Capstone Project2: Online News Popularity

1. Background

Internet has been the major platform for people sharing and receiving news. Different from traditional newspaper and magazines, online news is immediate and myriad. It is a challenge again for journalism to have their work stands out among countless competitor. For websites and social media, it is also important for them to understand the factors affects news popularity. They can put most popular news on the front page to attract the visitor by predicting the popularity before publication. Thus, it is significant to learn the factors affects online news popularity and predict it.

2. Dataset Describe

The dataset is consisted of 39,643 news articles from an online news website called Mashable collected over 2 years from Jan. 2013 to Jan. 2015. The dataset was initially 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 andretrieved using the provided urls.

It is downloaded from UCI Machine Learning Repository as https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity.

Each record contains the URL link of the news and 60 descriptive attributes, such as number of words in the title, number of images and weekday it is posted. The dataset has been initially preprocessed.

3. Data Wrangling

3.1 Read the data and check data info using pandas

The dataset was read as pandas dataframe. There is 1 object columns(url), 59 floating data and 1 int data (share). There is no missing value. The shares data is the feature that measuring news popularity. We will use all columns except for url.

3.2 Remove the white space in columns name

There is white space in the title of each column. Removing the space can help further processing.

3.3 Check outlier of the dataset

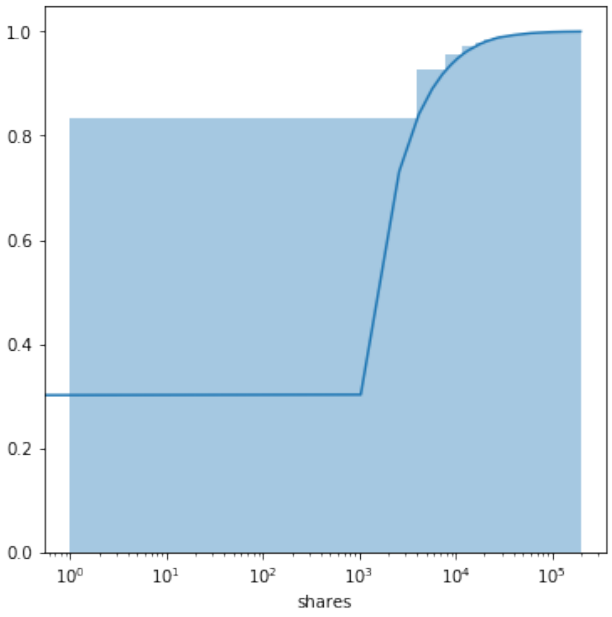
I plotted the distribution of columns with large stand deviation, which is timedelta, n_token_content and shares. There is no obvious outlier for timedelta and n_token_content. There is some outlier for shares, which is very high. I keep it because it is the target data (popular articles) of the analysis.

3.4 Add 'weekday' and 'data_channel' columns to aggregate the info

The weekday and data_channel are dummy variables now. I created new object columns to store the weekday and data_channel information. It will help for data exploratory.

3.5 CDF plot of shares

Following is the cumulative distribution function of shares. The aim is to find top 10% of shares. According to CDF plot, the articles with over 6000 times shares is among the top 10%.

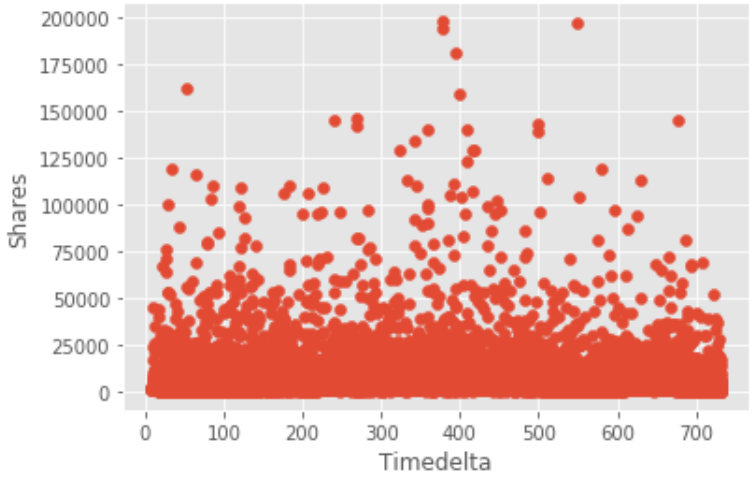


Then I created a column named popularity, stored 1 for more than 6000 shares and 0 for otherwise. This column will serve label for machine learning models.

