**DS502 Statistical Method for Data Science**

**Project Report**

**Online News Popularity Analysis**

**Group 6**

*Hiromi Bosman, Xiaoqun Wang, Zhaoning Su, Zili Ma, Congyang Wang*

# Abstract

With the rapid development of web in recent years, the prediction of online news popularity becomes a trendy research topic. In this paper, we attempt to find a way to analyze whether a piece of online news will be popular prior to their publication.

The data set we use to analyze is the “News Popularity” dataset found on UCI Machine Learning Repository, which has 61 features and about 40,000 observations, gathered from Mashable which is a digital media website founded by Pete Cashmore in 2005, We first standardize the data without few outliers, and do some basic statistical works to describe the data. Then we use five different classification methods: KNN, Naive Bayes, Random Forest, Logistic Regression and CART to classify these news in popular and unpopular one. After comparing the accuracy and ROC curve we find the Random Forest do the best prediction.

Next we try to improve the model, we use the PCA to reduce the feature dimensions and Boruta algorithm to select the best features. Then model is built based on original data features and new selected features.From the AUC result we find that Boruta gives us at least the same performance as the original ones.

Now we get a believable model to predict whether a piece of online news will be popular. In the future, we can also plan to track news over time, then we can use more sophisticated methods such as time series and trends analysis to improve the prediction accuracy.

# 2. Project Introduction

## ***2.1 Motivati***on and Objectives

Like the famous opening paragraph of Charles Dickens’ novel, A Tale of Two Cities said “It was the best of times, it was the worst of times.” Today is an information boom time, we met tons of news, photos, videos everywhere every second. For a media, how to write a meaningful and popular articles is a key thing to pursue, for a reader, to find a place where can read a meaningful article and willing to share to friends is a key thing to spend the time.

And, we found this dataset, *‘Online News Popularity Dataset’* from famous UCI Machine Learning Repository. In this project, we will use the statistical method and process workflow to deal with, test and analysis the data to :

1. **Data Preprocessing:** Original dataset may includes noise or redundant information, in order to improve our accuracy and efficiency, the first thing we need to do is feature selection. We will use PCA or Fisher Criterion to reduce the dimensions of data set and select the features which contribute to the model the most.
2. **Modeling:** For regression problem, we can use Linear Regression, Ridge Regression and Lasso Regression after doing some suitable transformation (Logarithmic) and interaction; For classification problem, we can use Logistic Regression, SVM, K-NN, LDA, Decision Tree, Random Forest. After modeling, we then carry on model evaluation to compare the performance of the models we implemented.
3. **Model Evaluation:** We use the adjusted R2, P-value and Residual Standard Deviation to evaluate our regression prediction, and apply K-fold cross validation to evaluate our classification prediction, and ROC curve for sensitivity/specificity analysis.

As a part of this world, we hope we can try to use what we learned from the class to dig in and find out what kind to topics the people love to watch and share, to find out if we can predict the number of shares of an article in social network, to find out if we can have a chance to increase the popularity of a digital media.

# 3. Dataset/feature Description

## ***3.1 About the Dataset***

\* The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.

\* Acquisition date: January 8, 2015

\* The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method. See their article for more details on how the relative performance values were set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 39797 | **Area:** | Business |
| **Attribute Characteristics:** | Integer, Real | **Number of Attributes:** | 61 | **Date Donated** | 2015-05-31 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | N/A | **Number of Web Hits:** | 70473 |

*Source:* [*https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity*](https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity)

