CIND 820 Big Data Analytics Project

Predicting the Popularity of Online News

Supervisor: Tamer Abdou

Jaime Ip 024783334

July 17, 2023

Ryerson University

Contents

Abstract

Getting news information online is a vital part of everyone's life. People can build influence by posting online. It raises people's interest to understand what make an online news popular and how to improve the level of popularity. The theme of this capstone project is forecasting the popularity of online news prior to publication based on a broad set of extracted features. The dataset for the project is Online News Popularity Dataset from the UCI Machine Learning Repository. The dataset is originally acquired and pre-processed by K. Fernandes et al. It extracts 61 attributes describing different aspects of each article, from a total of 39,644 articles published in Mashable website from 2013 to 2014. This project is focused on: 1) the relationship between the popularity and the features of article, 2) the best model for predicting the popularity of online news and 3) selecting the set of features to optimize the performance of the model. Python is the primary tool that is used in this project. Data visualizations are used to reveal the relationship between the variables. The prediction is formulated as a binary classification problem and 3 classification algorithms include logistic regression, Random Forests, and SVM are implemented. Recursive feature elimination (RFE) is used to filter out the least irrelevant features. Then grid search method is used to identify the set of features to optimize the prediction result.

Dataset link: UCI Machine Learning Repository: Online News Popularity Data Set

GitHub link: GitHub - j7ip/CIND820CapstoneProject

Literature Review

Expansion of internet has changed the way how people consume news. Most people now get their news information online instead of traditional media. That explains why prediction of online news popularity becomes a trendy research topic. K Fernandes et al. [1] formulate the prediction question as a binary classification task and propose an Intelligent Decision Support System (IDSS) to analyze the popularity of the articles. By extracting the features of articles, the IDSS first predicts if an article will become popular. Then, it optimizes a subset of the articles features to enhance the predicted popularity probability. Based on their evaluation, Random Forest generates the best result with a discrimination power of 73%. The best optimization method can make a mean gain improvement of 15 percentage points in terms of the estimated popularity probability. Ren and Yang [2] use K Fernandes et al.'s dataset and implement 10 different machine learning models. They apply 5-fold cross validation to compare the performances of these models. PCA and filter methods (mutual information and Fisher criterion) are used for feature selection. Random Forest is also best model for prediction in their evaluation of models. It achieves an accuracy of 70% with optimal parameters. Zhang [3] proposes a three- layer neural network and tries to improve the performance of the system by using feature scaling, bimodal distribution removal and evolutional algorithm. Feature scaling has improved the testing accuracy by nearly 15%. Combining evolutional feature selection with bimodal distribution removal can also increase the score by 4%. The final system achieves 70% accuracy, which is 3 % higher than the comparing approach conducted by K. Fernandes et al.

Tatar, Dias de Amorim, Fdida and Antoniadis [4] review the current studies relate to prediction of popularity and point out 3 areas that need to be addressed in future studies. These areas include predicting long term popularity evolution, building richer models, and beyond popularity predictions. Keneshloo, Wang, Han and Ramakrishnan [5] build regression model to forecast the popularity. They engineer several classes of features, which include metadata, contextual or content-based, temporal, and social. The evaluation shows that metadata features are the most important factor for predicting the performance

of news articles in general. In Zhang and Lin's study [6], the popularity is categorized into 3 levels (high, middle, and low) and PCA is applied for dimension reduction. They create a model based on Random Forest and assess the accuracy by the ROC value area.

In this project, I would replicate K. Fernandes et al. study to formulate the prediction question as a binary classification task and recognize a subset of features that can optimize the prediction result. Logistic regression is recognized by Kirasich et al. [7] as a method that consistently performs with a higher accuracy than random forest when the variance in the explanatory and noise variables increases. As this method is not used in K. Fernandes et al.'s study and this dataset has a large set of variables, I plan to have an experiment of using logistic regression in this project and see if it can generate better prediction result.

Data Description and EDA (Exploratory Data Analysis)

The Online News Popularity dataset is from the UCI Machine Learning Repository (UCI Machine Learning Repository: Online News Popularity Data Set). It is originally acquired and pre-processed by K. Fernandes et al.. It includes data extracted from 39,644 articles published in Mashable website during 2013-2014. Each sample has 61 variables. 60 of them are the features of the articles and 1 is target variable (i.e. number of shares which represents the popularity of the article). The feature set is categorized in Table 1. Based on the dataset statistic in Pandas Profiling Report, there is no null value in this dataset. 75% of the articles are shared not more than 2,800 times and median is 1,400 times.