4. Data Exploratory

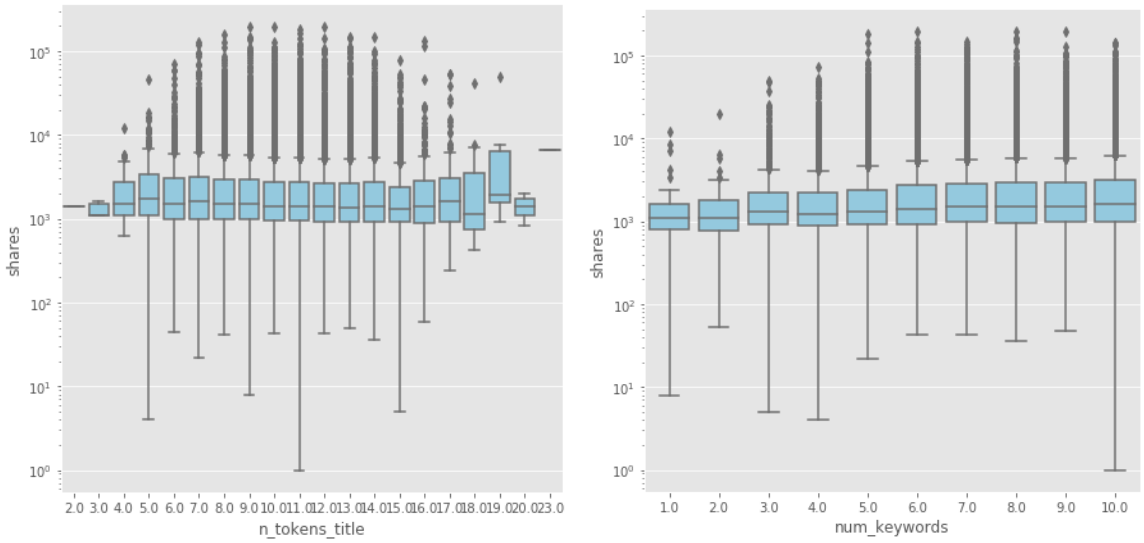
4.1 Explore the features that affect shares

4.1.1 Timedelta: Days between the article publication and the dataset acquisition



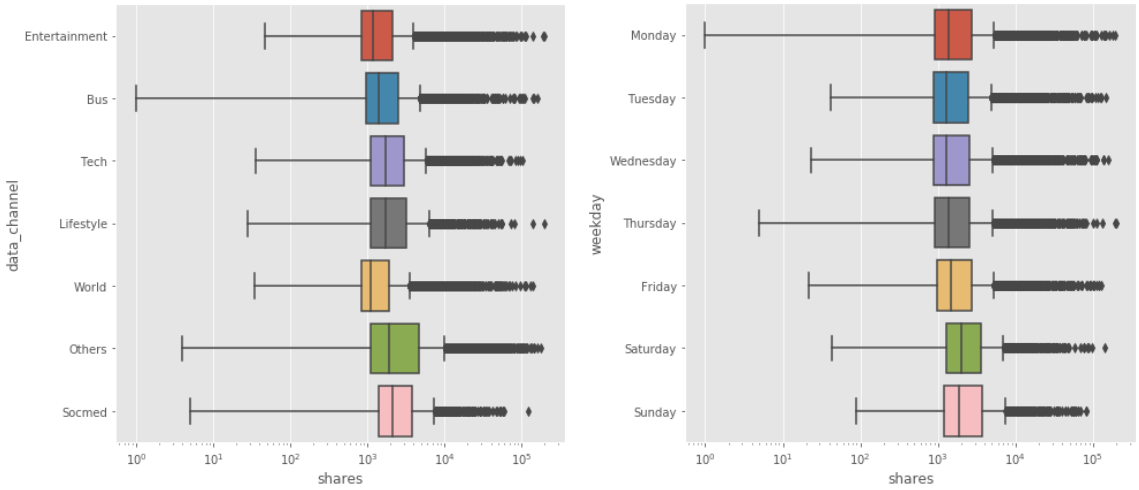
There is no obvious effect on shares.