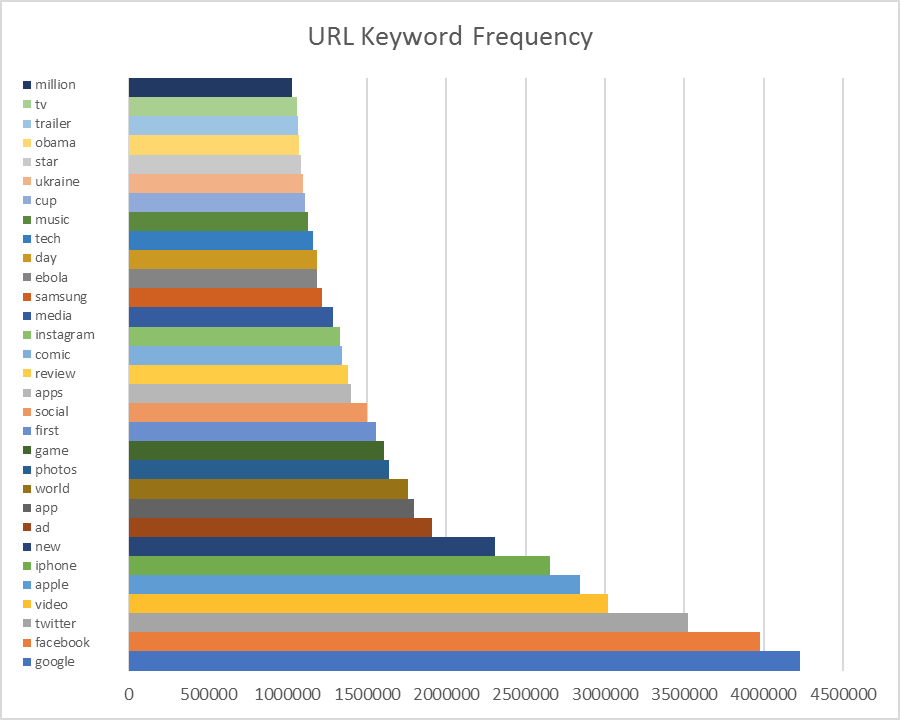
\* Attribute Information:

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

|  |  |
| --- | --- |
| **0. url:** URL of the article (non-predictive) | **31. weekday\_is\_monday:** Was the article published on a Monday? |
| **1. timedelta:** Days between the article publication and the dataset acquisition (non-predictive) | **32. weekday\_is\_tuesday:** Was the article published on a Tuesday? |
| **2. n\_tokens\_title:** Number of words in the title | **33. weekday\_is\_wednesday:** Was the article published on a Wednesday? |
| **3. n\_tokens\_content:** Number of words in the content | **34. weekday\_is\_thursday:** Was the article published on a Thursday? |
| **4. n\_unique\_tokens:** Rate of unique words in the content | **35. weekday\_is\_friday:** Was the article published on a Friday? |
| **5. n\_non\_stop\_words:** Rate of non-stop words in the content | **36. weekday\_is\_saturday:** Was the article published on a Saturday? |
| **6. n\_non\_stop\_unique\_tokens:** Rate of unique non-stop words in the content | **37. weekday\_is\_sunday:** Was the article published on a Sunday? |
| **7. num\_hrefs:** Number of links | **38. is\_weekend:** Was the article published on the weekend? |
| **8. num\_self\_hrefs:** Number of links to other articles published by Mashable | **39. LDA\_00:** Closeness to LDA topic 0 |
| **9. num\_imgs:** Number of images | **40. LDA\_01:** Closeness to LDA topic 1 |
| **…...** | **…...** |
| **18. data\_channel\_is\_world:** Is data channel 'World'? | **49. rate\_negative\_words:** Rate of negative words among non-neutral tokens |
| **19. kw\_min\_min:** Worst keyword (min. shares) | **50. avg\_positive\_polarity:** Avg. polarity of positive words |
| **20. kw\_max\_min:** Worst keyword (max. shares) | **51. min\_positive\_polarity:** Min. polarity of positive words |
| **21. kw\_avg\_min:** Worst keyword (avg. shares) | **52. max\_positive\_polarity:** Max. polarity of positive words |
| **22. kw\_min\_max:** Best keyword (min. shares) | **53. avg\_negative\_polarity:** Avg. polarity of negative words |
| **23. kw\_max\_max:** Best keyword (max. shares) | **54. min\_negative\_polarity:** Min. polarity of negative words |
| **24. kw\_avg\_max:** Best keyword (avg. shares) | **55. max\_negative\_polarity:** Max. polarity of negative words |
| **25. kw\_min\_avg:** Avg. keyword (min. shares) | **56. title\_subjectivity:** Title subjectivity |
| **26. kw\_max\_avg:** Avg. keyword (max. shares) | **57. title\_sentiment\_polarity:** Title polarity |
| **27. kw\_avg\_avg:** Avg. keyword (avg. shares) | **58. abs\_title\_subjectivity:** Absolute subjectivity level |
| **28. self\_reference\_min\_shares:** Min. shares of referenced articles in Mashable | **59. abs\_title\_sentiment\_polarity:** Absolute polarity level |
| **29. self\_reference\_max\_shares:** Max. shares of referenced articles in Mashable | **60. shares:** Number of shares (target) |
| **30. self\_reference\_avg\_sharess:** Avg. shares of referenced articles in Mashable |  |

## ***3.2*** One ***Result***

What topics are people most interested?



