Analysis and Prediction of News Popularity

553.636 Data Mining

Omotola Akeredolu, Ruoxi Liu, Kexin Wang, Weizhuo Wang

Introduction

In the recent decades, online news became a new popular source for people to obtain news, and for most of the times, people would only focus on the hot topics. In our project, we aimed to define whether a piece of news is "hot" or not.

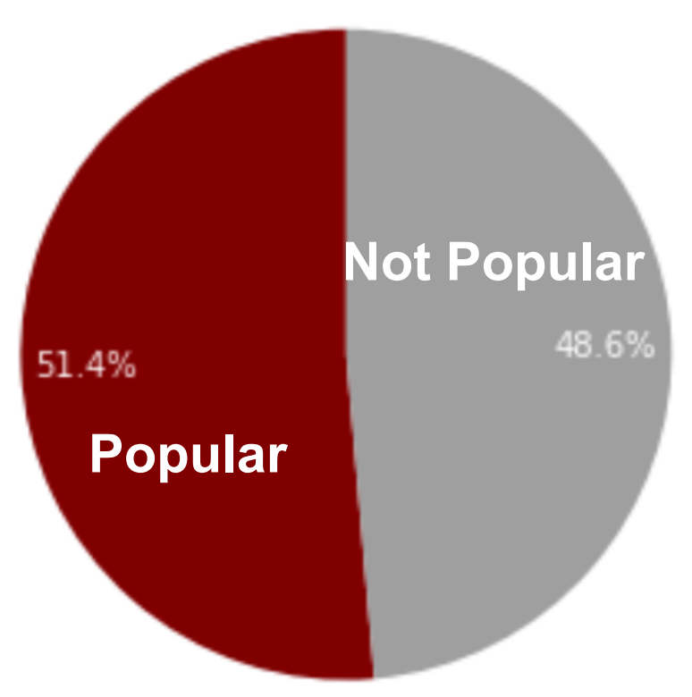
We used data mining algorithms such as KNN, Naive Bayes and Random Forest to perform classification, and compared the CV scores to determine the most appropriate model.