Table 1: List of Features by Category

Category	**Features**
Words	Number of words in the title/article, average word length, rate of non-stop words/unique words/ unique non-stop words
Links	Number of links Number of Mashable article links Min/avg/max number of shares of Mashable links
Digital Media	Number of images/videos

Time	Day of the week Published on a weekend
Keywords	Number of keywords Worse keyword (min./avg./max shares) Average keyword (min./avg./max shares) Best keyword (min./avg./max shares) Article category (Mashable data channel)
Natural Language Processing	Closeness to top 5 LDA topics Title subjectivity/sentiment polarity Article text subjectivity/polarity score and its absolute difference to 0.5 Polarity of positive/negative words (min./avg./max.) Rate of positive and negative words Pos. words rate among non-neutral words
Target	Number of article shares

Table 2: Pandas Profiling Report - Dataset statistics from

Dataset statistics

Number of variables	61
Number of observations	39644
Missing cells	0
Missing cells (%)	0.0%

Variable types

Categorical	1
Numeric	60

Table 3: Pandas Profiling Report - Data statistics of variable – number of shares of the articles

shares

Real number (ℝ)

Distinct	1454	Minimum	1
Distinct (%)	3.7%	Maximum	843300
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	3395.380184	Memory size	309.6 KiB

More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1
5-th percentile	584
Q1	946
median	1400
Q3	2800
95-th percentile	10800
Maximum	843300
Range	843299
Interquartile range (IQR)	1854

Descriptive statistics

Standard deviation	11626.95075
Coefficient of variation (CV)	3.424344291
Kurtosis	1832.672657
Mean	3395.380184
Median Absolute Deviation (MAD)	600
Skewness	33.96388488
Sum	134606452
Variance	135185983.7
Monotonicity	Not monotonic

As the dataset provides the number of shares but not the category of popularity (i.e. popular vs unpopular), I need to determine the threshold for number of shares as a popular news article online. The median of predictable variable (i.e.,1,400) is selected as the threshold to convert the target variable from numerical to categorical.

There are a lot of features in this dataset. Including all in the model will lead to overfitting problem and decrease the accuracy of the result. It is necessary to evaluate and eliminate unneeded or redundant features. By reviewing Table 1, I recognize some features may be related to the number of shares. These features include article category, published time of article, and number of words in the article. In general, people tend to select the articles related to what they are interested in and prefer articles that are not too lengthy. Besides, it is likely that people have more time to read and share more articles over weekend than weekdays. Figure 1 shows a higher proportion of popular news in four types of news, particularly technology and social media. However, entertainment and world news are less popular among Mashable readers and have higher unpopular proportion. Figure 2 shows a higher proportion of popular news published on the weekend. Among weekdays, Friday has the highest proportion of popular news and it is close to weekend. Scatter plot in figure 3 shows a negative correlation between the number of words in an article and number of shares i.e. the more words in an article, the less number of shares. Results shown in Figure 1-3 are align with the assumption that category of articles, published time of articles, and number of words in the article are relevant in predicting the popularity.

URL of the articles and time delta (i.e. days between the article publication and the dataset acquisition) are irrelevant in prediction of popularity and can be excluded. The variable n_tokens_ content represents number of words in the article content. There are 1,181 articles have zero word in the content i.e. no content. These records can be removed as they are not relevant to the analysis.