4.1.2 n_tokens_title: the Number of words in the title and num_keywords: the number of keywords in the metadata



There is no obvious effect on shares.

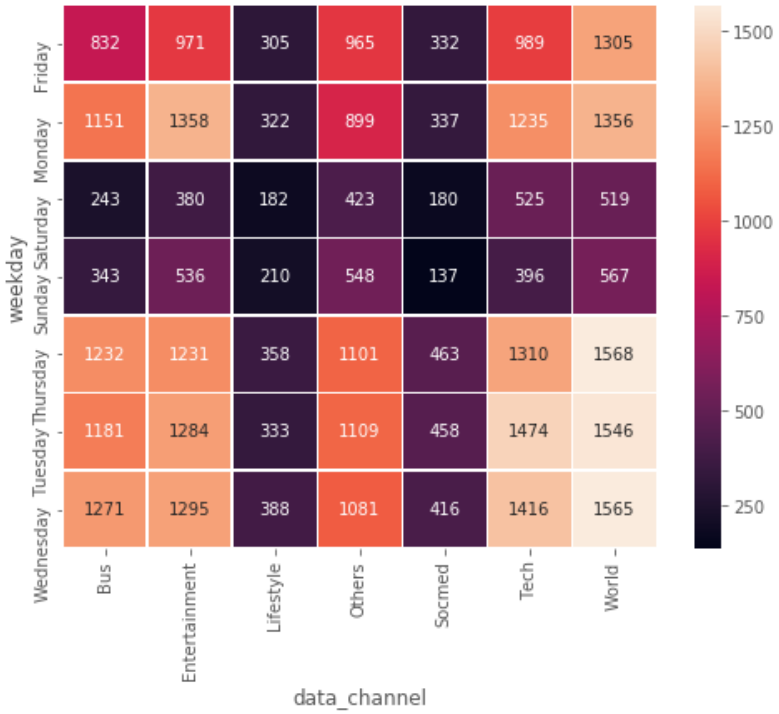
4.1.3 Data channel (includes World, Tech, Entertainment, Bus, Socmed, Lifestyle) and Weekday (through Monday to Sunday)



There is slightly higher mean for Saturday and Sunday. For data_channel, the world topic is slightly lower, while socmed is slightly higher than others. It has been noticed that the stand deviation of other in data_channel is higher than other topics. It could be more accurate of we category with other topics.