# 4. Methods Description(Regression / Classification)

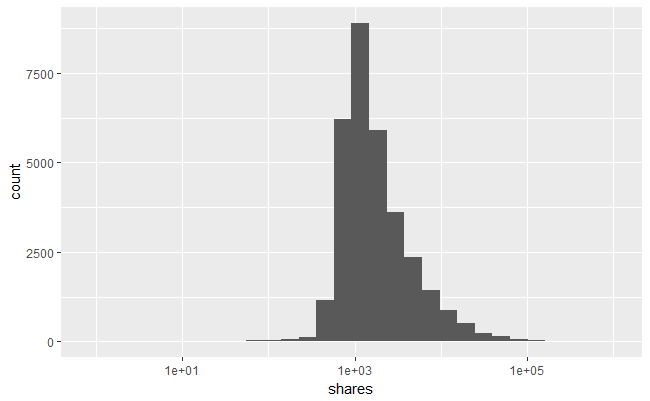
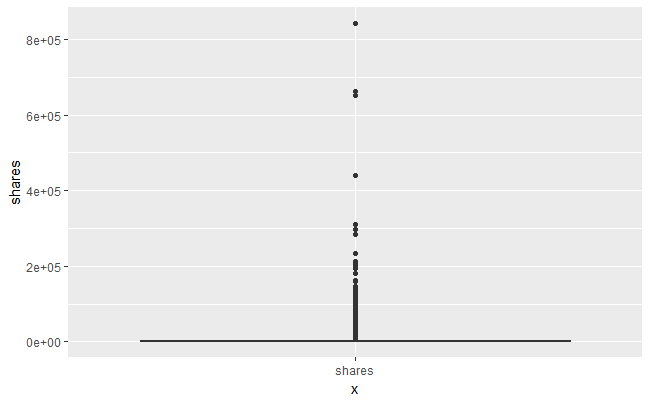
## 4.1 Data analysis

Before the conduction of data pre-processing, we used boxplot and histogram to explore the whole training data set. In the following part, we can the major problem we have here is the existence of outliers.

### 4.1.1 Target analysis

The distribution of the target value (Shares) in our news popularity data set is very uneven, many outliers can be found. As shown in the following figures, the maximum share number far beyond the median by several orders. Therefore we used the median value as the criteria to dividing results into 2 classes popular articles and non-popular articles to avoid predicting the share numbers.

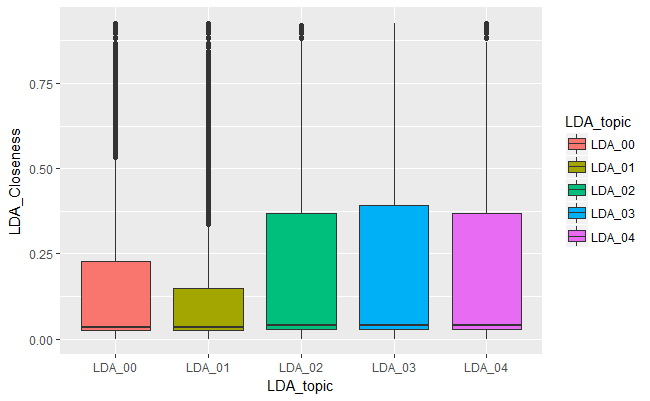
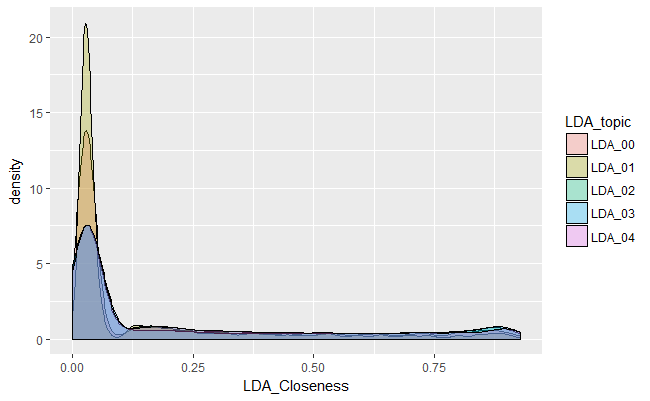