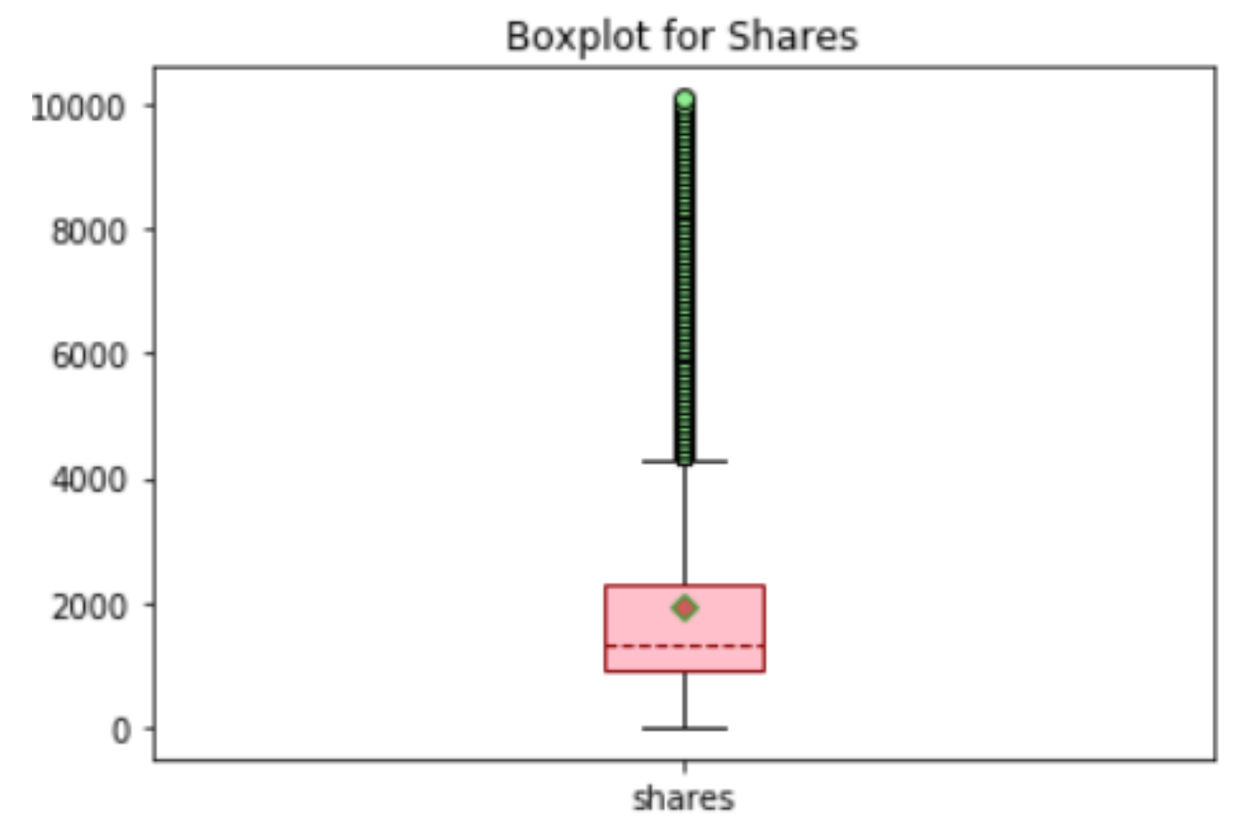
Data Exploration

- Our dataset is provided by the UCI machine learning repository.
- Total number of observations is 39797; total number of attributes is 61.
- Each attribute describes the news from a different perspective.
- Target variable ("class"), which has 2 values 0 and 1, is assigned based on the column "shares". For news with share number smaller than 1400, we regarded it as "not popular" and assigned 0 to it; Otherwise, we regarded it as "popular" and assigned 1 to it.