4.2 Explore the co-effect of features

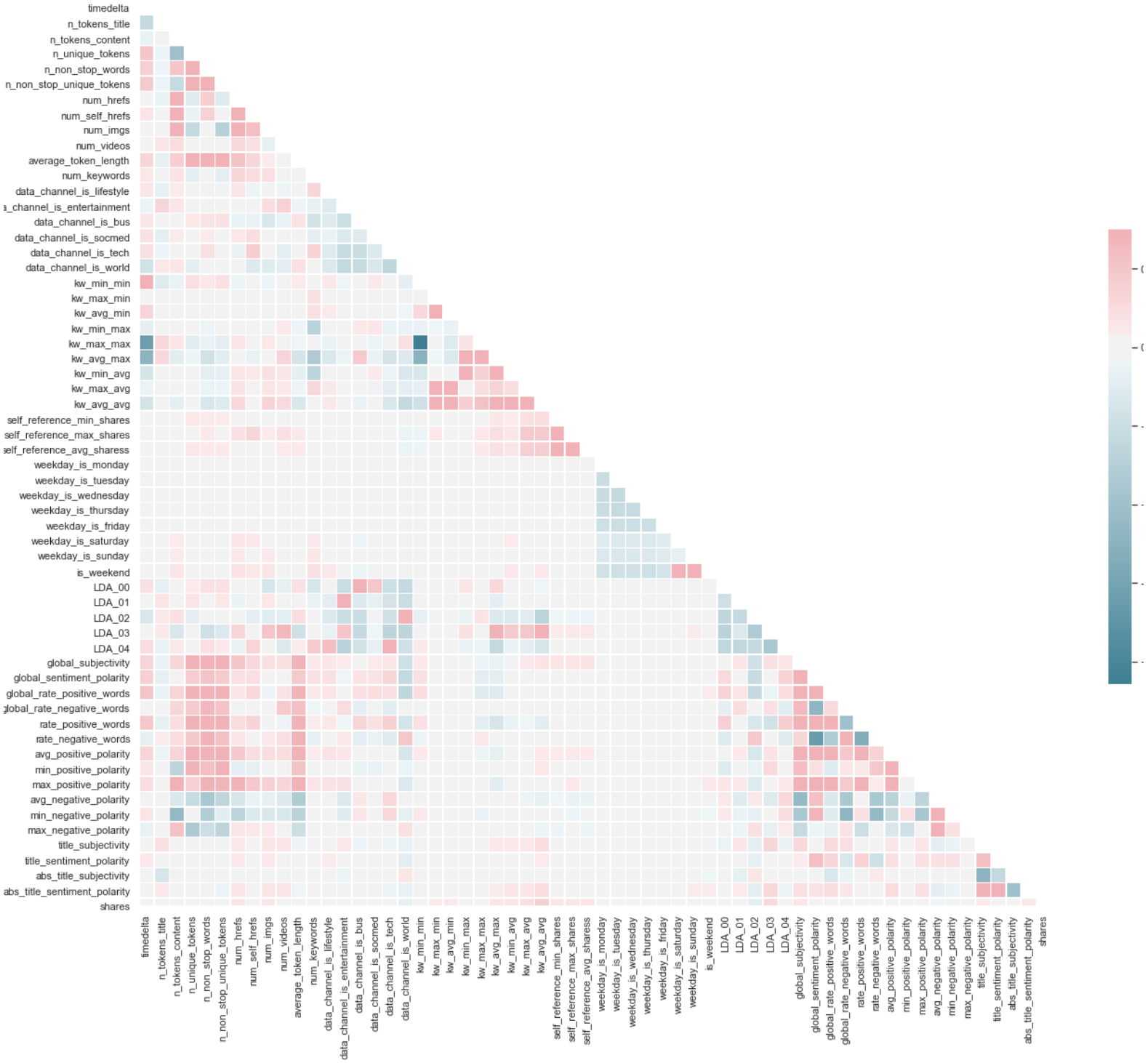
Weekday+Topic: Is there any preference of articles published during workday and weekend?



There are fewer articles published on weekends. Also, the lifestyle and socmed articles are less then others. Fewer articles may be the reason of higher mean of shares.

4.3 Correlation Heatmap of the features

Correlation heatmap is a direct method to explore the correlation between features.



Most of the features don't have correlation with shares. Only KW(keyword) and LDA (generated by topic model) is correlated with shares. Positive KW is positively correlated. LDA_3 is positively correlated, while LDA_2 is negatively correlated.

Most features are not strongly correlated with each other, except for few. For example, positive kw and negative kw.

Because there are no strongly corrected features, and no strong correlation between different features, I will use all feature for machine learning process.

