Project: Online News Popularity

by Zhaofeng, Xiang, Weixin, Kushal and Gokul

Introduction

Our Topic: Online News Popularity

We are in Information Era

- Online news: a crucial channel for people to get updated information worldwide.
- Millions of differents news being updated online everyday.
- Important to know readers’ preference of news.

Our Objective: Make Prediction For News Companies

- To make a prediction of the popularity of online news.
- We believe that our paper can be helpful for news companies to make strategies for attracting more viewers.

Our dataset:

- The Online News Popularity dataset from UCI repository

https://archive.ics.uci.edu/ml/datasets/online+news+popular

Methodology

- Data and sample
- Description of Variables
- Data pre-processing
- Exploratory analytical methods
  -Classification
  -Regression
- Data visualization Analysis

Data and sample

- From UCI Machine learning repository, detailed data includes date, href details, positive/negative polarity of its over all post, sentimental polarity, title polarity, number of tokens in title, number of keywords, and so on.
- The dataset were published by Mashable (www.mashable.com) and the acquisition date was on January 8, 2015.

Descriptions of variables

This table lists these attributes by category.

Table 2: List of attributes by category.

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

- 61 (58 predictive attributes, 2 non-predictive, 1 goal field) numbers of attributes in our dataset.
- To generate these variables, we have date, href details, positive/negative polarity of its over all post, sentimental polarity, title polarity, number of tokens in title, number of keywords, the number of shares and so on.

Data Pre-processing

Raw data :

Incomplete,

Inconsistent,

Lacking in certain behaviors or trends

Errors! ! !

Data pre-processing

Data Preprocessing

Refining: 7 channel catagories+several small topics

1	Channel
Business'	Media
	Business
	Small Business
	Paid Content
	Advertising
	Music
	Startups
	Marketing
	Mobile
	Gaming

2	Channel
lifestyle	Gadgets
	Marketing
	How To
	Apps & Software
	Dev & Design
	Gaming
	Lifestyle
	U.S.
	Mobile
	Startups
	Music
	Paid Content
	Small Business
	Conversations
	Sports
	Photography
	Memes
	Advertising
	Movies

3	Channel
'Entertainment'	Media
	Entertainment

4	Channel
Social Media	Social Media
	Movies
	Marketing
	Media
	Music
	How To
	Paid Content
	Mobile
	Startups
	Gaming

5	Channel
'Tech'	Tech
	Mobile
	Music
	Gaming
	Small Business
	Paid Content
	Gadgets
	Media
	Marketing
	Movies
	Startups
	Advertising
	How To

6	Channel
world	U.S.
	World
	Mobile
	Gaming
	Small Business
	Movies
	Gadgets
	Media
	Dev & Design
	Apps & Software
	Advertising
	Music
	Startups
	Paid Content

7	Channel
Watercooler	Watercooler
	Gadgets
	U.S.
	Mobile
	Media
	Movies
	Music
	Marketing
	Startups
	Advertising
	Gaming
	Apps & Software
	Paid Content
	Small Business
	Dev & Design

Data Preprocessing

Data preprocessing for regression (data reduction)

Data preprocessing for association (data reduction & data transformation)

Data preprocessing for classification (data discretization)