Distribution of Shares

Count	33510
Mean	2928
25%	930
50%	1400
75%	2500

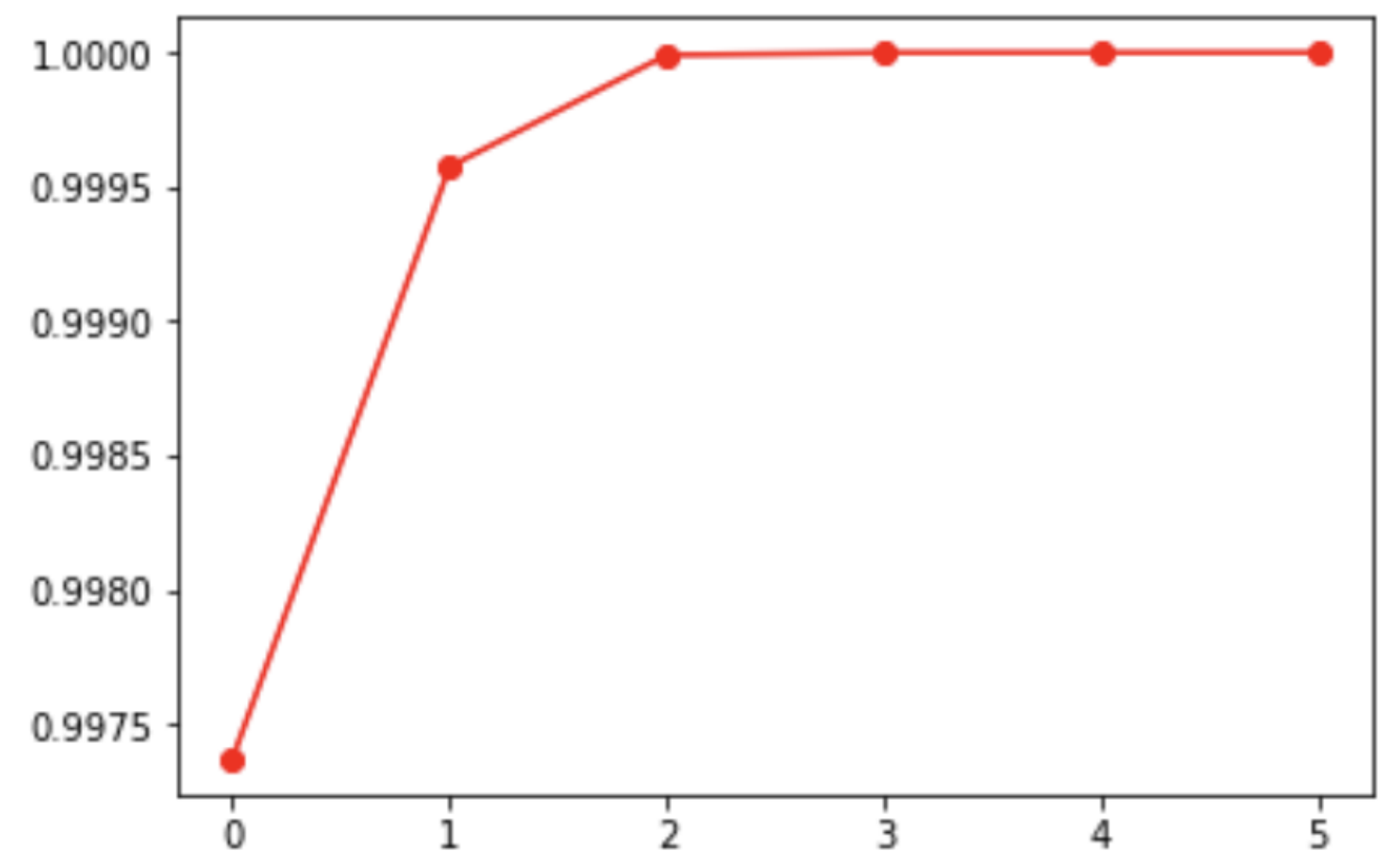




Methods and Models

We applied **KNN with 10 neighbors, Naive Bayes (flat priors), QDA, LDA, Logistic Regression and Random Forest with 300 trees**. We firstly fitted the original data and calculated the Cross-Validation scores for each method.

