CIND 820 Big Data Analytics Project

Predicting the Popularity of Online News

Supervisor: Tamer Abdou

Jaime Ip 024783334

June 5, 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 number of share and the features of article, 2) the best model for predicting the number of shares (i.e. popularity), 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 with cross validation (RFECV) 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

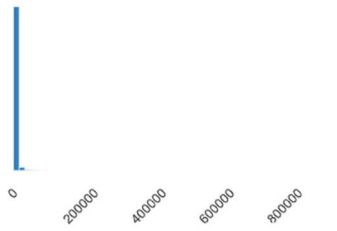
Categorical	1
Numeric	60

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	309.8 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.

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

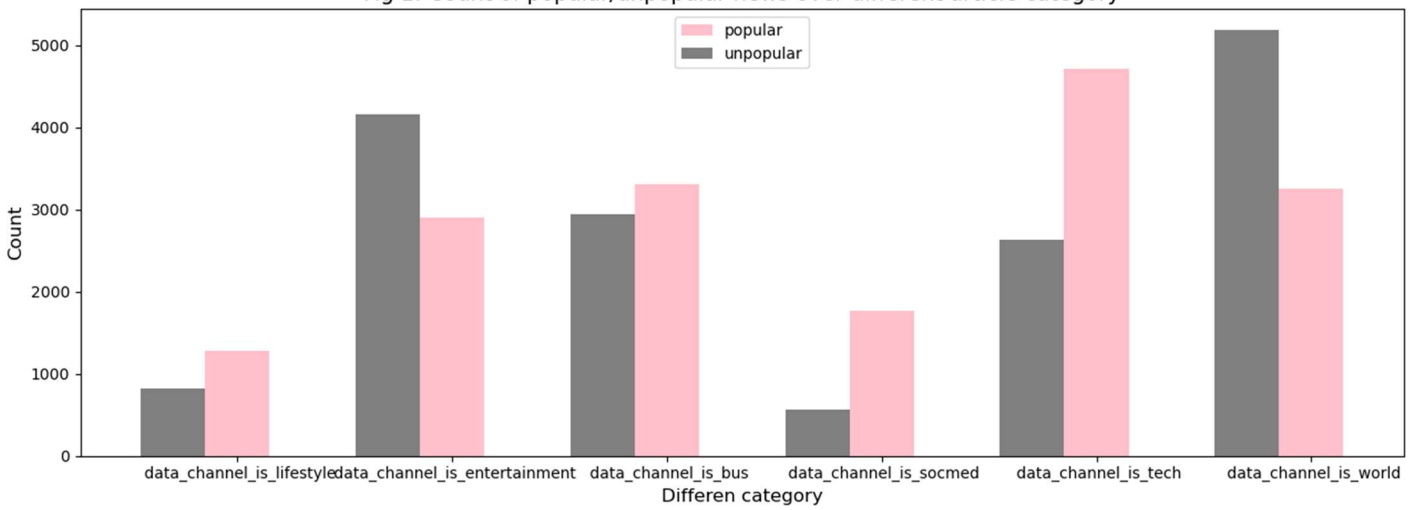


Fig 2:Count of popular/unpopular news over different day of week

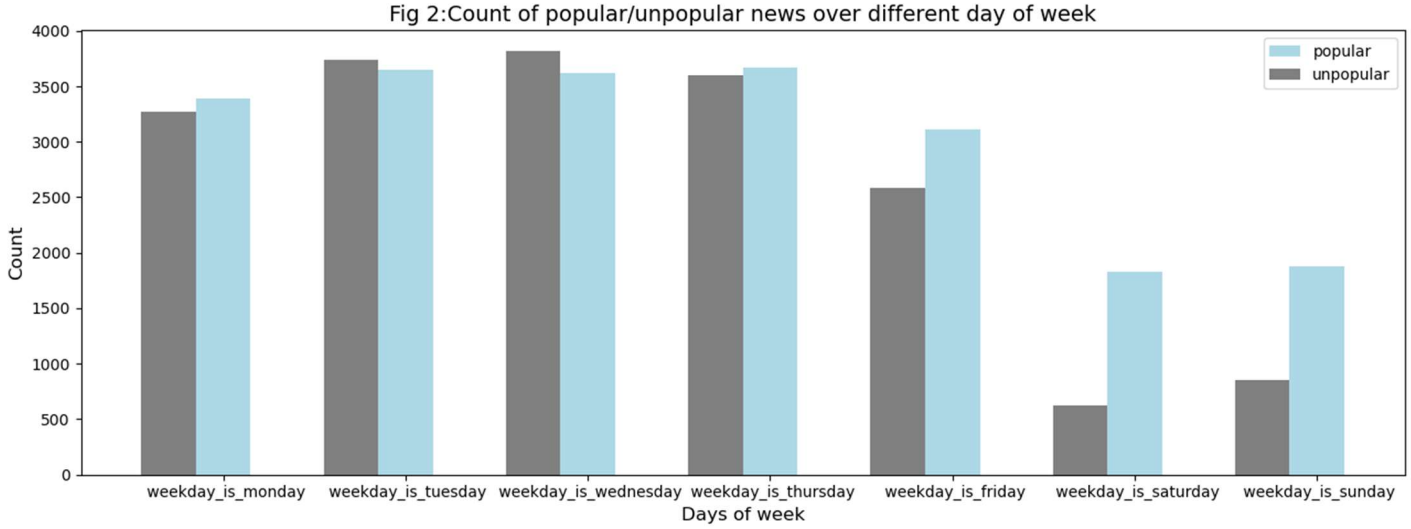
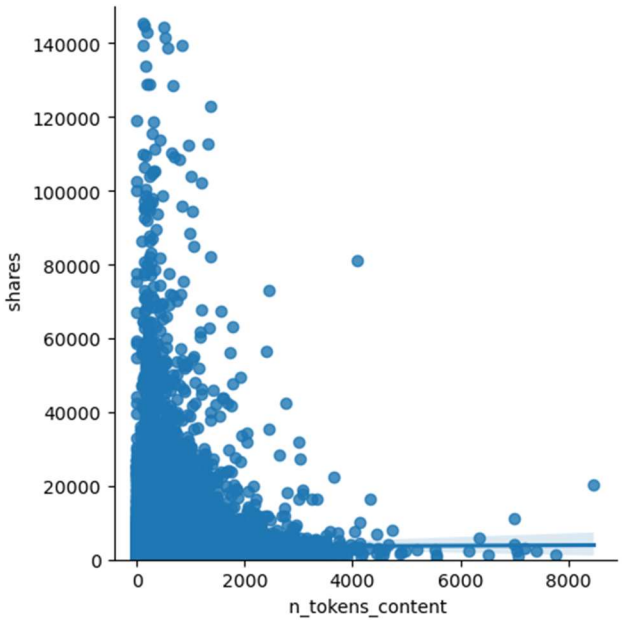
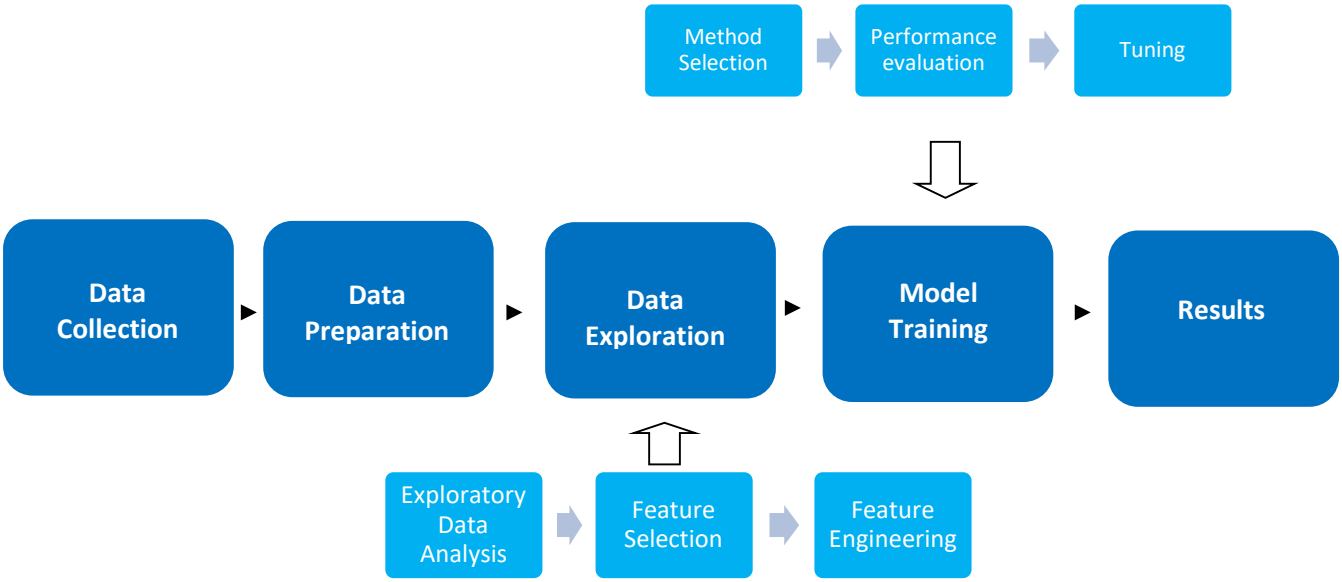


Fig 3: Scatter plot – Number of words vs Number of Shares



Approach



Step 1: Data collection

- Obtain the dataset from UCI Machine Learning Repository. The dataset is is originally acquired and pre-processed by K. Fernandes et al.

Step 2: Data Preparation

- Run Panda Profiling Report to get the data statistics overview of the dataset which includes missing value, outliners, quartile statistics, histogram for each variable etc. Identify any variable data need to be converted/transformed/corrected is necessary in this stage.
- This dataset does not have any missing value. The target variable is number of shares. I need to determine the threshold for number of shares as a popular news to convert the target variable from number into Boolean (i.e. popular or unpopular) and saved as a new feature. The median of predictable variable (i.e.,1,400) is selected as the threshold.
- All the column headings have an appended space on the left. These spaces may lead to errors when running codes in python. Removing the space can avoid unnecessary errors.

Step 3: Data Exploration

- **Exploratory Data Analysis**: analysis the relationship between variables and target variable by using data visualization. Bar charts and scatter plot reveal the relationship between article category, published day, and number of words are related to the number of shares.
- **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 already recognized as features that are not relevant in predicting the popularity. Besides, I can use recursive feature elimination with cross validation (RFECV) to recognize the most relevant features as well.
- **Feature Engineering**: I will create a new feature "Popularity" by converting the target variable shares from number to Boolean (i.e. popular and unpopular) based on the threshold of 1,400.

Step 4: Model Training

- **Method Selection**: I formulate this prediction problem as a binary classification problem (i.e. popular or unpopular) and plan to implement several classification algorithms include logistic regression, Random Forests, and K-nearest neighbors.
- **Performance evaluation**: Evaluate the performance of the classification models by reviewing the metrics score.
- **Tuning**: Use grid search method to determine which set of model parameters give the best performance.

Step 5: **Results**

- Compare the refined models and make conclusion which one 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.