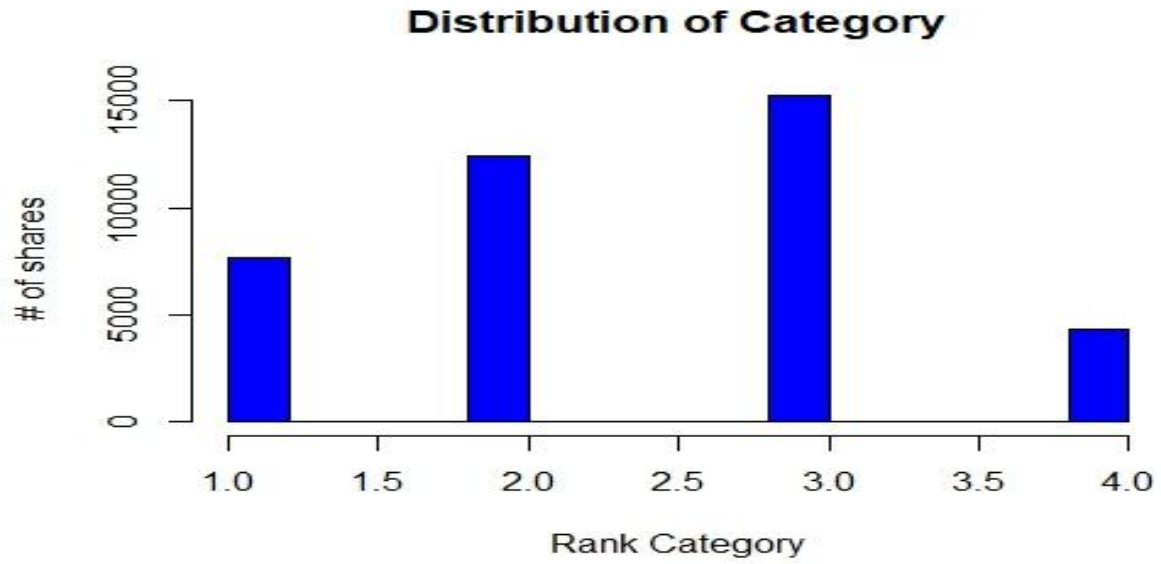
Data preprocessing for visualization (data scraping)

Exploratory Analytical Methods

Classification

Data Preprocessing for Classification

Feature Engineering to categorize Number of Shares to class labels 1, 2, 3 & 4.

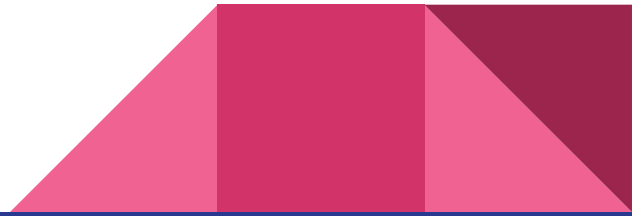


Data Preprocessing for Classification (Contd.)

Data Cleansing for missing values and noise.

Scaling of attributes.

Data transformation using Principal Component Analysis (PCA) to select key attributes.



Why Classification Model - C5.0?

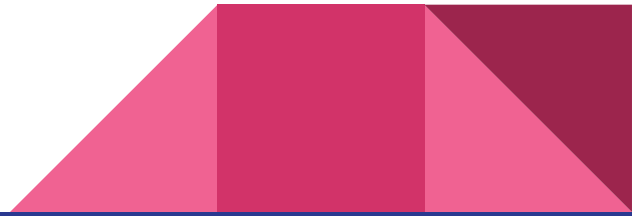
C4.5 has a better handling for both discrete and continuous attributes.

C4.5 algorithm prunes the tree after creation.

C5.0 is significantly faster than C4.5

C5.0 fetches similar results as C4.5 with considerably smaller decision trees.