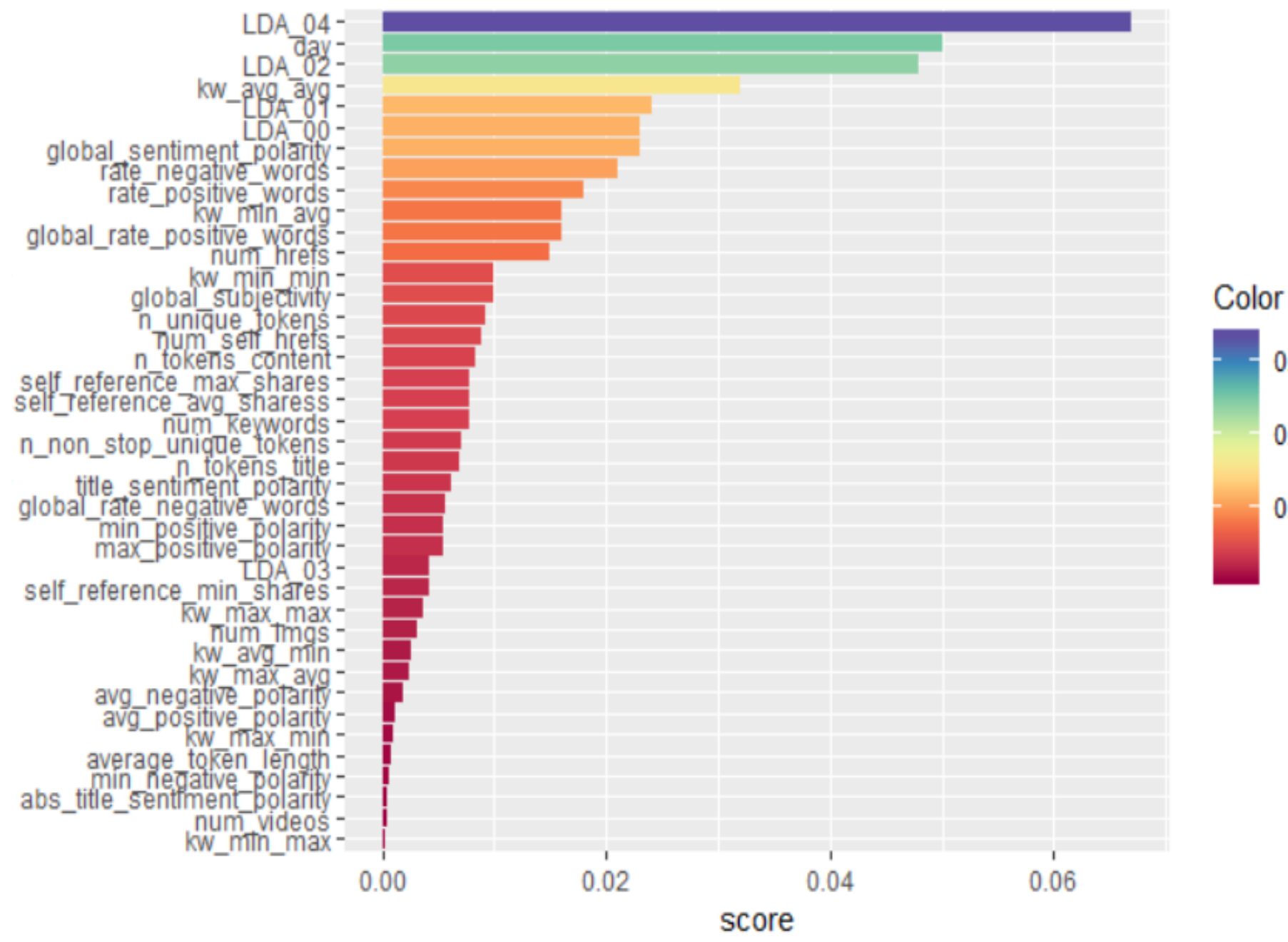
In seek of improving model accuracy further, we tried to use PCA and whitening to pre-process the data and redid classification. As suggested by the scatter plot, the transformed data after PCA kept 3 components.