Fig 1: Count of popular/unpopular news over different article category

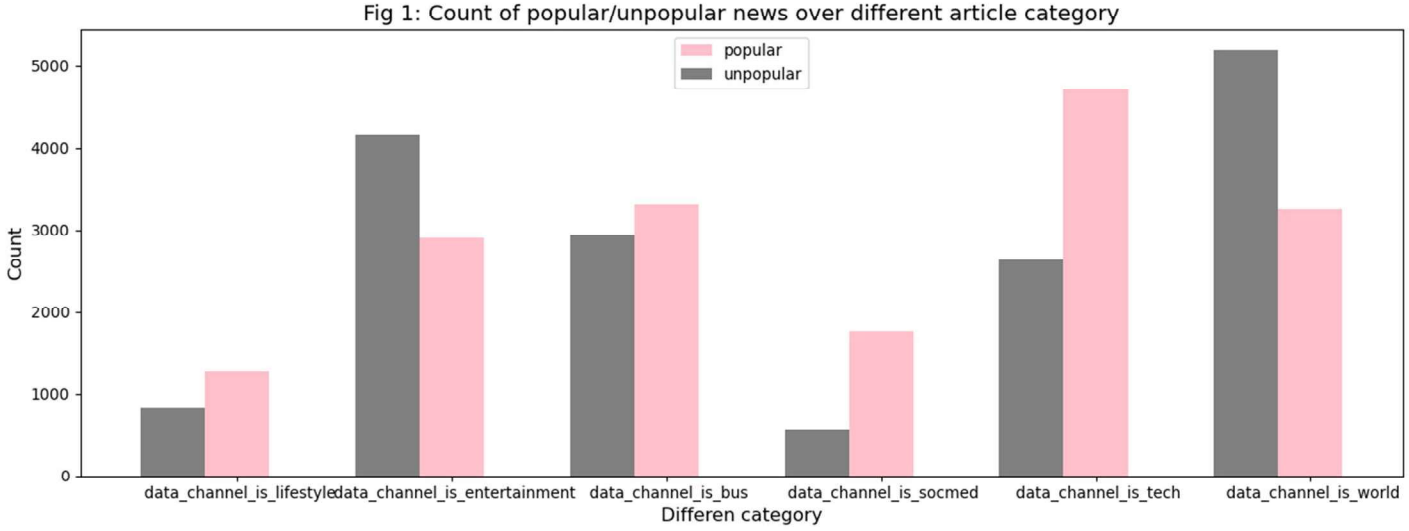


Fig 2:Count of popular/unpopular news over different day of week

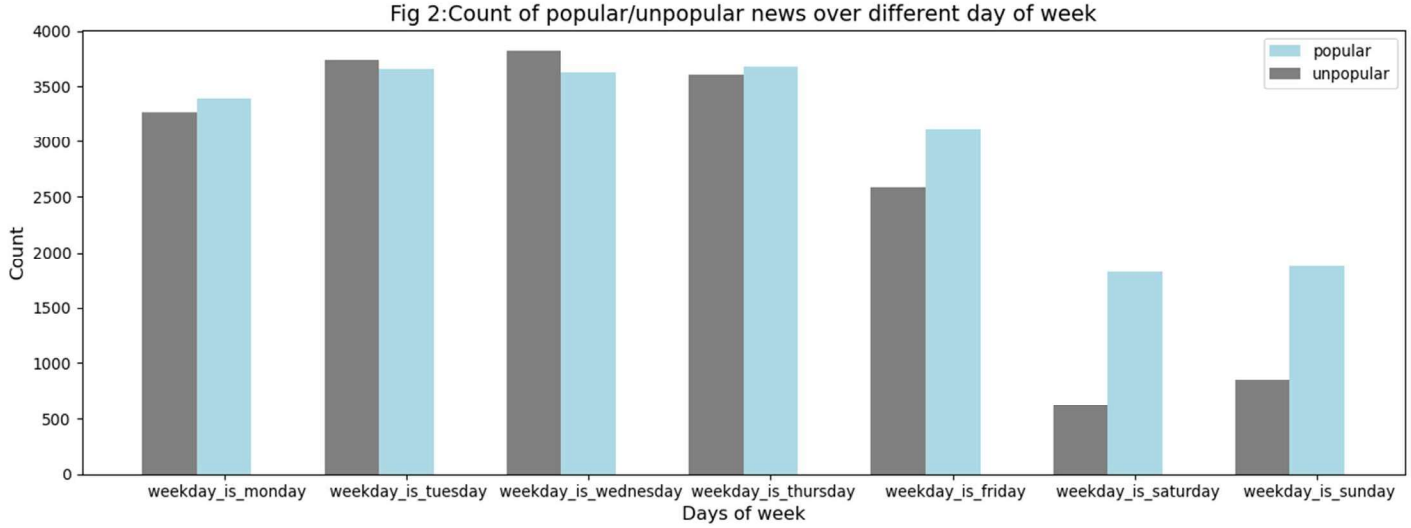
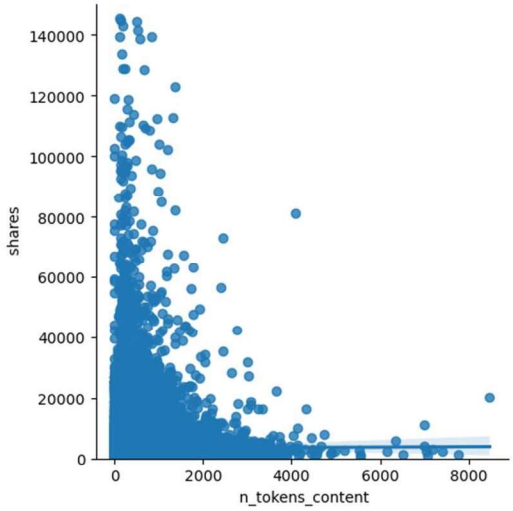
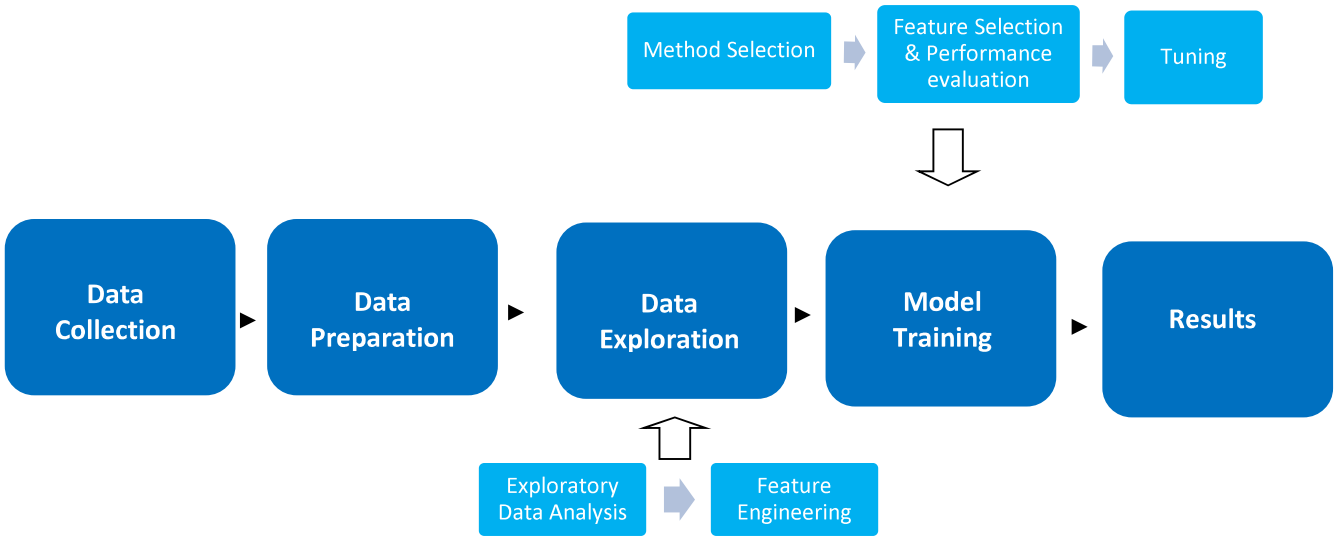


