Insights of Online News Popularity Data Set Data Mining Project Report

Yogesh KUMAR PILLI
MLDM - M1
yogesh.kumar@etu.univ-st-etienne.fr
Université Jean Monnet
Saint-Étienne, Loire, France

ABSTRACT

Over the past few decades Internet has drastically changed the way ideas, stories, facts are propagated across readership. We as a society have evolved technologically. Having said that, although it's easier to publish articles online with little or no cost involved, yet finding the readers for it is an altogether different task. How and why does an article get popular is the real mystery.In this study, I have intend to use a dataset to identify the criteria that make articles popular. I have also intend to do a comparative study of various Data Mining algorithms on 'Online News Popularity Dataset' and see how they fair with each other in determining whether an article will be popular or not.

KEYWORDS

Data Preprocessing,Regression,Classification,Machine Learning Algorithms

ACM Reference Format:
Yogesh KUMAR PILLI and MLDM - M1 . 2020. Insights of Online News Popularity Data Set Data Mining Project Report. In ,. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM UNDERSTANDING AND MOTIVATION

Online articles are becoming more and more prominent, as they have already crowded out more tradition news outlets, and are continuing to expand. There are very few articles that go viral, but those select few that do, get a substantially larger number of shares and views than most other articles.More importantly, the demographic for viewers of online articles is dis-proportionally skewed towards teens and young adults, a prime target for many companies to run advertisement campaigns on social media[7].As a well-known parameter, the number of shares determines the popularity of the news which in turn results in the revenue through advertisements. The key challenge in predicting the popularity of a blog post is it vastly depends on human behaviour as mentioned in the research "A survey on predicting the popularity of web content"[1].

MLDM-M1, Online New Popularity
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DATA UNDERSTANDING

The data source is from UCI Machine learning library[2]. The well-known attributes such as the number of keywords, the category of the blog post was easily understandable. The major hindrance was on the initial understanding of the analytics information provided in the data set such as LDA, polarity, subjectivity which demanded more knowledge on the text mining techniques which was briefly inferred from "Large scale sentiment analysis for news and blogs"[4]. So, the variety in the data was also a critical factor in deciding the next steps.

Following features are of critical importance when understanding the data set::

- It is a Multivariate data set as is evident from the Number of Attributes which is 61.
- Attributes are all Integer or Real barring URL and time delta which are Non-Predictive in nature.
- Total Number of Rows/Instance: 39797
- Total Number of Columns/Attributes: 61

3 DATA PREPERATION

To be able to make sense of the data, all of data sets are engineered for features.

Data Pre-Processing

Processing of the data to plays a vital role in understanding the data as well as making the model more robust.The dataset was split into training and test sets, with an 70/30 ratio. All the 58 predictors and the response variable were analyzed for the detection of missing values, outliers and collinearity between predictors.Lets Study our data as stated below parameters Does the subjects and Publishing days of news matters

From the Figure 1, You can see that all subjects looks similar regarding share numbers.

Publishing day didn't show much influence on shares neither.In order make our data clear I have get ridden all the unwanted indicators.