Feature Selection

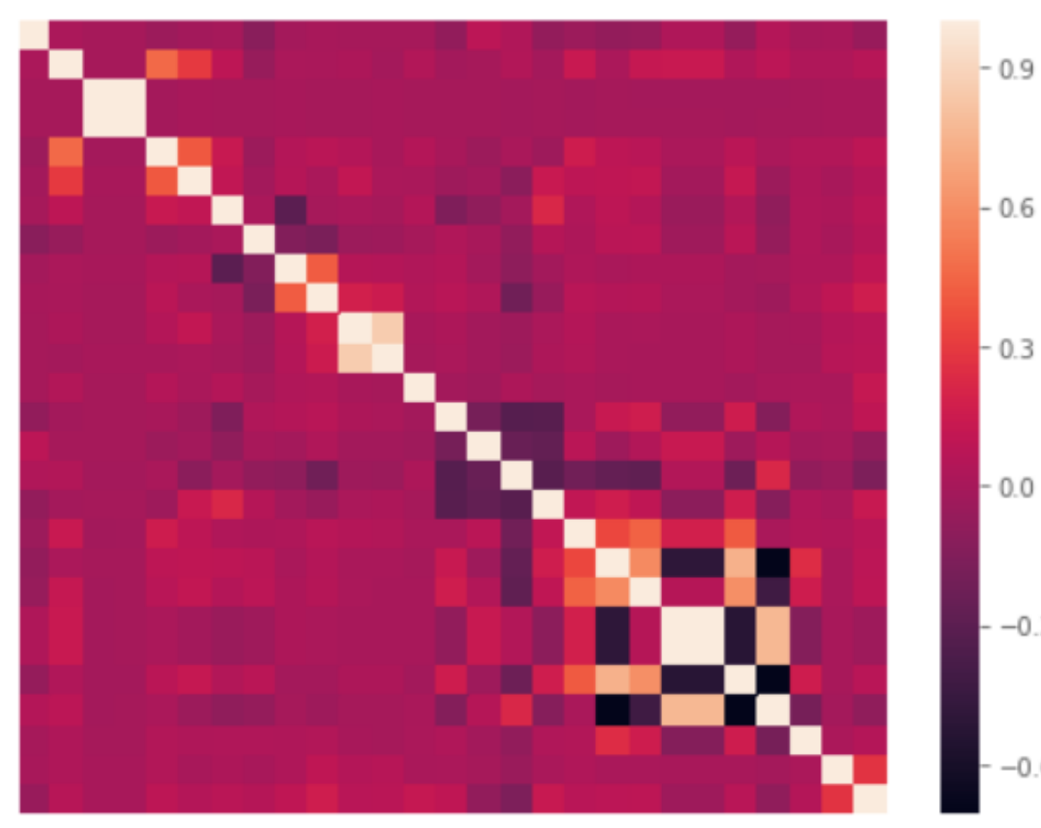
I. **Fisher Score:**

Fisher criterion is an effective way in feature ranking. A larger fisher score indicates greater discriminative power of the corresponding feature. We kept **25** features with highest Fisher Score here.

