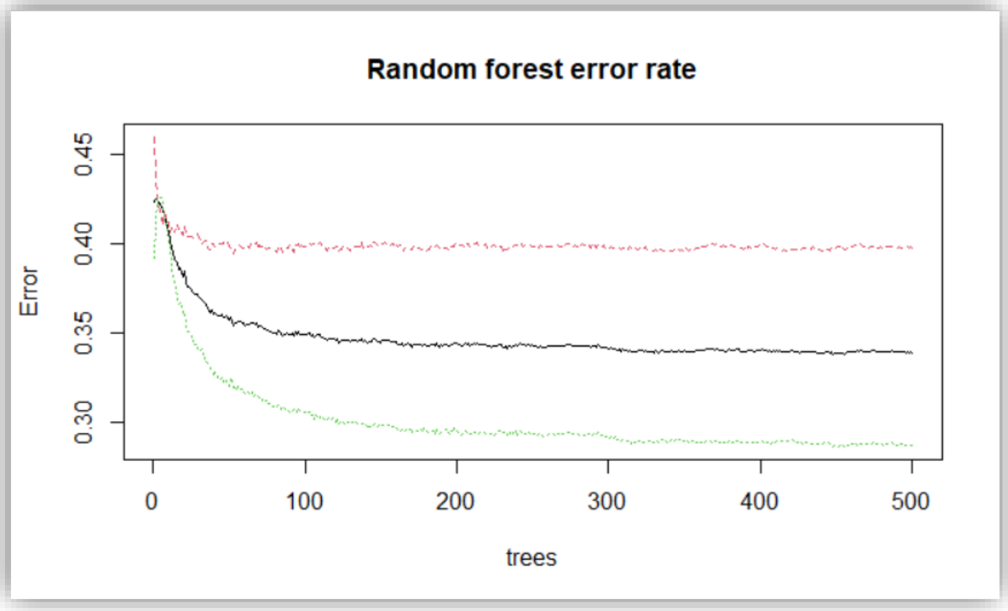
The result of random forest classifier to finding out the news popularity by using the confusion matrix is:

	Actual Value		
Predicted Value	0	1	Total
0	3604	1937	5541
1	2374	4908	7282
Total	5978	6845	12823

Model performance measure:

Accuracy	66.38%
Sensitivity	60.29%
Specificity	71.70%
Balanced Accuracy	65.99%

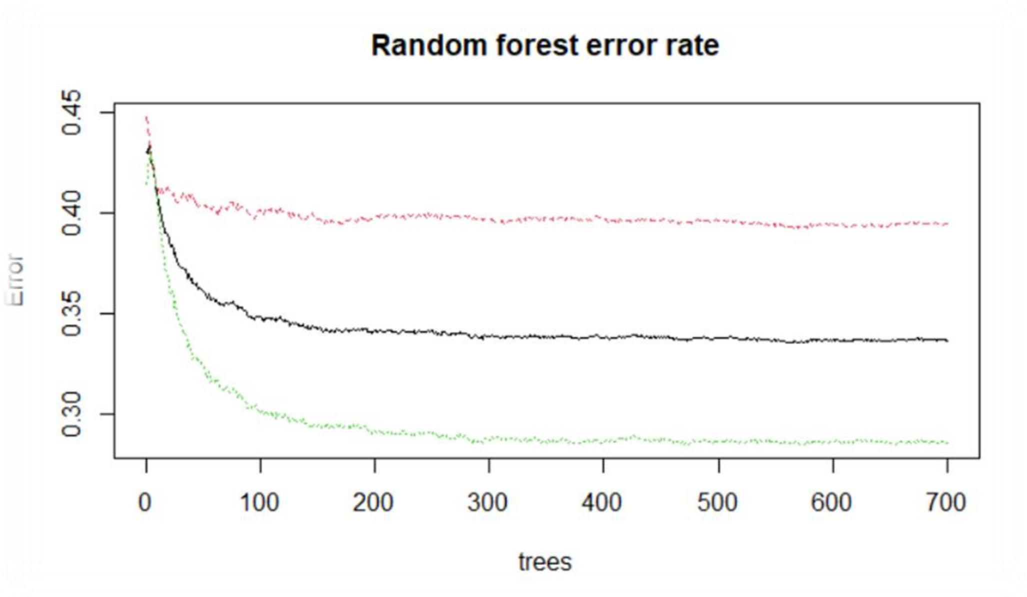
The random forest had better accuracy and lower error rate as compared to decision trees. If we generate more number of trees in random forest the, error rate goes on decreasing. We have created random forest 500 decision trees and 8 splits.



Here the black line represents the error rate for our test set in general while the green and red line represent the error rate of Yes and No label respectively. All three seem to be decreasing with the increase in the number of trees in random forest. But after a certain number of trees, the error rate line seems to be becoming flatter.

So, I ran the same train and test set for 700 trees in random forest. Which gave us an accuracy of 66.68%, which gives a slight improvement from 500 trees.

Below is plot of error rate, when 700 trees in random forest.