Figure 1: Distribution of share's with the different subject of news

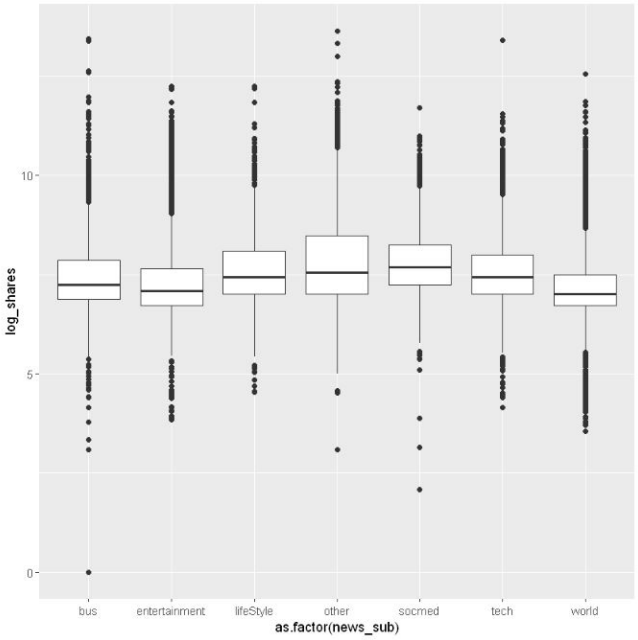
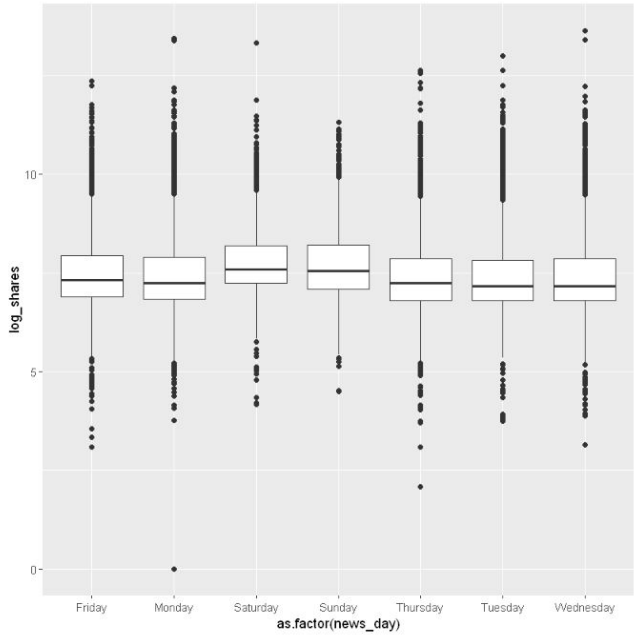


Figure 2: Distribution of share's with different days in a week



Preprocessing through PCA

A full feature set may include much noise. I had first attempted PCA for dimension reduction but it did not provide any improvements for our models,As PCA is a commonly used as dimensionality reduction algorithm[6].However, the PCA results could only made our models perform worse. This is because the original feature set is well-designed and correlated information between features is limited.

Figure 3: Principal component analysis on first two components

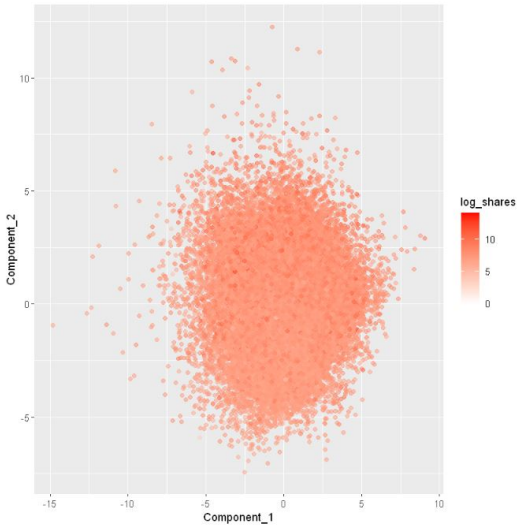
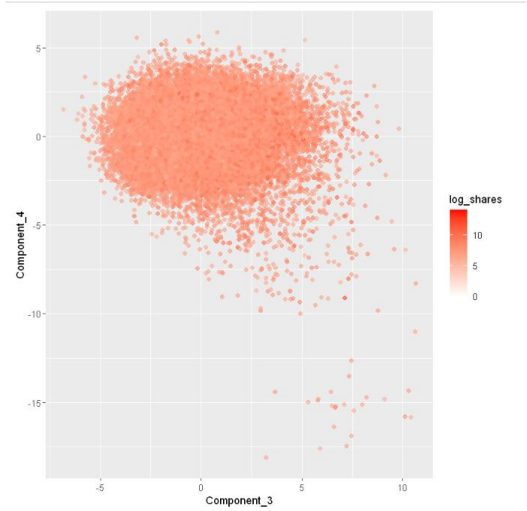


Figure 4: Principal component analysis on 3rd and 4th components



By the comparing both the figures 3 and 4 ,The variance of the share number is not aligned with the first 4 major components of the variance from independent variables.