Fig 3: Scatter plot – Number of words vs Number of Shares



Approach



Step 1: Data collection

The dataset is obtained from UCI Machine Learning Repository. It is originally acquired and pre-processed by K. Fernandes et al. It extracts 61 attributes describing different aspects of each article, from a total of 39.644 articles published in Mashable website from 2013- 2014.

Step 2: Data Preparation

Panda Profiling Report is run to get the data statistics overview of the dataset. The report provides data included missing value, outliners, quartile statistics and histogram for each variable. Cleaning data is also required to ensure the accuracy in the prediction result.

This dataset does not have any missing value or duplicate value. All the column headings have an appended space on the left. These spaces may lead to errors when running codes in python and hence are also removed.

The target variable is number of shares and the median is 1,400. A threshold for number of shares as a popular news is required and the median is selected.

Step 3: Data Exploration

The first research question is the relationship between the popularity and the features of article is focused in this step. By understanding their relationship and characteristics of the features, the appropriate classification models can be selected.

3.1 Exploratory Data Analysis:

Data visualization is used to analysis the relations between features and target variable. Bar charts and scatter plot indicate that number of shares is related to article category, published day, and number of words.

The variable n_tokens_ content represents number of words in the article content. There are 1,181 articles have zero word in the content i.e. no content. These records are removed as they are not relevant to the analysis.

The comprehensive Exploratory Data Analysis is in previous section Data Description and EDA (Exploratory Data Analysis).

3.2 Feature Engineering:

As classification models are used in this project, the target variable is converted from numerical to categorial and saved as a new variable shares_level. Data is divided into 3 classes in the variable shares level. Group 1 is unpopular (0-950 shares). Group 2 is normal (950- 1400 shares) and Group 3 is popular (>1400 shares). The proportion of these 3 groups:

Group	Proportion
1 - Unpopular (0-950 shares)	25.3%
2- Normal (950-1,400 shares)	25.5%
3 - Popular (>1,400 shares)	49.2%

There are also 14 categorical features, which can be divided into 2 groups:

Day of week	Article category
weekday_is_monday	data_channel_is_lifestyle
weekday_is_tuesday	data_channel_is_entertainment
weekday_is_wednesday	data_channel_is_bus
weekday_is_thursday	data_channel_is_socmed
weekday_is_friday	data_channel_is_tech
weekday_is_saturday	data_channel_is_world
weekday_is_sunday	
is_weekend	

The value of these categorical features is either 1 or 0. Feature "is_weekend" is not necessary as columns for each day in weekend are exist. The other 7 features for day of the week (Monday – Sunday) are consolidated into 1 new single feature named "weekday". (1=Monday, 2= Tuesday, 3= Wednesday, 4= Thursday, 5=Friday, 6=Saturday, 7=Sunday). Same process is done for the other 6 features related to article category (1=lifestyle, 2=entertainment, 3=bus, 4=socmed, 5=tech, 6=world, 0= not belong to any of these 6 categories) and the consolidated feature is named as data_channel. The total number of categorical features reduce from 14 to 2 after consolidation.

Step 4: Model Selection

The second research question is selecting the best model for prediction, which is the focus in this step. First, the dataset is split into training and testing sets (training 70%: 30% testing). Three classification algorithms include 1) Logistic Regression 2) Random Forests and 3) SVM are selecte. Normalizing features is necessary before training and testing the models. As the

dataset has outliners and negative values, Min Max Scaler and Box-Cox are not optimal methods. Standard Scaler is used for this dataset.