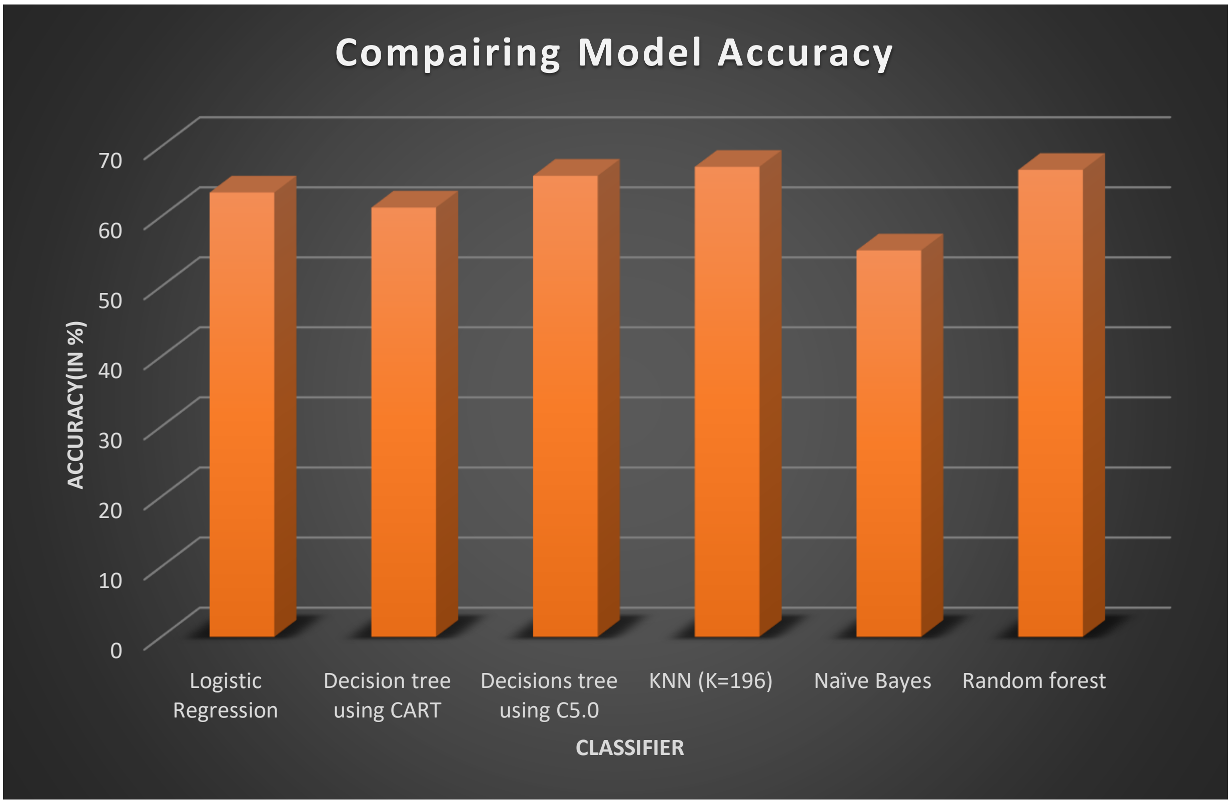


In the above plot we can see that, the decrease in error rate for random forest starts to approach 0 after the number of trees has crossed a certain count.

Performance of different classifier:

Following table shows the performance of different classifier

Classifier	Accuracy	Sensitivity	Specificity
Logistic Regression	63.44	56.79	69.24
Decision tree using CART	61.29	59.4	62.94
Decisions tree using C5.0	65.83	57.86	72.8
KNN (K=196)	**67.11**	**57.87**	**80.42**
Naïve Bayes	55.17	85.56	28.62
Random Forest	**66.68**	**60.29**	**72.24**



We observed that the random forest algorithm and KNN algorithm perform best for this dataset as compare to other classifiers. Decision tree using C5.0 algorithm and logistic regression also gives good accuracy. Moreover, the accuracy of Naïve Bayes gives poor accuracy.

CONCLUSION:

The goal of the project is the predicting the popularity of the news article. We use data cleaning and data pre-processing and prepare the best data for data modelling. We used regression and classification on the data. Since the linear model couldn't produce better results because of the variance in the data. Then classification algorithms are applied. We found that random forest and KNN performs slightly better than other classifiers. While Naïve bayes perform poorly on data.

The performance of Random forest and KNN are good for prediction of the popularity of online news article.

The conclusion based on Exploratory data analysis is:

- Data channel of type Social media, Lifestyle and Technology are the most popular types of data channel. While data channel Entertainment and World are less popular (36% chances of being popular).
- News published on the weekends is more popular as compared to those published on the weekdays.
- The longer the articles, the higher chances of sharing the articles
- The longer the title, the less chances of sharing the articles
- There is negative correlation between shares and Number of images and number of videos.
- If the articles have high number of links, then chances of sharing the articles (or popularity of article's) is also high
- Having a lot of internal links, may increase the chances of sharing
- The keywords in the articles are high (above 5), then chances of sharing the articles (or popularity of article's) is high.

Scope for further study:

For better prediction we also use to try another classification technique like SVM, AdaBoost, etc.

This problem can also be tackled as a multiclass classification problem; we can have 5 classes i.e., obscure (1) articles that are shared very few times, mediocre (2), popular (3), super popular (4) and viral (5).

In this project 'number of shares' of online news articles on social networking website is used as the variable for prediction of popularity. We can also use any other attribute for predicting the popularity.
We can also use any other variable selection method to select most important features from data like PCA.

References:-

- *The data set for the analysis is taken from the UCI repository.*
  Dataset https://archive.ics.uci.edu/ml/datasets/online+news+popularity

- *K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 — Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.*

-

Appendix

Attribute Information:

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field

0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.
Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information.
Subjectivity is also a float which lies in the range of [0,1]
Keywords are words that capture the essence of your **paper**. **Keywords** make your **paper** searchable and ensure that you get more citations. Therefore, it is important to include the most relevant **keywords** that will help other authors find your paper

Code:

Project_Code.txt