Figure 5: Heat Map to Check the correlation matrix of the 43 left columns

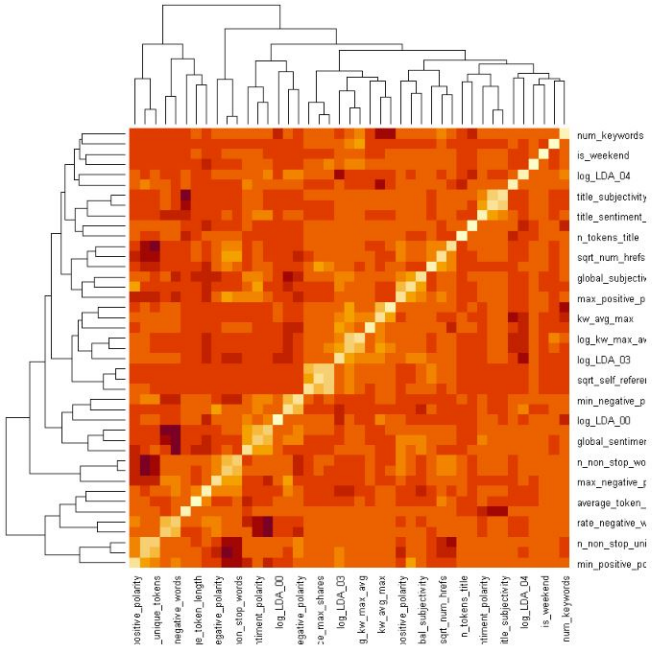
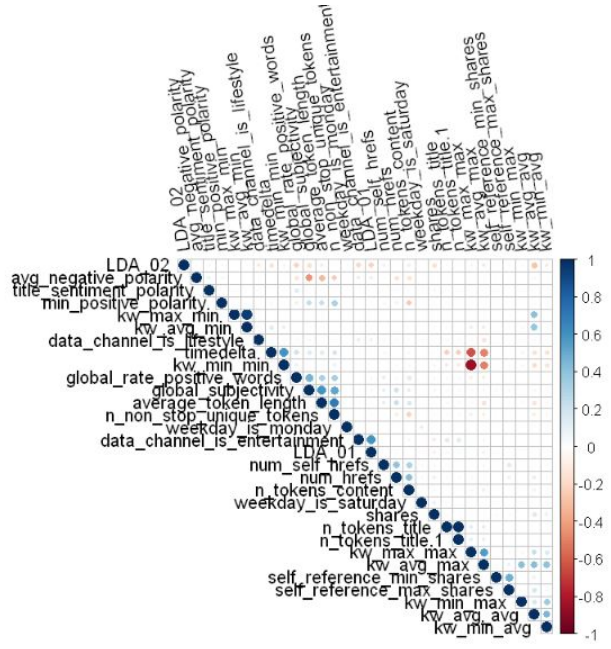


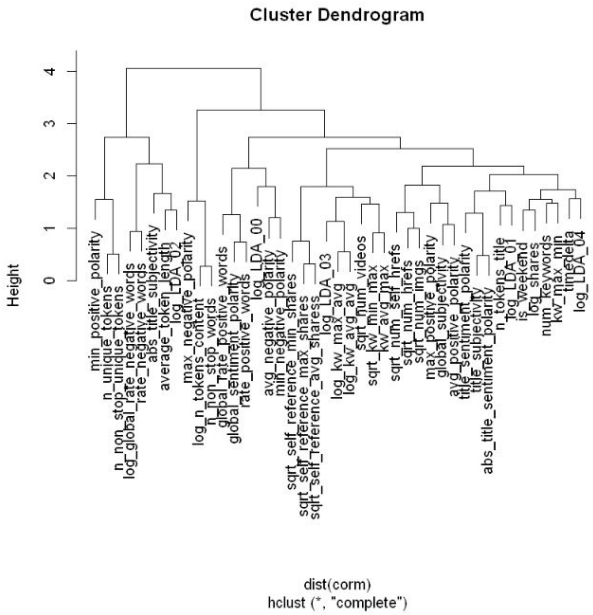
Figure 6: Correlation of different variables with our dependent variables



Lets further look at the correlations of different variables with our dependent variables

From the heapmap above in Figure 5 ,I can tell there indeed are some groups of variables which are pretty close to each other.

Figure 7: Description of the different clusters



4 MODELING

After preparing the data by removing all the outliers and irrelevant data completely, the dependent variables were identified in the previous section.We start by training different models and identifying the ones that reveal the best accuracy. Before Training the model, it's important to identify the fact that I have here dealing with a large number of attributes. Thus for effective accuracy I need to train models using algorithms that support large number of attributes. Over considerable tests I have finally shortlisted 4 different algorithms. CART, C50, KNN and Naive Bayes.

Regression

The selected attributes, as listed in the previous section are all numeric. 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. For this data is partitioned at 70-30% for training and testing. The linear model produced the mean R-Square value of 0.02213, which is also very low as compared.