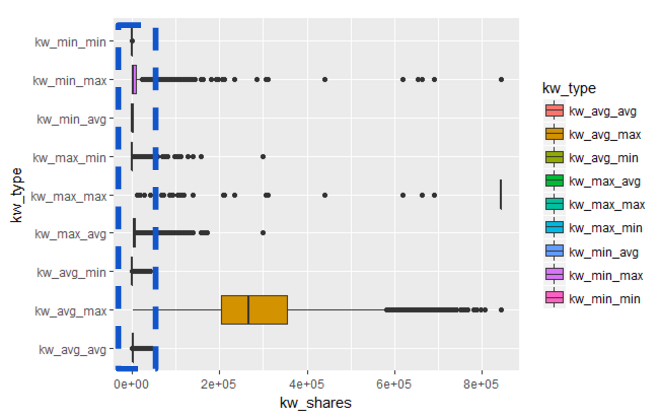
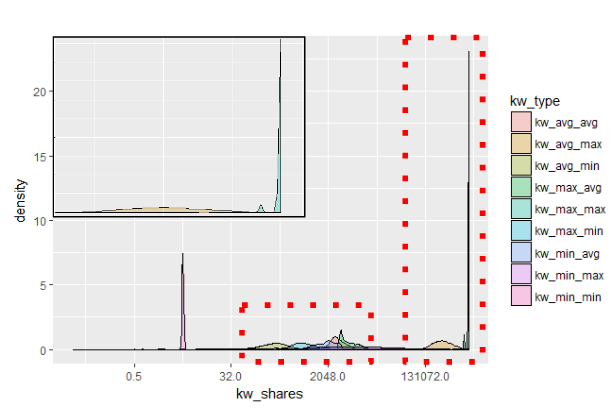
### 4.1.2 Feature analysis

The outlier problems are serious in the original data set.

The 1st example comes from the LDA-topic-related features which are used to describe how close the posted article’s topic is to a specific LDA topic. From both the boxplot and density curve plot, the outlier observations are very common in LDA\_00 and LDA\_01. We cannot deal with these observation by simple removement.



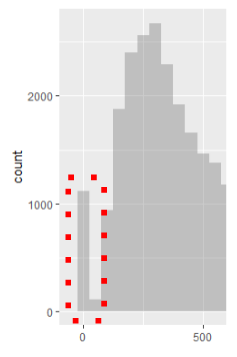
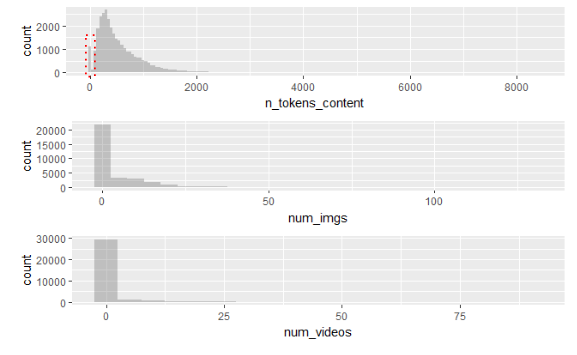
The 2nd example of outliers comes from the keyword-related features.These features are used to describe the maximum, minimum and average shares numbers of best, worst and average keywords. From the following figures, we can the main distribution peaks of these feature are usually located in the extreme value part (the extreme distribution patterns are more obvious in the boxplot.) Still, we can see many outliers existing in this features.



### 4.1.3 Unreasonable values

We also observed some feature which doesn’t make sense in the prediction process. The following figure shows three features’ distribution.“N\_token\_content” is used to take account for the number of words in the posted news. “num\_imgst” is used to take account for the number of images in the posted news.“num\_videos” is used to take account for the number of videos in the posted news.

It is reasonable for a online article without any image and video. However, it is very surprised that there are more than 1000 pieces of news without a single word. (The right figure is the zoomed-in version of the upper left figure)



## 4.2 Data Preprocessing

In this part, We cleaned and transformed our data in the following way:

|  |  |
| --- | --- |
| ***Conditions*** | ***Processing*** |
| Features with rare outliers | Remove the observations with the outlier value |
| Features with many outliers | Log Transformation |
| Features where most values are constant | Remove the feature |
| Feature values are unreasonable | Remove the observation |