Fig 4: Distributions of variables

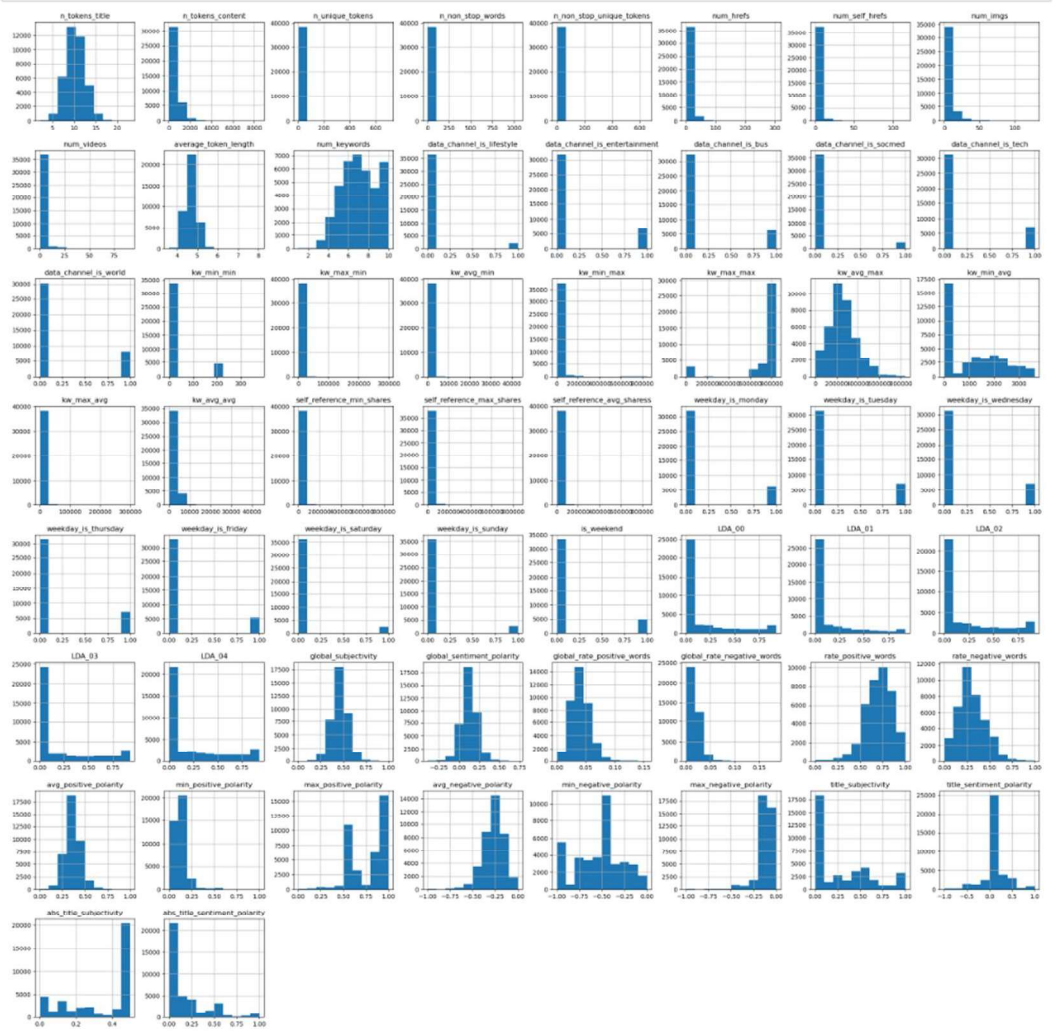


Fig 5: Features contain negative values

```
# Finding negative values.

negcols=df2_num.columns[(df2_num<0).any()]
negcols
```

```
Index(['kw_min_min', 'kw_avg_min', 'kw_min_avg', 'global_sentiment_polarity',
       'avg_negative_polarity', 'min_negative_polarity',
       'max_negative_polarity', 'title_sentiment_polarity'],
      dtype='object')
```

4.1 Method Selection

K Fernandes et al. formulate the prediction question as a binary classification task and conclude Random Forest generates the best result. As I would replicate their study to formulate the prediction question as a classification task, Random Forest is selected for this project. Kirasich et al. compared Random Forest and Logistic Regression for binary classification in their study in 2018. They concluded that Logistic Regression method consistently performs with a higher accuracy than random forest when the variance in the explanatory and noise variables increases. Hence, Logistic Regression is also selected and compared to the result generated by Random Forest model. Another popular classification algorithm, SVM, is also used in the testing.

The drawbacks of converting the target variable to categorical and using classification methods include losing a level of details and misinterpretation. When the threshold to categorize the number of shares as popularity is changed, the result can be very different. As this project is focused on replicating K. Fernandes et al,'s study, using the same approach would allow me to have a direct comparison to the original study result. Hence, classification approach is chosen.

4.2 Feature selection & Performance evaluation

A. First testing

Models are built and run with full set of features in the first testing. Based on the metrics, the models need to be adjusted in order to get a better result, especially Group 2.

Random Forest

	precision	recall	f1-score	support
1	0.48	0.39	0.43	2904
2	0.35	0.10	0.16	2968
3	0.58	0.85	0.69	5667
accuracy			0.54	11539
macro avg	0.47	0.45	0.43	11539
weighted avg	0.50	0.54	0.49	11539

Logistic Regression

	precision	recall	f1-score	support
1	0.48	0.27	0.35	3619
2	0.00	0.00	0.00	2253
3	0.53	0.89	0.66	5667
accuracy			0.52	11539
macro avg	0.34	0.39	0.34	11539
weighted avg	0.41	0.52	0.44	11539

SVM

	precision	recall	f1-score	support
1	0.00	0.00	0.00	2904
2	0.00	0.00	0.00	2968
3	0.49	1.00	0.66	5667
accuracy			0.49	11539
macro avg	0.16	0.33	0.22	11539
weighted avg	0.24	0.49	0.32	11539

Random Forest

```
Accuracy values for 10-fold Cross Validation:
 [0.32809443 0.430217   0.46463219 0.48389493 0.47594757 0.48710088
 0.49931409 0.50587444 0.52735675 0.51216105]

Final Average Accuracy of the model: 0.47
```