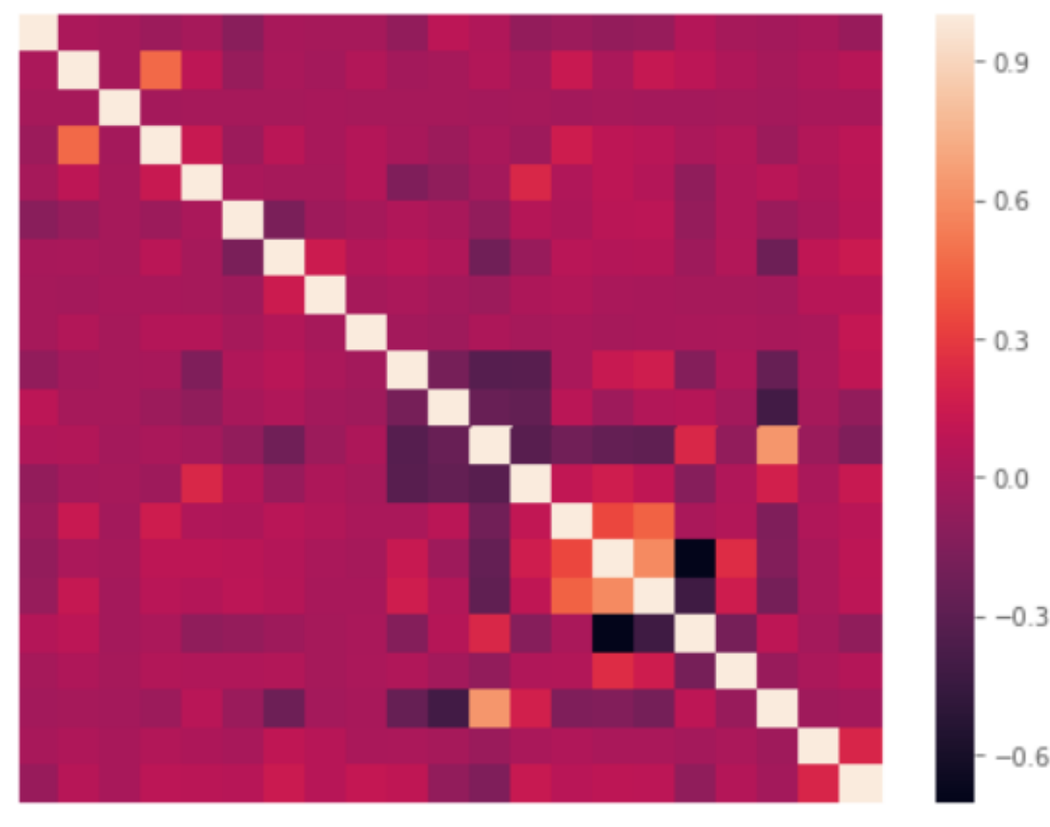


II. **Correlation Examination:**

We analyzed correlation between the predictor variables. The result plot implies some variables are strongly correlated. For example, the correlation between "Rate of unique words in the content" and "Rate of unique non-stop words in the content" **almost close to 1.** Then we deleted several features and finally kept **19** features, which gave us a better performance than the other choices.



