CIND 820 Big Data Analytics Project

Predicting the Popularity of Online News

Supervisor: Tamer Abdou

Jaime Ip 024783334

June 26, 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 K-nearest neighbors 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

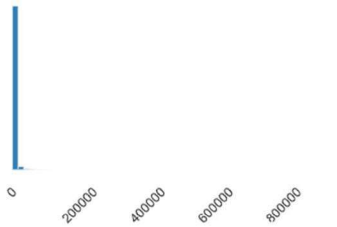
Categorical	1
Numeric	60

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 number into Boolean.

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 are removed as they are not relevant to the analysis.

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

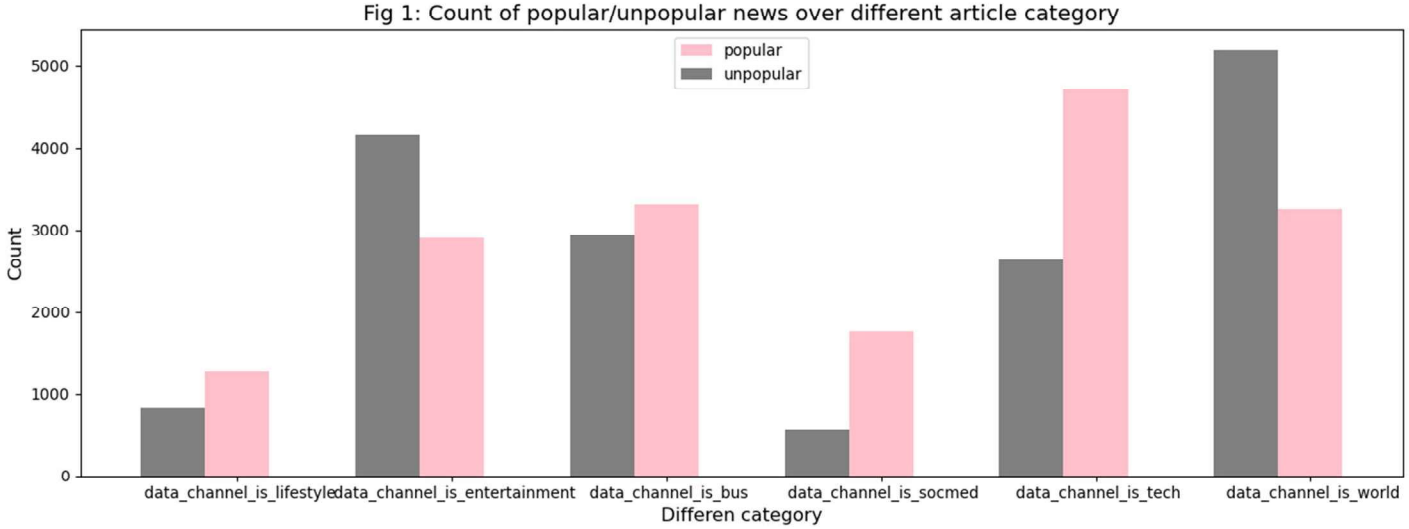


Fig 2:Count of popular/unpopular news over different day of week

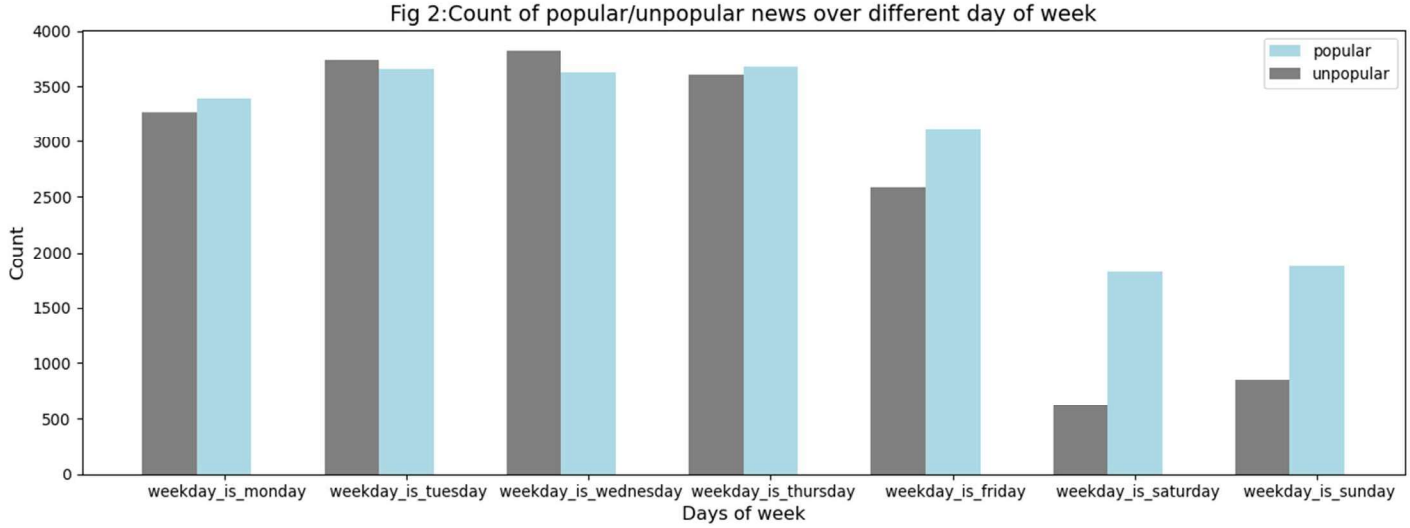
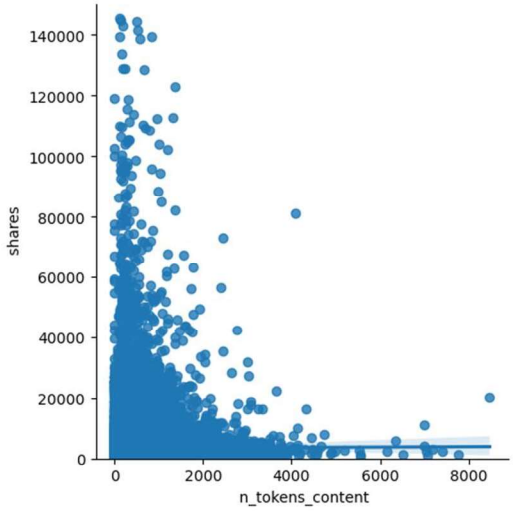
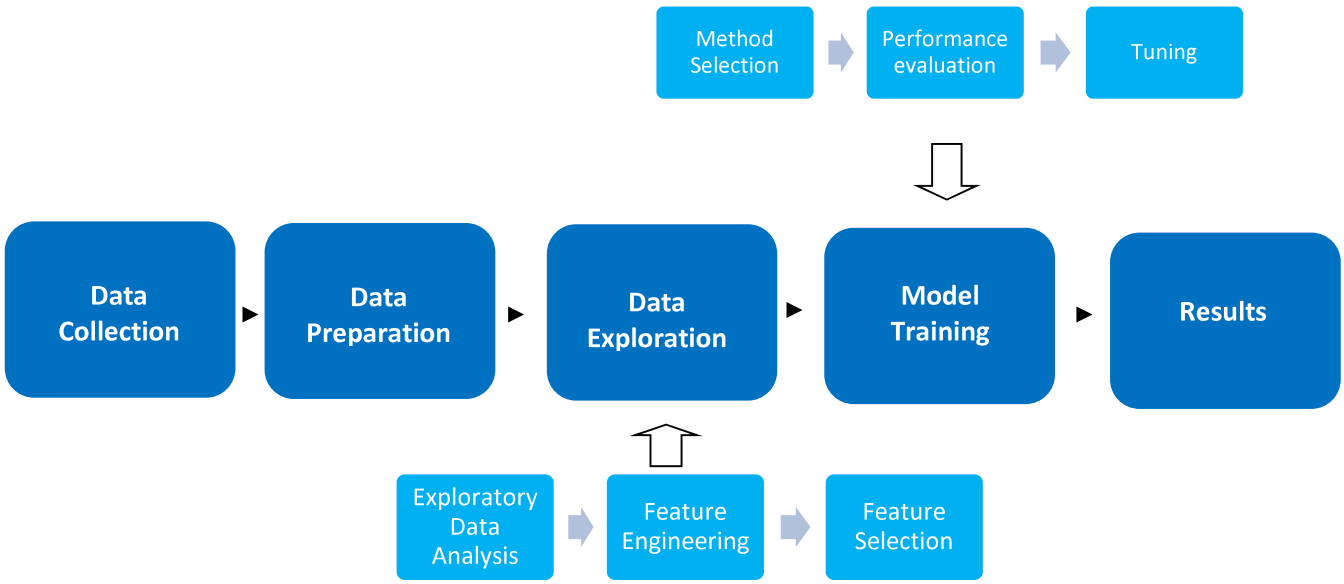