Logistic Regression

```
Accuracy values for 10-fold Cross Validation:
 [0.32546695 0.34423876 0.38292824 0.3804597  0.37762233 0.37969276
 0.37441732 0.3777141  0.37336387 0.38919531]

Final Average Accuracy of the model: 0.37
```

SVM

```
Accuracy values for 10-fold Cross Validation:
 [0.32398904 0.32398904 0.32398904 0.32412977 0.32412977 0.32412977
 0.32412977 0.32441613 0.32441613 0.32412977]

Final Average Accuracy of the model: 0.32
```

Feature importance list is also generated. The 10 least important features are recognized:

num_imgs	0.013443
min_positive_polarity	0.013420
abs_title_subjectivity	0.012505
abs_title_sentiment_polarity	0.012232
max_positive_polarity	0.011841
num_keywords	0.010974
num_videos	0.008331
kw_max_max	0.007345
kw_min_min	0.005792
n_non_stop_words	0.000000

B. Second testing

The least important 10 features are removed and tested in the Random Forest model. Accuracy remains at the same level after these features are eliminated. It proves that these features have minimal impact on the results and can be eliminated to improve the efficiency of running the models. Fig 5 also shows that the accuracy remains stable after the estimator number in Random Forest model reaches around 50. Current number of estimators is 100 and hence will not be increased as no improvement in accuracy is expected.

Random Forest (before removing the least important features)

	precision	recall	f1-score	support
1	0.48	0.39	0.43	2904
2	0.35	0.10	0.16	2968
3	0.58	0.85	0.69	5667
accuracy			0.54	11539
macro avg	0.47	0.45	0.43	11539
weighted avg	0.50	0.54	0.49	11539

Before

```
Accuracy values for 10-fold Cross Validation:
 [0.32809443 0.430217   0.46463219 0.48389493 0.47594757 0.48710088
 0.49931409 0.50587444 0.52735675 0.51216105]

Final Average Accuracy of the model: 0.47
```

Random Forest (After removing the least important features)

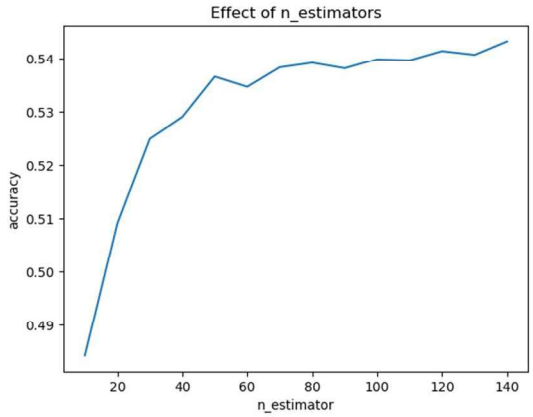
	precision	recall	f1-score	support
1	0.47	0.39	0.42	2904
2	0.34	0.11	0.16	2968
3	0.58	0.84	0.68	5667
accuracy			0.54	11539
macro avg	0.46	0.44	0.42	11539
weighted avg	0.49	0.54	0.48	11539

After

```
Accuracy values for 10-fold Cross Validation:
 [0.33805128 0.41013241 0.46950889 0.48223797 0.47821066 0.47414417
 0.49713834 0.51062035 0.51664666 0.5145445 ]

Final Average Accuracy of the model: 0.47
```

Fig 5: Relationship of accuracy and number of estimators in Random Forest Model



C. Third testing

The metrics report in the first testing indicates that Group 2 has zero in each metrics when running Logistic Regression and SVM. The scores are better but still low in Random Forest model. Group 2 is the articles that are shared 950-1400 times. Same as Group 1, it is also not classified as popular and is used to provide more details by breaking down the unpopular group into 2 levels. However, we can still predict if an article is popular (i.e. > 1400 shares) without this subgroup of unpopular articles. As a result, the shares level is revised and has 2 groups (Group 1: < 1400 and Group 2: >1400) in the third testing. The result is improved in all 3 models while Random Forest model has the best result among all. Its average accuracy in 10- fold cross validation increased from 0.47 to 0.63. Besides, there is no zero in the metrics reports for both Group 1 and 2 in all 3 models and look more reasonable than results in the first testing.

Random Forest has the best result in all 3 testings. It is the best model for predicting the popularity of online articles. Below is the summary and details of the accuracy and 10-fold cross validation result of the third testing:

	Average Accuracy - 10-fold Cross Validation	Accuracy
Random Forest	0.63	0.65
Logistic Regression	0.58	0.59
SVM	0.4	0.54

Random Forest (1st testing)

	precision	recall	f1-score	support
1	0.48	0.39	0.43	2904
2	0.35	0.10	0.16	2968
3	0.58	0.85	0.69	5667
accuracy			0.54	11539
macro avg	0.47	0.45	0.43	11539
weighted avg	0.50	0.54	0.49	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.32809443 0.430217   0.46463219 0.48389493 0.47594757 0.48710088
 0.49931409 0.50587444 0.52735675 0.51216105]