Using Random Forest algorithm may lead to the problem of overfitting



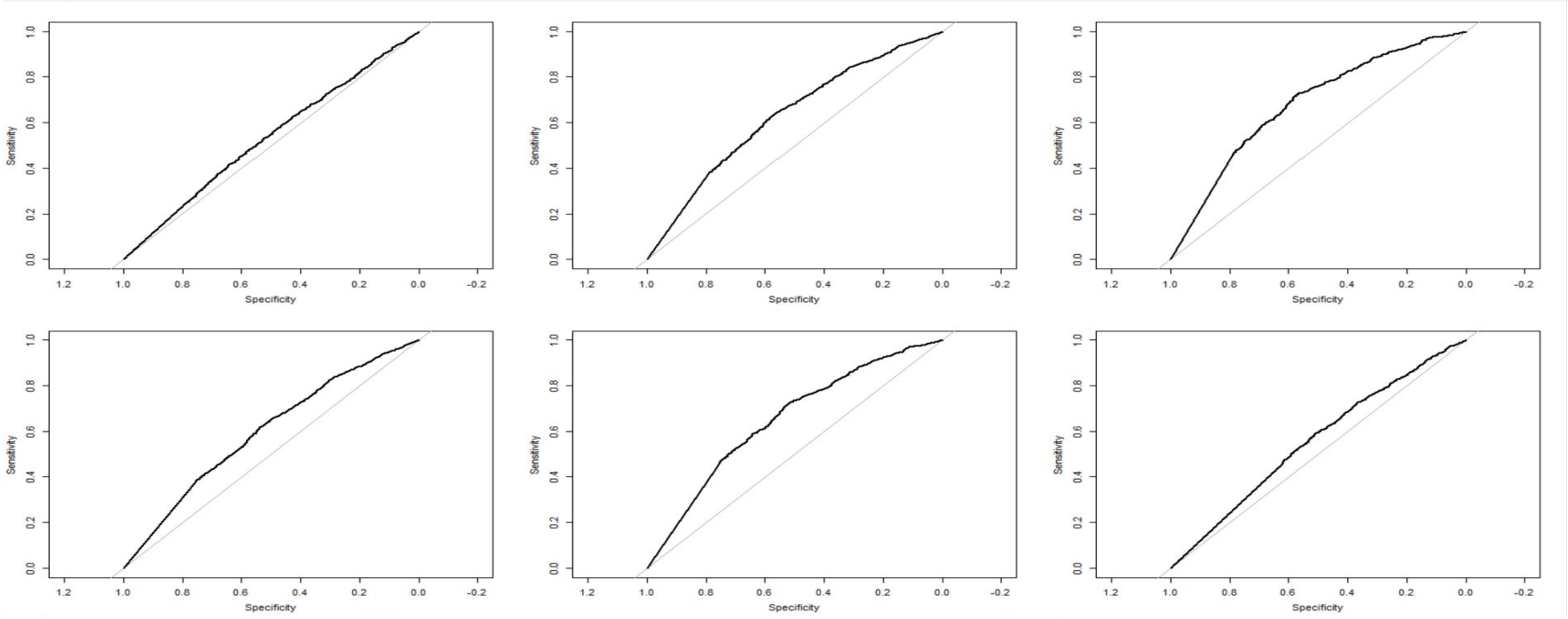
Results of Classification Model - C5.0

Confusion Matrix for C5.0 Model

Prediction \ Reference	1	2	3	4
1	369	377	262	73
2	615	874	680	151
3	503	1172	1943	563
4	29	56	100	64

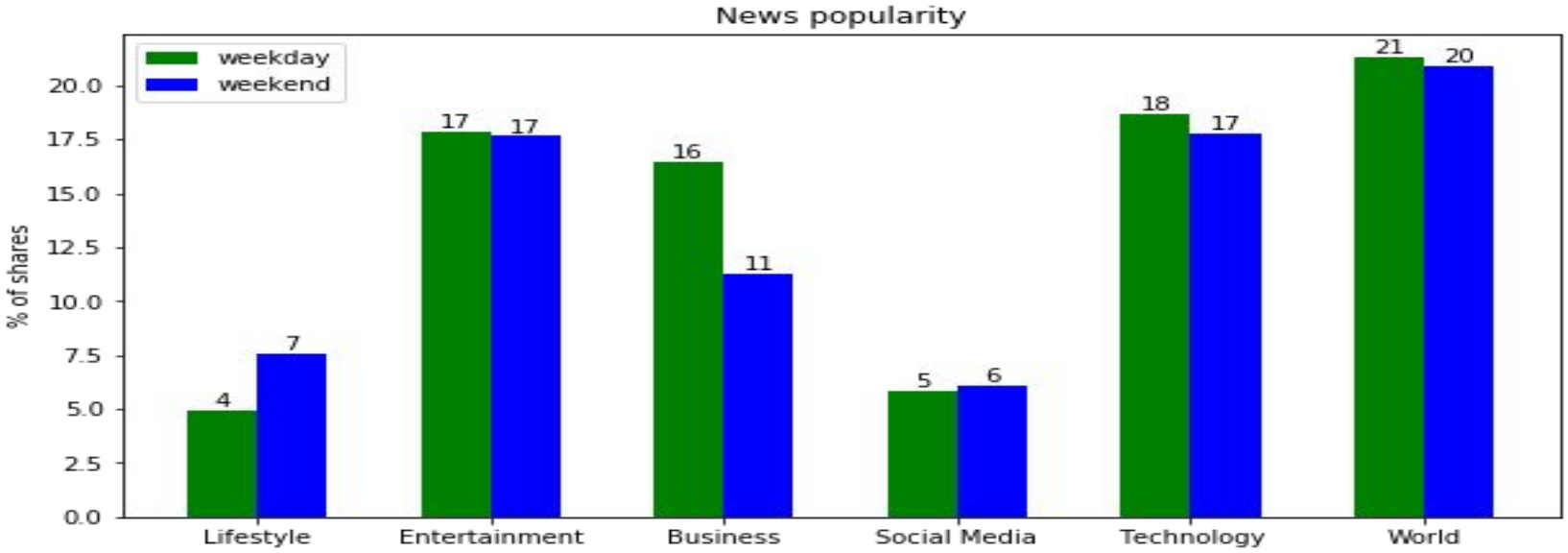
Results of Classification Model - C5.0 (Contd.)