Figure 8: Description of the Residuals and Standardized residual over Fitted values

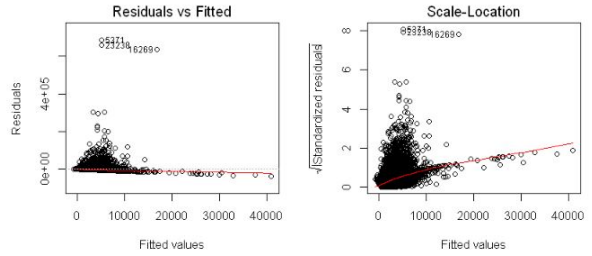
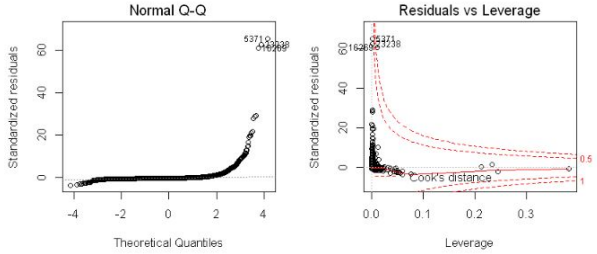


Figure 9: Description of the Theoretical Quantiles and Leverage over Standardized residual



5 MODEL EVALUATION AND DEPLOYMENT

Naive Bayes Evaluation and Deployment

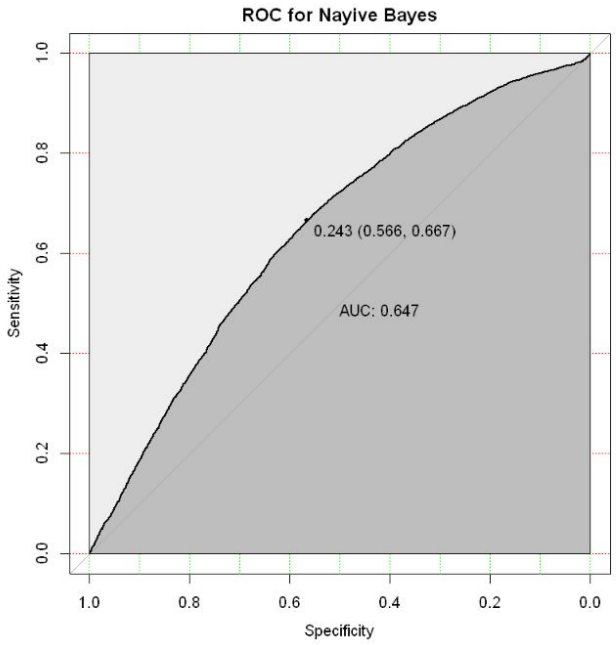
I implemented the Naive Bayes model first with a few basic features that I finalised important variables in the data set. **First approach:** I used a version of the algorithm that supports numeric attributes and assumes the values of each numerical attribute are normally distributed. This is a strong assumption, but I wanted to check the result.

Figure 10: Confusion Matrix of Naive Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4197 3714
         1 1325 2659

               Accuracy : 0.5764
                 95% CI : (0.5674, 0.5853)
    No Information Rate : 0.5358
    P-Value [Acc > NIR] : < 2.2e-16
```

I calculated summary stats(mean and standard deviation) of each numerical feature by class values.This of course, gave me a poor accuracy of **57.64%** on the validation set.Lets have a look in to the confusion matrix and the ROC details as below

Figure 11: ROC of Naive Bayes



K-nearest Neighbor-kNN Evaluation and Deployment

The KNN or k-nearest neighbor algorithm is a type of instance-based learning, where new data are classified based on stored, labeled instances. We work around with different values for k to find the one that best fits our model.Considering several experiments, I have taken k=10.Because it gives best accuracy

- I work around with different values for k to find the one that best fits our model.
- Confusion Matrix was developed to validate the accuracy for every trained Model with best k

Displaying the confusion Matrix and ROC Curve for KNN3 as below in Figures

Classification and Regression Trees- CART Evaluation and Deployment

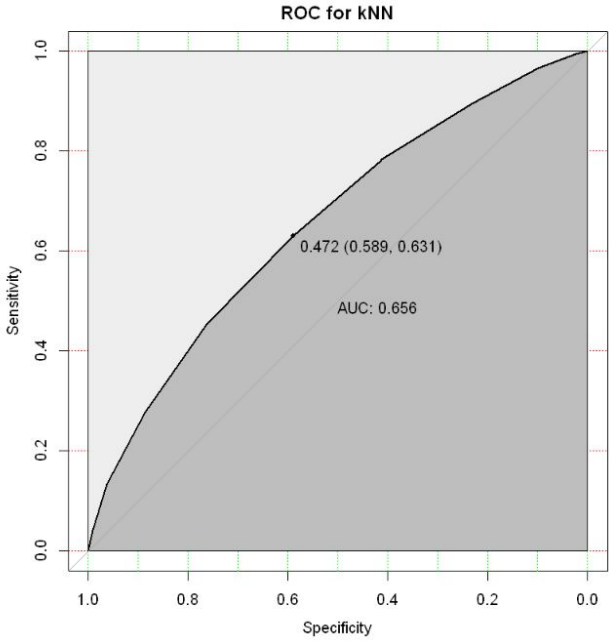
Classification and Regression Trees or CART can be used for logistic regression. The dataset we are working here with requires a better training model than the lazy learning KNN model.

- CART forms an intelligent decision tree that improves the overall accuracy of the training model.

Figure 12: Confusion Matrix of kNN

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3254 2356
         1 2268 4017

               Accuracy : 0.6113
                 95% CI : (0.6024, 0.62)
    No Information Rate : 0.5358
    P-Value [Acc > NIR] : <2e-16
```