5. Machine Learning

5.1 Model selection

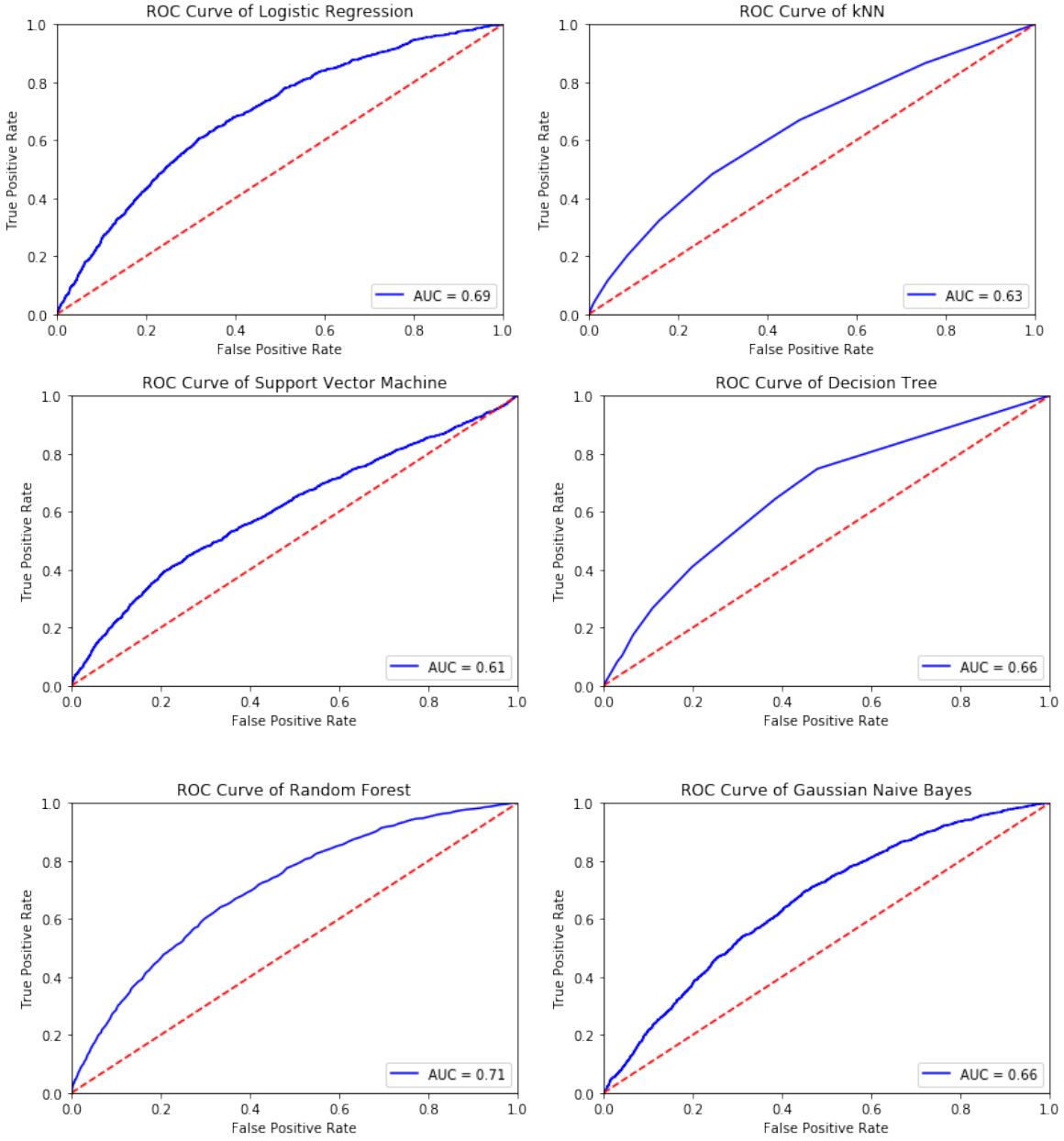
- Logistic Regression
- K-Nearest Neighbor
- Support Vector Machine
- Decision Trees
- Random Forest
- Naive Bayes

5.2 Model Evaluation by accuracy score

Logistic Regression	0.9
K-Nearest Neighbor	0.9
Support Vector Machine	0.9
Decision Trees	0.9
Random Forest	0.9
Naive Bayes	0.83