## 

## 4.3 Results of different modeling methods for classification

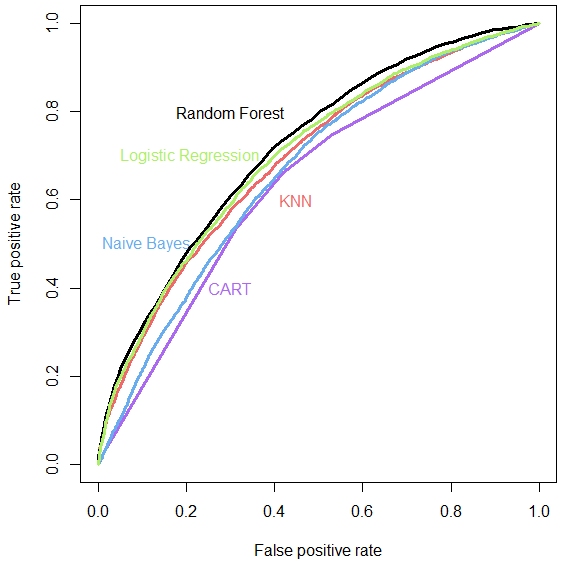
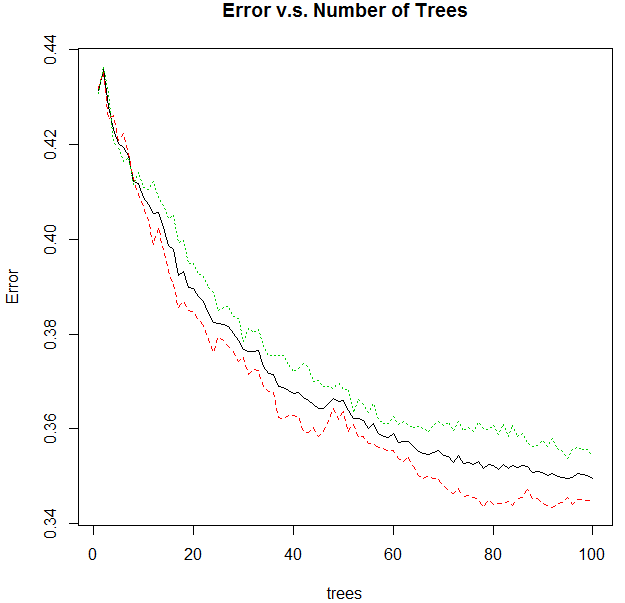
Here we carried out 5 different classification methods including KNN, Naïve Bayes, Random forest, logistic regression and CART. The results given in formats of ROC and CM are shown below. In general, all methods have accuracy that is larger than 60%. In terms of accuracy, the best result is given by random forest with the highest accuracy equaling to 66.2%. CART and Naïve Bayes gave us the same lowest accuracy rate equaling to 61.92%.

|  |  |  |
| --- | --- | --- |
| **Modelling method** | **ROC** | **CM** |
| ***KNN*** | KNN_ROC_K300.png | KNN_ConfusionMatrix_K300.png |
| **Modelling method** | **ROC** | **CM** |
| ***Naive Bayes*** | NaiveBayes_ROC.png | NaiveBayes_ConfusionMatrix.png |
| ***Random Forest*** | RandomForest_ROC.png | RandomForest_ConfusionMatrix.png |
| ***Logistic***  ***Regression*** | LogicRegre_ROC.png | LogicRegre_ConfusionMatrix.png |
| ***CART*** | CART_ROC.png | CART_ConfusionMatrix.png |

We tuned the process of RF to optimize the number of trees. In the final modeling, we used 100 trees. The testing error vs number of trees is shown in the following left figure. Also we can know the ranking of features according to their importance in the RF model to analyze which attributes have a greater impact on judging whether the online news is popular. The rank of the features in our RF model is shown in this table below, from which we can see that in this case the average keyword (average shares) and the minimum shares of referenced articles in Mashable are two most important variables in this classification.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Mean Decrease Accuracy (#Rank)** | **Feature** | **Mean Decrease Accuracy (#Rank)** |
| kw\_avg\_avg | 18.65(#1) | kw\_min\_avg | 12.41(#5) |
| Self\_reference\_  min\_shares | 16.21(#2) | LDA\_00 | 12.29(#6) |
| kw\_max\_avg | 15.97(#3) | Self\_reference\_  max\_shares | 12.20(#7) |
| Self\_reference\_  avg\_sharess | 13.49(#4) | ... | ... |