Figure 13: ROC of kNN



- Since CART can be used to classify and predict the class label and also to identify any numerical values through Regression

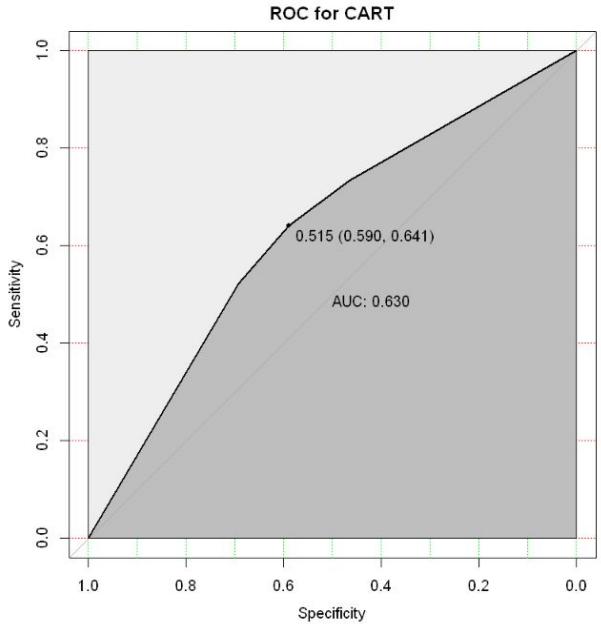
C.50-Evaluation and Deployment

So far we have worked on developing training Models using Naive Bayes,KNN and CART methods. The accuracy in either cases has been ordinary. Next we develop a training model using C5.0-Decision Trees. C5.0[3] is decision trees and rule-based models for Predictions.

Figure 14: Confusion Matrix of CART and Accuracy of Model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3258 2290
         1 2264 4083