Final Average Accuracy of the model: 0.47
```

Random Forest (3rd testing)

	precision	recall	f1-score	support
1	0.65	0.67	0.66	5872
2	0.65	0.63	0.64	5667
accuracy			0.65	11539
macro avg	0.65	0.65	0.65	11539
weighted avg	0.65	0.65	0.65	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.43671524 0.58735631 0.63785726 0.652909   0.64110356 0.66855455
 0.6679816  0.67393784 0.67979784 0.6566545 ]

Final Average Accuracy of the model: 0.63
```

Logistic Regression (1st testing)

	precision	recall	f1-score	support
1	0.48	0.27	0.35	3619
2	0.00	0.00	0.00	2253
3	0.53	0.89	0.66	5667
accuracy			0.52	11539
macro avg	0.34	0.39	0.34	11539
weighted avg	0.41	0.52	0.44	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.32546695 0.34423876 0.38292824 0.3804597  0.37762233 0.37969276
 0.37441732 0.3777141  0.37336387 0.38919531]

Final Average Accuracy of the model: 0.37
```

Logistic Regression (3rd testing)

	precision	recall	f1-score	support
1	0.59	0.66	0.62	5872
2	0.60	0.52	0.56	5667
accuracy			0.59	11539
macro avg	0.59	0.59	0.59	11539
weighted avg	0.59	0.59	0.59	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.39161971 0.59800763 0.64109214 0.55347359 0.58105554 0.59925339
 0.58071078 0.60201626 0.62231344 0.59331224]

Final Average Accuracy of the model: 0.58
```

SVM (1st testing)

	precision	recall	f1-score	support
1	0.00	0.00	0.00	2904
2	0.00	0.00	0.00	2968
3	0.49	1.00	0.66	5667
accuracy			0.49	11539
macro avg	0.16	0.33	0.22	11539
weighted avg	0.24	0.49	0.32	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.32398904 0.32398904 0.32398904 0.32412977 0.32412977 0.32412977
 0.32412977 0.32441613 0.32441613 0.32412977]

Final Average Accuracy of the model: 0.32
```

SVM (3rd testing)

	precision	recall	f1-score	support
1	0.53	0.73	0.62	5872
2	0.55	0.33	0.41	5667
accuracy			0.54	11539
macro avg	0.54	0.53	0.51	11539
weighted avg	0.54	0.54	0.52	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.32398904 0.32427528 0.41008421 0.42149832 0.4371186  0.42910944
 0.4427739  0.42614476 0.41251802 0.40103047]

Final Average Accuracy of the model: 0.4
```

D. Fourth testing

The feature importance list is generated by Pandas Series and the least important 10 features are removed in the second testing. In order to locate the opportunity to further improve accuracy and efficiency of the Random Forest model, Recursive Feature Elimination (RFE) is

also run to compare to the feature importance list generated by Pandas Series.

Fig 6: Feature importance by RFE

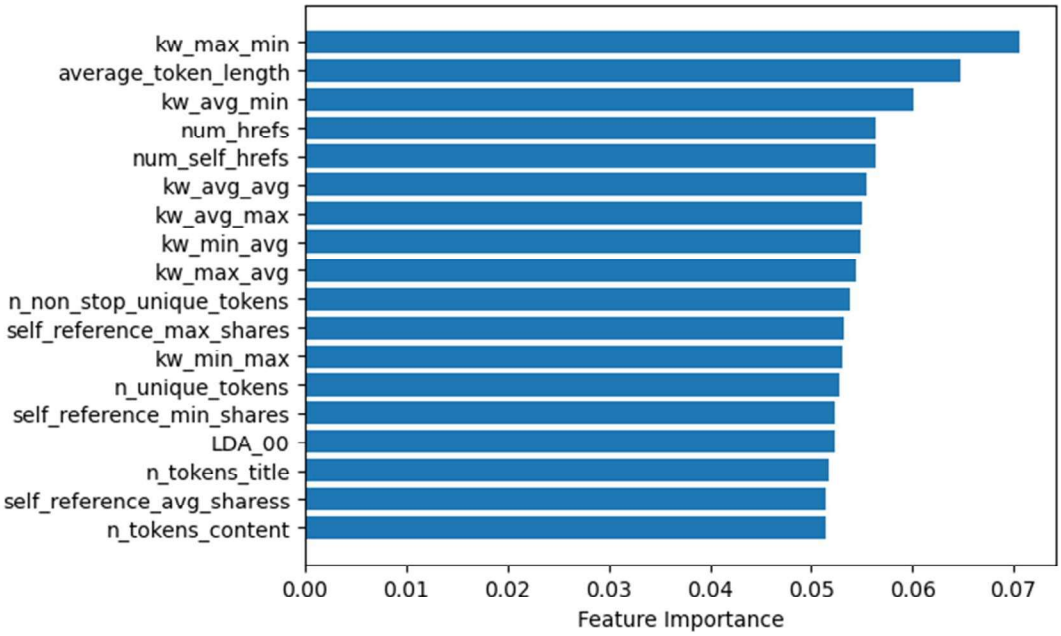


Fig 7:The least important 10 features by RFE

```
'title_sentiment_polarity', 'weekday', 'data_channel, min_negative_polarity',
       'max_negative_polarity', 'title_subjectivity, rate_positive_words,
n_tokens_title, num_self_hrefs, kw_min_max
```

Fig 8: Revised least important 10 feature importance list by Pandas after the first 10 least important are removed

rate_negative_words	0.023812
kw_min_avg	0.023808
n_tokens_title	0.019302
kw_min_max	0.018290
title_sentiment_polarity	0.018090
min_negative_polarity	0.017483
max_negative_polarity	0.016920
title_subjectivity	0.016524
data_channel	0.015558
num_self_hrefs	0.015522

There are 7 common least important features in Fig 7 and Fig 8. They are eliminated in the this testing and the below result indicates that the average accuracy of 10-fold cross validation

increase from 0.63 to 0.65.