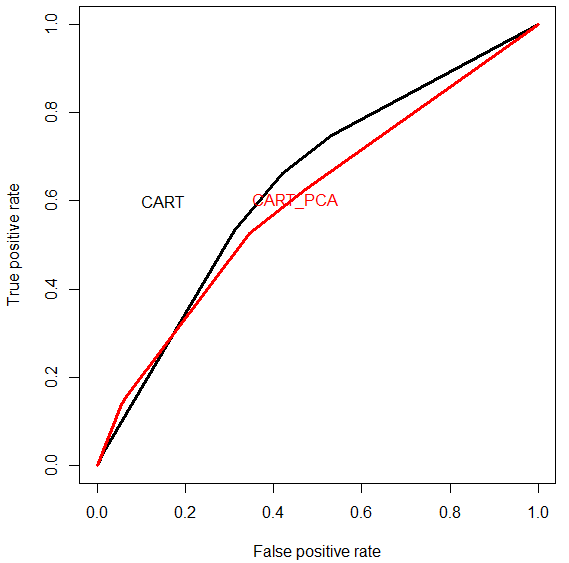
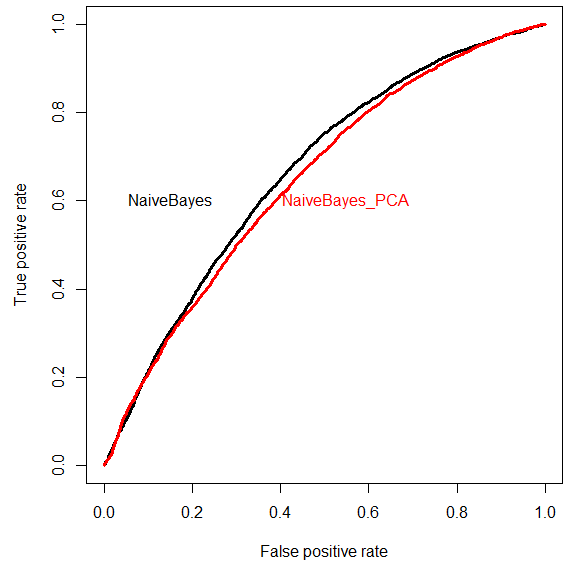
The Combination of ROCs of different model in below right figure can show our cooperation result more clearly. AUC of KNN, Naïve Bayes, Random forest, logistic regression and CART are 0.693, 0.665, 0.719, 0.703, 0.607.

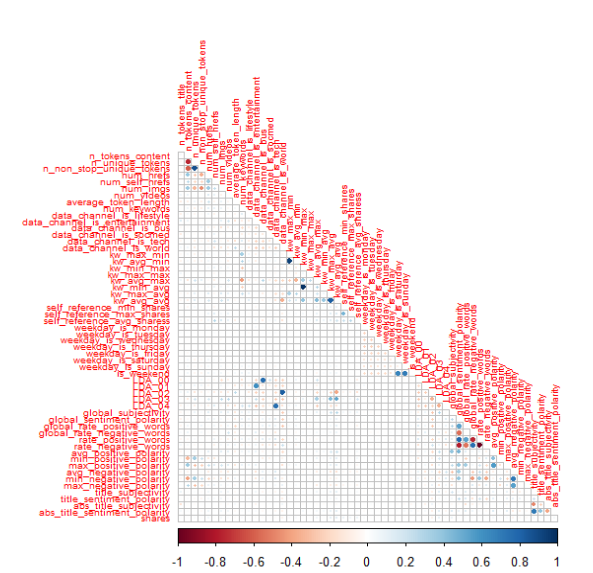


## 4.4 Dimension Reduction by PCA

### 4.4.1 Dimension Reduction by PCA

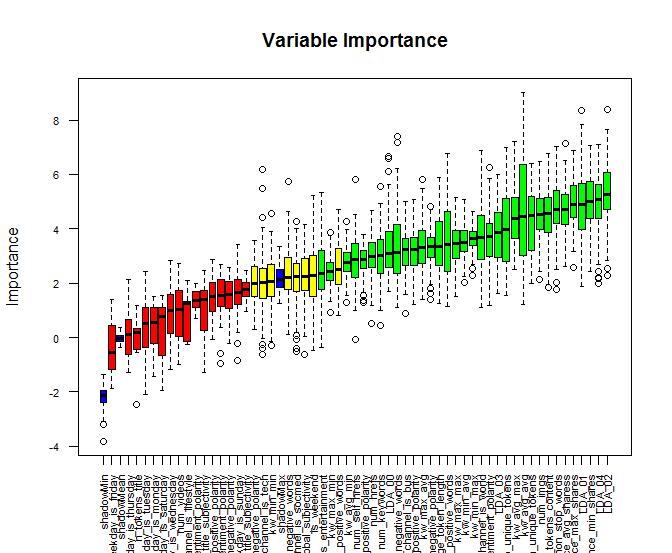
We tried to reduce the feature dimensions by PCA, the testing ROC results based on CART and Naive Bayes are shown as follows. The results indicate the PCA can reduce the feature number without losing too much accuracy, but in general, PCA cannot give us a better prediction by reducing the redundancy. The reason of the failure of PCA is the original feature set is well designed and the correlated between features is very limited. The correlation plot of the features is shown in the lower figure. We can see the appearance of highly correlated terms is rare and localized.