               Accuracy : 0.6172
                 95% CI : (0.6083, 0.6259)
    No Information Rate : 0.5358
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2306
```

Figure 15: ROC of CART



- C5.0 models are rule based and they work on intelligently retaining the predictions and results as they go over the training data recursively.
- The number of trials have an adverse effect on the speed of the training model but with greater trials the accuracy has a marked change.

6 CONCLUSION

The business problem was started with the aim of predicting the reach/popularity of the news article. After multiple

Figure 16: Decision Trees

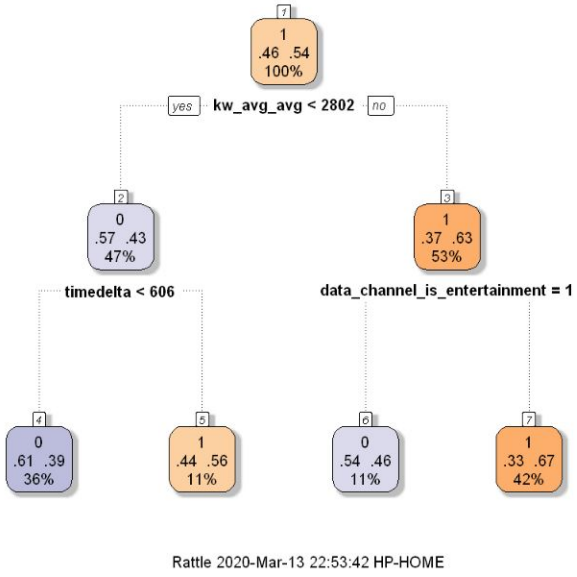


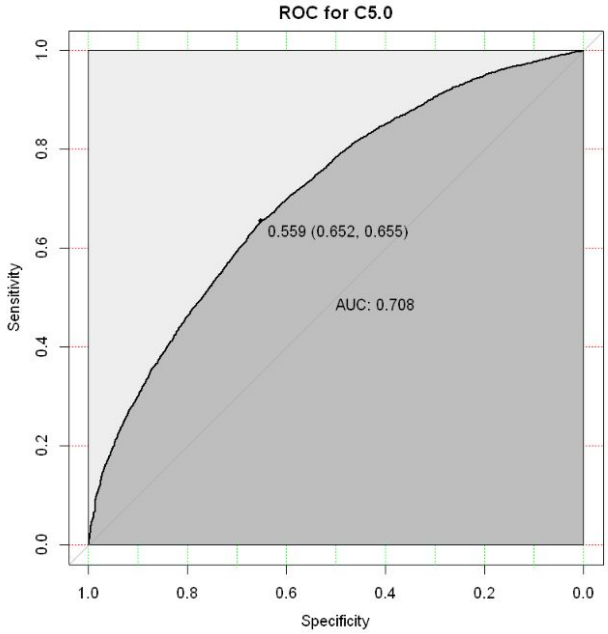
Figure 17: Confusion Matrix of C5.0 and Accuracy of Model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3211 1817
         1 2311 4556

               Accuracy : 0.653
                 95% CI : (0.6443, 0.6615)
    No Information Rate : 0.5358
    P-Value [Acc > NIR] : < 2.2e-16
```

levels of cleaning and pre-processing the data is stabilized for model building. Since the linear model couldn't produce better results because of the variance in the data, various number of bins are used, and classification algorithms are applied. Upon analyzing various models, the suited random forest model is fine tuned.

- A comparative study by training different classification models on the dataset and using them to predict output class on testing data has revealed that C5.0 is the best model among all considered models for an accurate prediction of the popularity of an online news article.
- Though 2 categories classification provides better and relevant results, it assumes popularity as a definitive output rather than a ranking methodology. In the next steps, instead of categorizing the news article, a ranking mechanism can be built using the bag of words and other text mining, clustering methodologies[5].
- The ranking methodology can be improved over the period using the reinforcement learning by adding the words from the popular articles to the bag of words normal distribution. In the weighted regression, a more complex fit for the variance could be tried in future work.

Figure 18: ROC of C5.0



REFERENCES

[1] Serge Fdida Alexandru Tatar, Marcelo Dias de Amorim and Panayotis Antoniadis. 2017. A survey on predicting the popularity of web content. *Springer* 50, 1 (Jan. 2017). https://doi.org/10.1145/1188913.1188915

[2] Paulo Cortez Kelwin Fernandes, Pedro Vinagre and LPedro Sernadela. 2015. *Online News Popularity Data Set.* Retrieved March 2, 2015 from https://archive.ics.uci.edu/ml/datasets/online+news+popularity

[3] Max Kuhn <mxkuhn@gmail.com>. 2020. *C5.0 Decision Trees and Rule-Based Models.* https://topepo.github.io/C5.0/

[4] Manjunath Srinivasaiah Namrata Godbole and Steven Skiena. 2017. Large-Scale Sentiment Analysis for News and Blogs. *JACM* 54, 2, Article 5 (April 2017), 50 pages. https://doi.org/10.1145/1219092.1219093

[5] He Ren and Quan Yang. 2017. Predicting and Evaluating the Popularity of Online News. Retrieved 2017 from http://cs229.stanford.edu/proj2015/328_report.pdf

[6] Li Sun. 2015. A more perfect union. Article. Retrieved March 21, 2015 from https://rstudio-pubs-static.s3.amazonaws.com/22671_778c16d46da6489c9f88cd7c12b20ed3.html

[7] Sheng Yu and Subhash Kak. 2018. A Survey of Prediction Using Social Media, 2005. *Springer* 3, 1, Article 4 (Jan.-March 2018). https://doi.org/10.1145/1057270.1057278