Random Forest (3rd testing)

	precision	recall	f1-score	support
1	0.65	0.67	0.66	5872
2	0.65	0.63	0.64	5667
accuracy			0.65	11539
macro avg	0.65	0.65	0.65	11539
weighted avg	0.65	0.65	0.65	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.43671524 0.58735631 0.63785726 0.652909   0.64110356 0.66855455
 0.6679816  0.67393784 0.67979784 0.6566545 ]

Final Average Accuracy of the model: 0.63
```

Random Forest (4th testing)

	precision	recall	f1-score	support
1	0.65	0.67	0.66	5872
2	0.65	0.63	0.64	5667
accuracy			0.65	11539
macro avg	0.65	0.65	0.65	11539
weighted avg	0.65	0.65	0.65	11539

```
Accuracy values for 10-fold Cross Validation:
 [0.55414652 0.62739455 0.64463444 0.65966882 0.64297687 0.65835566
 0.66067502 0.67986555 0.67802917 0.65334519]

Final Average Accuracy of the model: 0.65
```

4.3 Tuning:

Random search method is run to determine which set model parameters give the best performance. One advantage of this method over Grid Search method is it can be more efficient if the search space is large as it only samples a subset of the possible combinations rather than evaluating them all. In these cases, it may not be feasible to explore the entire search space using Grid Search method.

Below is the best hyperparameters, accuracy, AUC and the best score found by Random Search. Accuracy of testing set is around 0.64, which are slightly lower than the result of 0.65 in 4th testing. AUC is 0.07 higher than accuracy for training set and 0.05 higher for testing set. AUC is the area under the ROC Curve with x axis represents the false positive rate and y axis represents true positive rate. Possible reasons of higher ROC than accuracy include 1) imbalanced data, 2) the threshold is not chosen correctly. Overall, the result is acceptable while still space to further improve.

```
{'criterion': 'entropy', 'max_depth': 9, 'max_features': 2, 'min_samples_leaf': 17, 'min_samples_spli
t': 3, 'n_estimators': 17}

Accuracy of Random forest train : 0.7023102065072054
Accuracy of random forest test : 0.6443365976254442
AUC of random forest train : 0.7745251412620454
AUC of random forest test : 0.6954234299729443
Best score is: 0.6406936536411617
```

Step 5: **Results**

Same as K Fernandes et al. and Ren and Yang's studies, Random Forest generates the best result in the testings while accuracy rate of 0.65 in the fourth testing and 0.64 in the refined model by Random Search are lower than K Fernandes et al. and Ren and Yang's results (0.67 and 0.70 respectively). Feature scaling, feature elimination, and modification of number of classes of target variable are used to improve the performance of the models.

As the accuracy is lower than 70% and AUC is lower than 80%, there's still some space for further improvements by testing other methods and approaches of feature scaling and elimination. Classification approach also has an inherent limitation of losing a level of details and possible misinterpretation. Also, the source of the dataset is the articles published in Mashable website from 2013 to 2014. As people's interest continue to change, what found in projects that use this dataset may not be applicable in predicting the popularity of articles that published in later years. Working on predicting long term popularity evolution in future studies as Tatar, Dias de Amorim, Fdida and Antoniadis [4] suggested can improve the performance of the model.

References

[1] Fernandes, K, Vinagre, P, & Cortez, P. (2015) A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Portuguese Conference on Artificial Intelligence.

[2] Ren, H., Yang, Q. (2015.) Predicting and Evaluating the Popularity of Online News, Department of Electrical Engineering, Standford University.

[3] Zhang, S. (2018). Online News Popularity Prediction. Research School of Computer Science, Australian National University.

[4] Tatar, A., De Amorim, M.D., Fdida,S., & Antoniadis,P. (2014) A survey on predicting the popularity of web content. Journal of Internet Services and Applications, 5(1):1–20, 2014.

[5] Keneshloo, Y., Wang, S., Han, E.H., & Ramakrishnan, N.(2016) Predicting the Popularity of News Articles. 2016 IEEE International Conference on Big Data, Wahington, DC, USA, 2016, pp. 2400-2409, doi:10. 1109/BigData.2016.7840875.

[6] Zhang, Y., Lin, K., (2021) Predicting and Evaluating the Online News Popularity based on Random Forest. Journal of Physics: Conference Series, Volume 1994, Issue 1, idf. 012040, 5 pp.

[7] Kirasish, K., Smith,T., Sadler, B. (2018) Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review: Vol.1: No.3, Article 9.