******



### 4.4.2 Feature selection by Boruta Library

Boruta library for R is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest. This package derive its name from a demon in Slavic mythology who dwelled in pine forests. The following figures shows the comparison results between feature-selected modeling and non-feature-selected modeling. We first use Boruta to select the best features. Then model is built based on original data features and new selected features.



In the Boruta boxplot above, it shows the feature importance by showing different colors. Blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Red, yellow and green boxplots represent Z scores of rejected, tentative and confirmed attributes respectively. Out of 59 features, Boruta confirmed 42 features and rejected 16 features. We used those 42 features to do the classification.

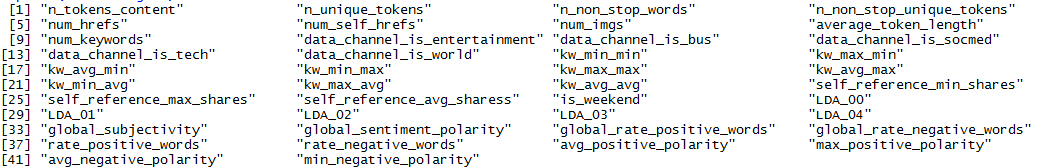
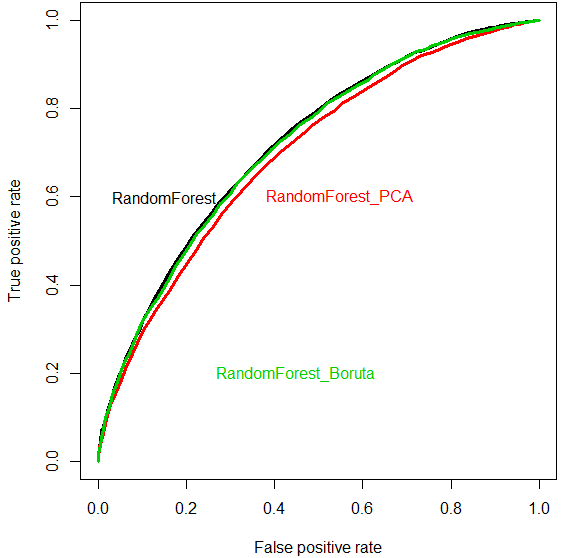
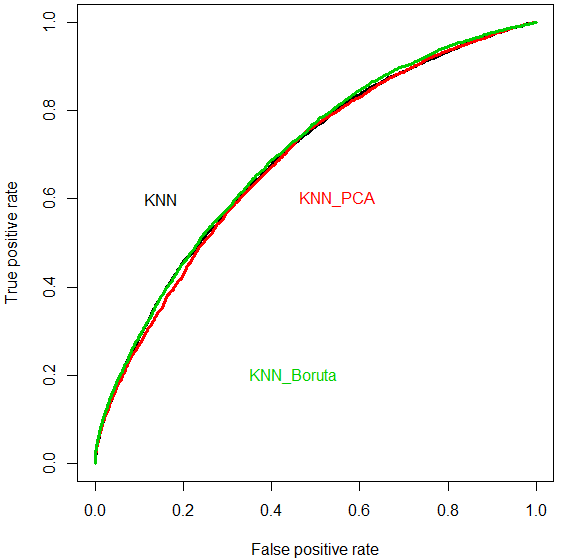
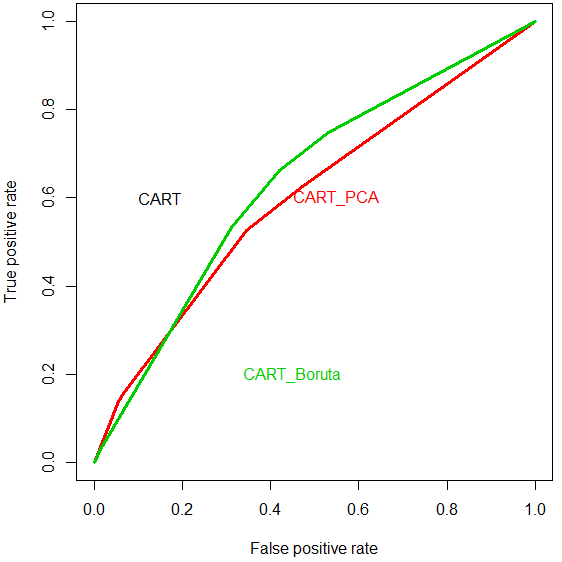