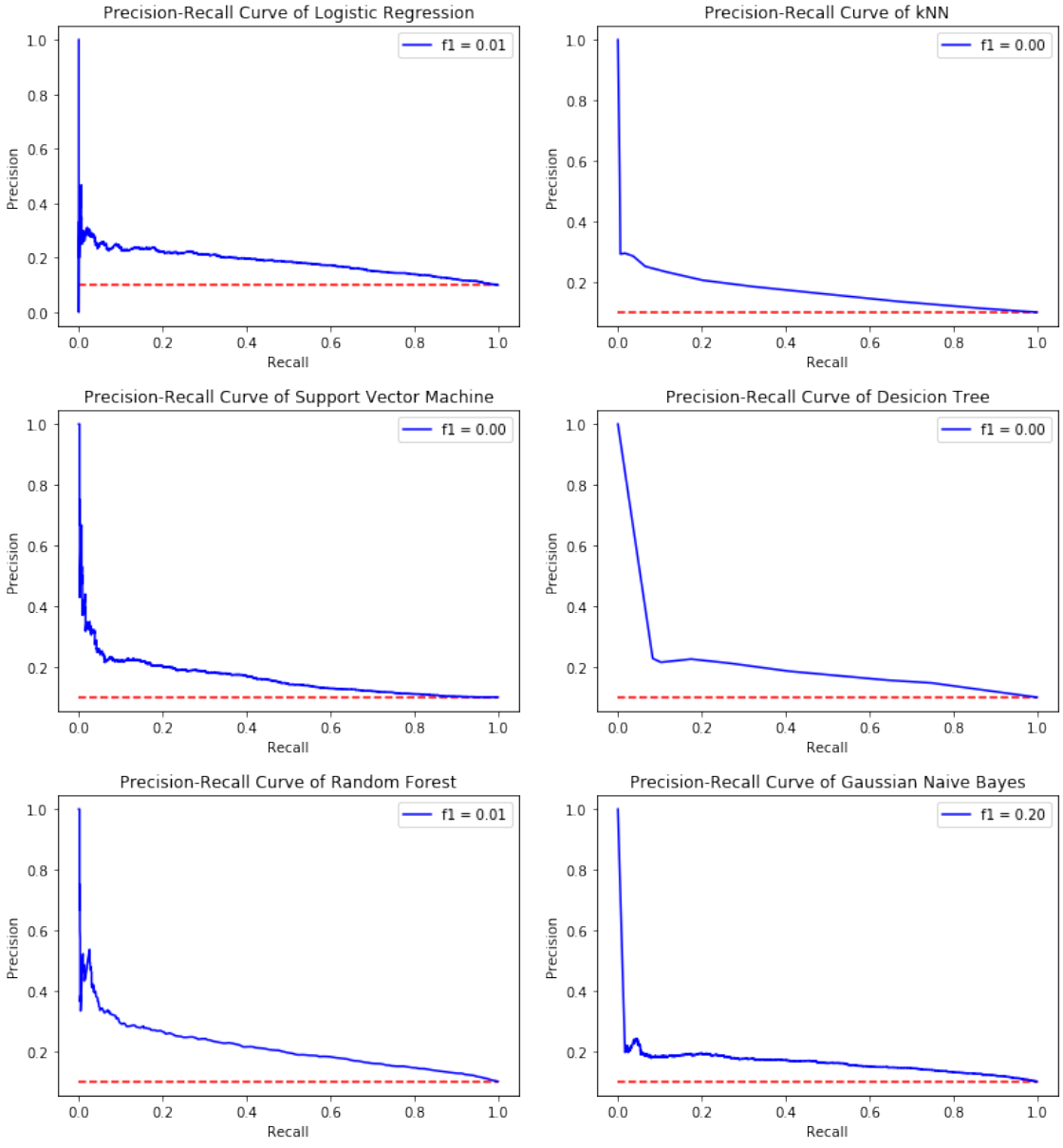
Because the dataset is unbalanced (10% data are popular), the accuracy score is not a proper metric for model evaluation. The same problem exists when we use confusion metrics. So I will give up confusion metrics.