Before Feature Selection

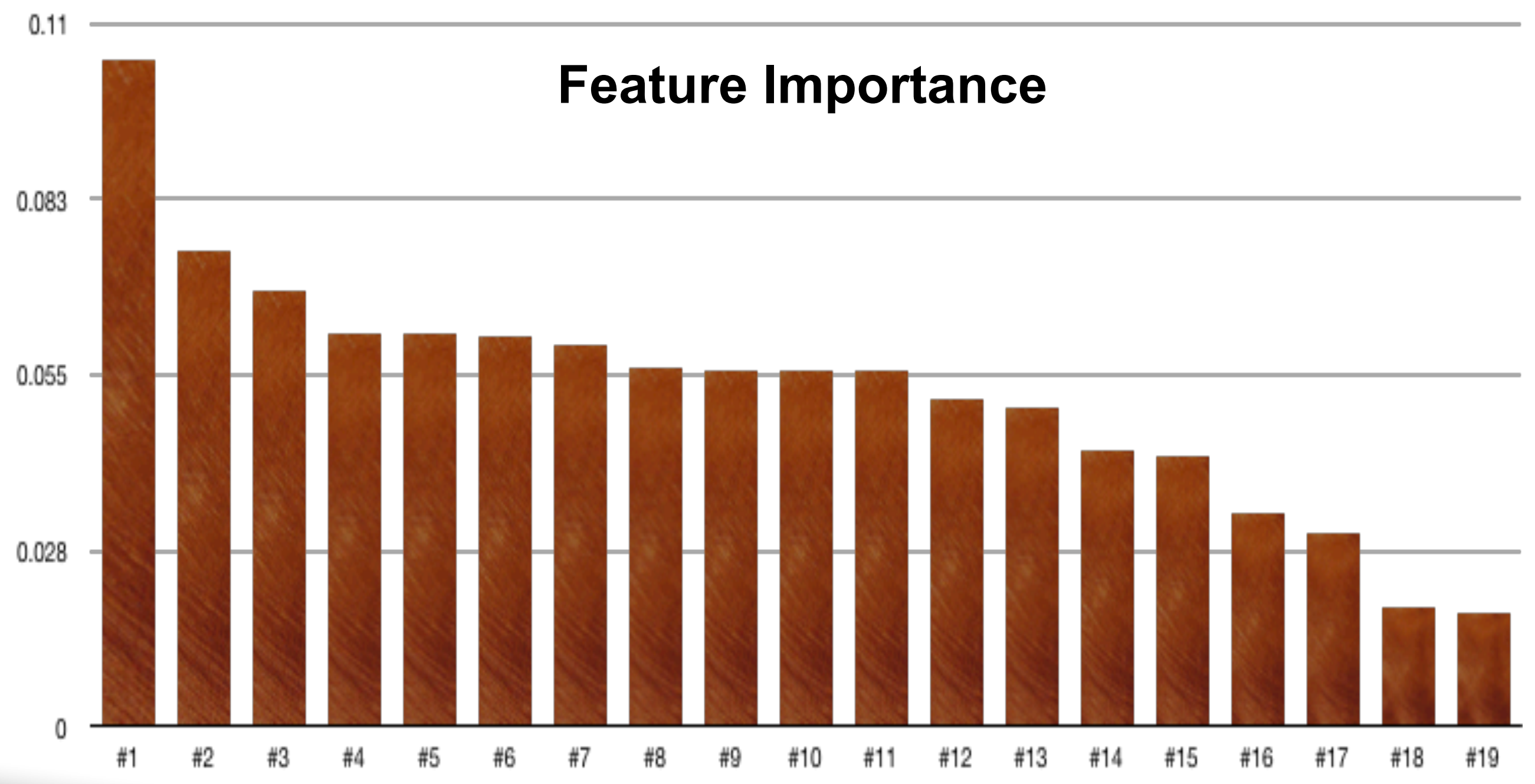


After Feature Selection

Classifier Comparison

Cross-Validation Scores		
	Without PCA	PCA
KNN (K=10)	0.54437	0.53258
Naive Bayes	0.53709	0.52841
QDA	0.54329	0.52889
LDA	0.61170	0.58768
Logistic Regression	0.60949	0.59371
Random Forest (300 Trees)	0.65364	0.60190

Feature Importance					
#1	Avg. keyword (avg. shares)	0.104221	#11	Text subjectivity	0.055595
#2	Avg. shares of referenced articles in Mashable	0.074363	#12	Day in a week when article published	0.051281
#3	Closeness to LDA topic 2	0.068016	#13	Rate of negative words among non-neutral tokens	0.050062
#4	Rate of unique words in the content	0.061305	#14	Channel	0.043187
#5	Closeness to LDA topic 4	0.061249	#15	Number of links	0.042497
#6	Closeness to LDA topic 0	0.060852	#16	Title polarity	0.033447
#7	Closeness to LDA topic 1	0.059724	#17	Number of words in the title	0.030153
#8	Number of words in the content	0.056258	#18	Number of keywords in the metadata	0.018569
#9	Rate of positive words in the content	0.055799	#19	Worst keyword (min. shares)	0.017717
#10	Text sentiment polarity	0.055703			