Figure: Selected features

From the table below, we can see, the model built by Boruta selection gives us at least the same performance as the original ones provided. It is not surprised that Boruta has almost exactly same AUC as the AUC given by RF model, because Boruta is based on RF method too. The AUC result is shown in the following table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature selection** | **AUC result Of different Selection methods** | | |
| **Method** | **PCA** | **Boruta** | **Original** |
| **KNN** | 0.6345 | 0.6413 | 0.6391 |
| **CART** | 0.5918 | 0.6192 | 0.6192 |
| **NB** | 0.5886 | 0.5952 | 0.5962 |
| **RF** | 0.6435 | 0.6581 | 0.6579 |
| **LR** | 0.6425 | 0.6484 | 0.6496 |

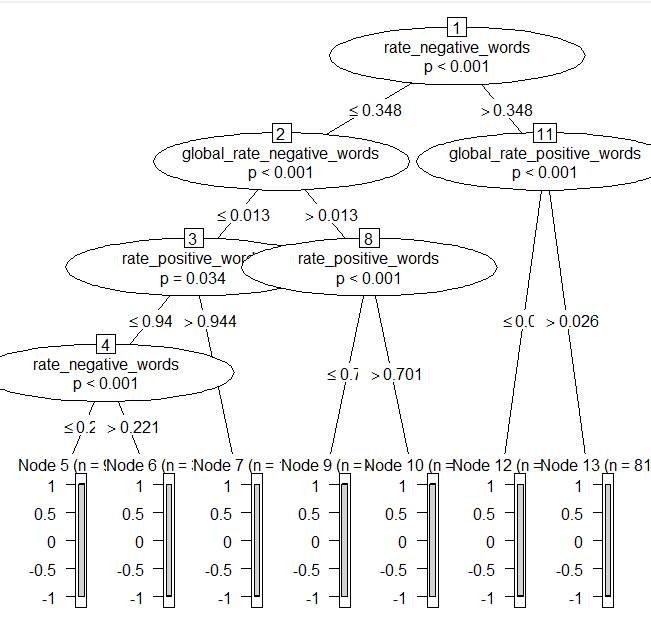
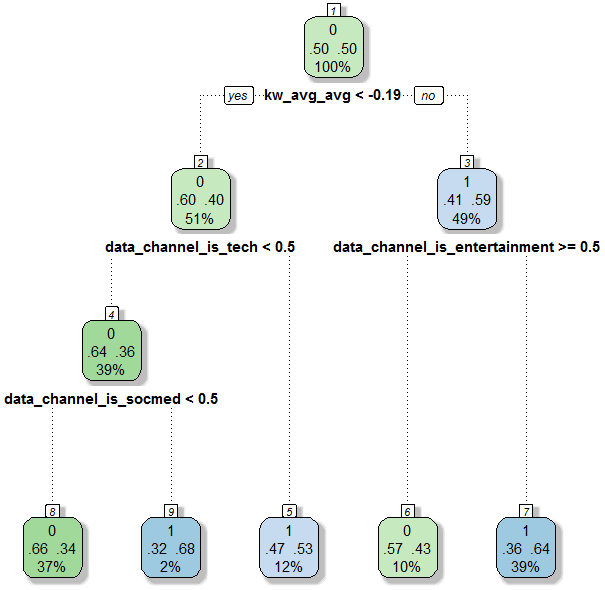


# 5. The Conclusion

The tree model based on classification tree and CART are shown as below. The key variables selected by the 2 models for splitting are different.

For the C-tree model, the key factors determining whether a piece of news is popular online selected by the model are positive and negative word-related features like “rate\_negative words” and “rate\_positive words”. While the key factors selected by CART trees are quite different. In this model, Average keyword’s average share number and the channel type of the news are considered as more important.

These various and low correlated tree model may explain why both of them cannot give us a good accuracy. However, the best option of the data set with the large variance in tree models is just RF model. Our result also agree with our idea that RF has the best performance on our data set.



# 6. Future Work

According to our statistical learning, we can predict whether a piece of online news will be popular with features we have analysed that are known prior to an article publication. Here we try a lot of methods to predict the classification, and the best result was achieved by a Random Forest (RF), with an overall area under the Receiver Operating Characteristic (ROC) curve of 71.9%. In order to enhance the expected popularity, we can analyze more features in the future. Otherwise we can also plan to track news over time, then we can use more sophisticated methods such as time series and trends analysis to improve the accuracy.

# 7. References

"UCI Machine Learning Repository: Online News Popularity Data Set." *UCI Machine Learning Repository: Online News Popularity Data Set*. Web. 11 Dec. 2016.