5.3 Model Evaluation by ROC curve and AUC



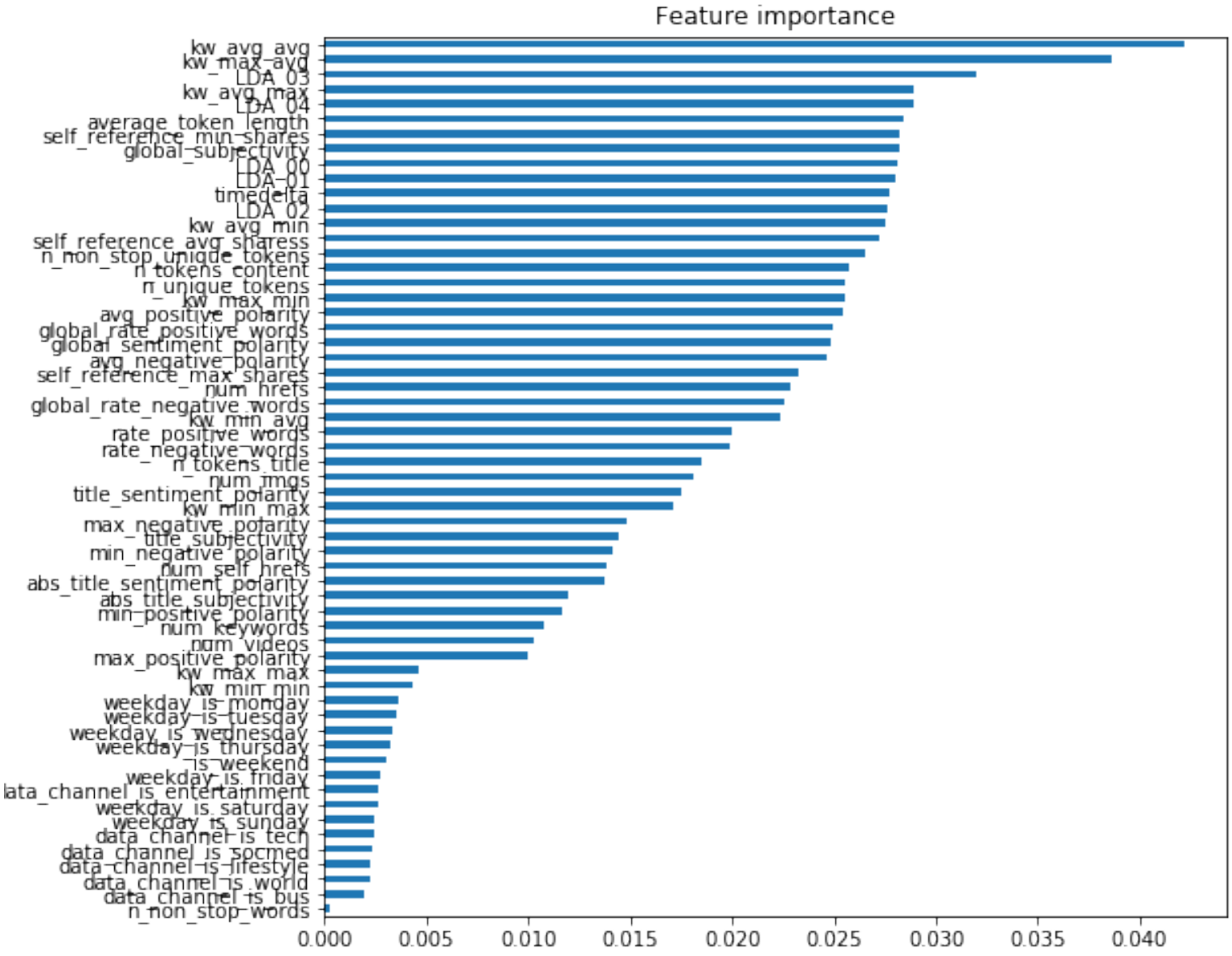
Within all models, random forest has best performance with AUC=0.71

5.4 Model Evaluation by Precison-Recall Curve and f1-score



The Gaussian Naïve Bayes model has obvious advantage than other when evaluated by precision-recall curve. The f1-score is 0.2 vs 0-0.01. Considering the imbalance of dataset, precision-recall can better describe the result of prediction. So I will choose the one with higher f1 score than higher AUC and accuracy score, which is Gaussian Naïve Bayes model.

5.5 Feature Importance



Keyword and LDA is the top features that affect news popularity. Weekday and data channel have least influence on new popularity.

6. Conclusion & Next step

6.1 The model can be applied to comparing the possible popularity of online news. Even though it is not accurate to predict if or not the news will be popular, the performance is not bad of predicting the possibility to be popular, which is enough for us to determine the one that standing out. Then the platform can either promote the article or put on front page to attract the attention of readers.

6.2 The Gaussian Naïve Bayes model is the best choice over all selected classifiers. The precision-recall curve weight the most through all evaluation result, because the dataset is unbalanced. There are a lot false negative, and precision-recall curve is most sensitive to the error in such case. Even though random forest model has better performance when evaluated by ROC cruve, Gaussian Naïve Bayes is still better choice because it cause much less computer calculation power.

6.3 The top important features are keyword and LDA. Both features are generated by statistic models. The regular categories have very little affect on popularity, like data channel and weekday. Thus, it is hard for writers to modify the articles for higher popularity purposely, and leaves all decision for network platform.

6.4 Keep monitoring the performance after applying the model. If there is large error in a month, we may need to change the features and train the model again.