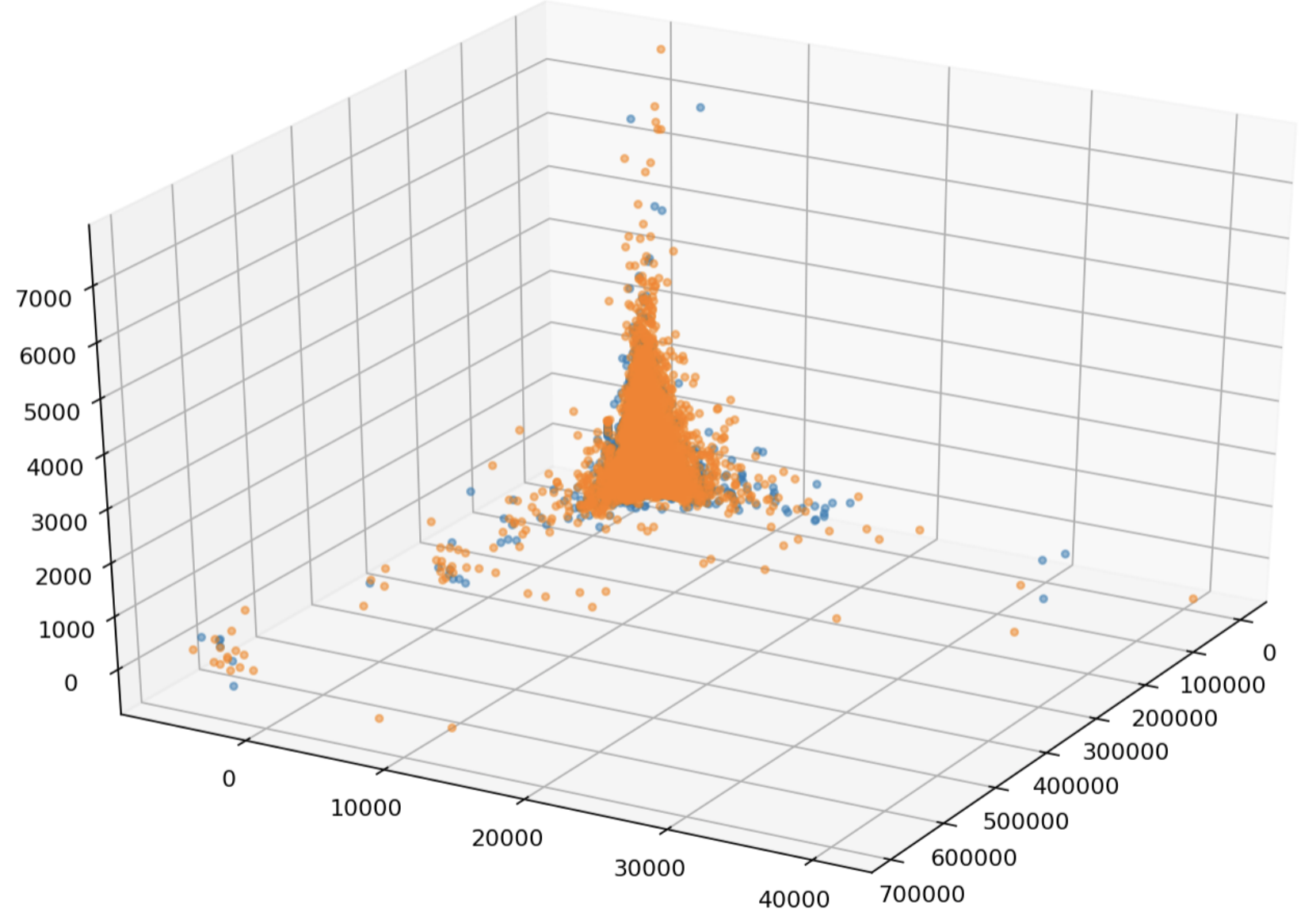
The best model is the Random Forest with 300 decision trees, maximum depth of 15 and maximum feature of 5. The accuracy is approximately 65%. Logistic Regression does a decent job as well, with the accuracy being approximately 60%. The least efficient model is Naive Bayes. The accuracy is approximately 53%.

The feature importance is then calculated from the Random Forest model. And it shows that the Average Keyword is the most important feature and the Average Shares of Referenced Article is the second most important feature. Note these two variables are related to the services that Mashable provides: articles often reference other articles published in the same service; and articles have meta-data, such as keywords.

PCA

In our study, PCA did not improve the results in any cases. Since variables in our data have low correlation, performing PCA will lead to information loss and consequently worsen the classification results. Feature importance table suggests that the principal components mainly use 5 variables. By plotting the components, we see the two classes are highly overlapped.

PCA Feature Importance			
	PC1	PC2	PC3
Avg. shares of referenced articles in Mashable	0.999973	-0.007277	0.000233
Avg. keyword (avg. shares)	0.007277	0.999904	-0.005205
Number of words in the content	-0.000197	0.005103	0.999894
Worst keyword (min. shares)	-0.000111	-0.010570	-0.009046
Number of links	0.000002	0.000765	0.010188



Conclusion and Future Work

Conclusively , this project was focused on predicting the popularity of news articles based on the number of shares of the articles. Given the high dimensionality of our data, we used fisher score to select the most relevant features. We then used several classification algorithms to classify the data. Our results show that the accuracy of the classification models is quite low with Random Forest being the highest with an accuracy of 65%. We then used PCA to process the data but it failed to improve the classification as expected.

To improve accuracy, there is limited room in model selection, but much room in feature selection. Although the original dataset contains 61 features, many of them are highly correlated and thus are removed. We believe that by adding additional features that could describe the news effectively, we will be able to improve the accuracy.

In the future, we could implement an efficient data cleaning method instead of manually removing and deleting features with missing data during the preprocessing stage. Also, due to the size of the data and limited computing power, we could not run some classification models like SVM. In the future we could use a computer with higher computation power to run these other classification models and see how they compare to our current results.