Fig 3: Scatter plot – Number of words vs Number of Shares



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 performed to ensure the accuracy in the prediction result.

This dataset does not have any missing 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 removed.

The target variable is number of shares and the median is 1,400. A threshold for number of shares as a popular news to convert the target variable from number into Boolean (i.e. popular or unpopular) is required. The median is selected as the threshold.

Step 3: Data Exploration

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).

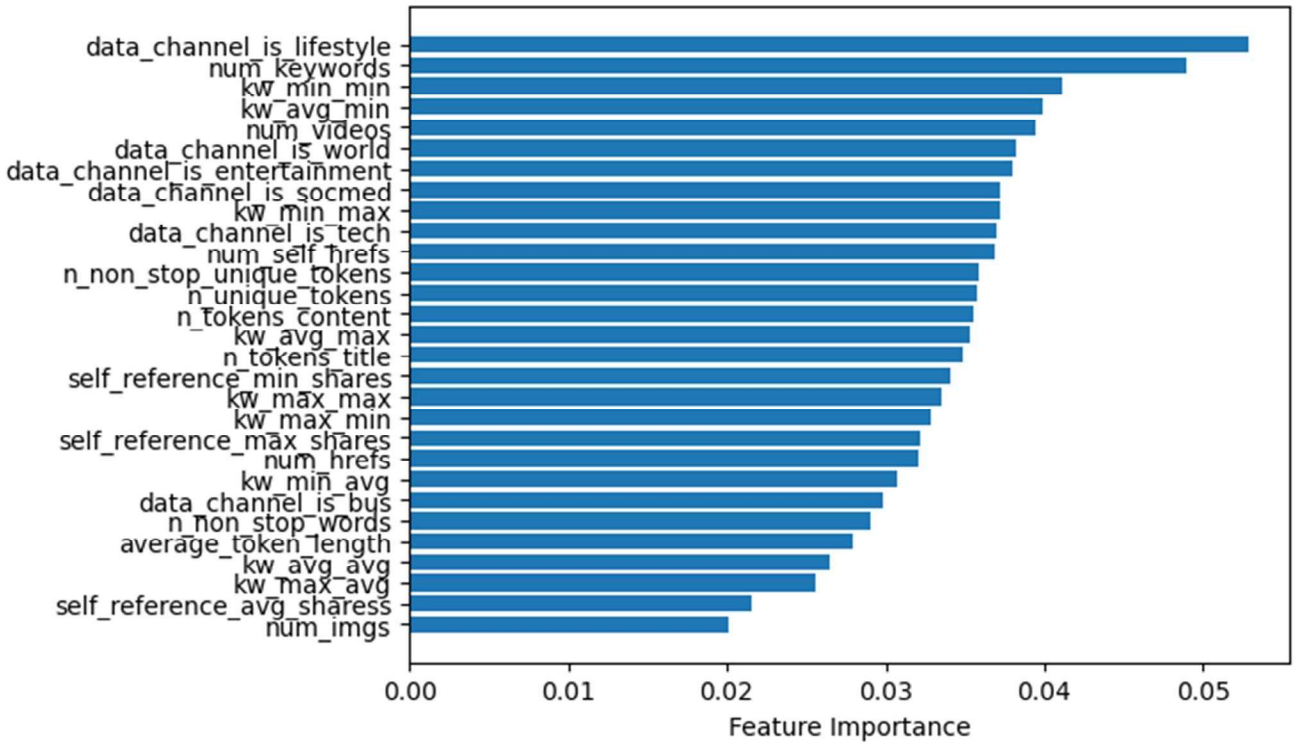
3.2 Feature Engineering:

There are both categorical and numerical features in the dataset. The value of the categorical features is either 1 or 0. The numerical features have different range of values. If not applying feature scaling to them, then the models will give higher weightage to the features with large values, resulting in a biased model. The scaling method used is MinMaxScaling. It translates each numeric features to a value between zero and one. Target variable is converted from number into Boolean based on the threshold decided in step 1 as well.

3.3 Feature Selection:

Including all features can lead to overfitting problem in model, It is necessary to remove unneeded or redundant features. URL and time delta are recognized as irrelevant in predicting the popularity in EDA and are removed in this step. Besides, Recursive Feature Elimination (RFE) is used to recognize the most relevant features as well and features with the highest importance are shown in Figure 4.

Fig 4: Feature importance



The graphs in EDA show that article category (news, particularly technology and social media are popular while entertainment and world news are less popular), published day, and number of words are related to the number of shares. While number of words (n_tokens_content) and article category (data_channel_is_lifestyle/world/entertainment/socmed/tech/bus) are on the feature importance list, all published day features are excluded. It is necessary to review and use other methods as well. There are 3 general types of feature selection: Filter Method, Wrapper Method and Embedded Method. One of each method will be run and the results will be compared in order to get the best set of features in the next stage of this project.

Step 4: Model Training

4.1 Method Selection

As I decide to replicate K. Fernandes et al. study to formulate the prediction question as a binary classification task, so I select several classification algorithms include logistic regression, Random Forests, and K-nearest neighbors. The target variable, number of shares, is numeric variable and is converted to a categorical variable in step 3.2. Feature Engineering and methodology like multiple linear regression is not used in this project. Based on K Fernandes et al. study, Random Forest generates the best result with a discrimination power of 73%. 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, both methods are selected in this project and the results will be compared to see if K Fernandes et al.'s conclusion may be challenged.

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. This project is focused on replicating K. Fernandes et al,'s study and classification approach is used in their study. Using the same approach would allow me to have a direct comparison to the original study result. Hence, classification approach is chosen in this project.

4.2 Performance evaluation

Models are built and run with full set of features in first test. Below is the results with sample size of 1%, 10%, and 80% of data. For both KNN and Random Forest methods, results and sample size are positively related. However, the result is the best with simple size at 10% for Logistic Regression. Beyond a certain point, the increase in accuracy will be small and not worth the effort and expense. Further testing of sample size for logistic regression will be conducted in the next round.

Logistic Regression

	Sample Size		
	1%	10%	80%
Accruacy - Training	0.564	0.595	0.582
F1 - Training	0.570	0.635	0.626
AUC - Training	0.565	0.591	0.576

KNN

	Sample Size		
	1%	10%	80%
Accruacy - Training	0.510	0.543	0.562
F1 - Training	0.543	0.567	0.590
AUC - Training	0.508	0.542	0.560

Random Forest

	Sample Size		
	1%	10%	80%
Accruacy - Training	0.607	0.640	0.660
F1 - Training	0.613	0.663	0.689
AUC - Training	0.608	0.637	0.657

By reviewing the sample size group with the best result of three methods , accuracy and AUC of all methods are low and need to be improved. F1 score is acceptable and has the space to be further improved.

4.3 Tuning:

Grid search method will be used in next stage of the project to determine which set model parameters give the best performance.

Step 5: Results

Refined models will be compared and a conclusion of which model and selected set of features can generate the most accurate prediction.

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.