ROC curve



Research Question

Does publication of certain data channel on weekends/weekdays have impact on sharing and popularity?



Exploratory Analytical Methods

Association

Association

```
Apriori
=======

Minimum support: 0.5 (4375 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 4

Best rules found:

 1. n_tokens_content=t 6142 ==> n_non_stop_words=t 6142    <conf:(1)> lift:(1) lev:(0) [22] conv:(22.46)
 2. n_tokens_content=t  global_subjectivity=t 6132 ==>  n_non_stop_words=t 6132    <conf:(1)> lift:(1) lev:(0) [22] conv:(22.43)
 3. global_sentiment_polarity=t 4452 ==>  global_subjectivity=t 4448    <conf:(1)> lift:(1) lev:(0) [15] conv:(3.97)
 4. n_non_stop_words=t  global_sentiment_polarity=t 4441 ==>  global_subjectivity=t 4437    <conf:(1)> lift:(1) lev:(0) [15] conv:(3.96)
 5. n_tokens_content=t 6142 ==>  global_subjectivity=t 6132    <conf:(1)> lift:(1) lev:(0) [17] conv:(2.49)
 6. n_tokens_content=t  n_non_stop_words=t 6142 ==>  global_subjectivity=t 6132    <conf:(1)> lift:(1) lev:(0) [17] conv:(2.49)
 7. n_tokens_content=t 6142 ==>  n_non_stop_words=t  global_subjectivity=t 6132    <conf:(1)> lift:(1.01) lev:(0) [39] conv:(4.47)
 8. rate_positive_words=t 4658 ==>  n_non_stop_words=t 4647    <conf:(1)> lift:(1) lev:(0) [6] conv:(1.42)
 9. global_subjectivity=t  rate_positive_words=t 4642 ==>  n_non_stop_words=t 4631    <conf:(1)> lift:(1) lev:(0) [5] conv:(1.41)
10. global_sentiment_polarity=t 4452 ==>  n_non_stop_words=t 4441    <conf:(1)> lift:(1) lev:(0) [5] conv:(1.36)
11. global_subjectivity=t  global_sentiment_polarity=t 4448 ==>  n_non_stop_words=t 4437    <conf:(1)> lift:(1) lev:(0) [5] conv:(1.36)
12. n_non_stop_words=t  num_keywords=t 4392 ==>  global_subjectivity=t 4378    <conf:(1)> lift:(1) lev:(0) [5] conv:(1.31)
13. global_sentiment_polarity=t 4452 ==>  n_non_stop_words=t  global_subjectivity=t 4437    <conf:(1)> lift:(1) lev:(0) [20] conv:(2.23)
14. num_keywords=t 4411 ==>  global_subjectivity=t 4396    <conf:(1)> lift:(1) lev:(0) [4] conv:(1.23)
15. rate_positive_words=t 4658 ==>  global_subjectivity=t 4642    <conf:(1)> lift:(1) lev:(0) [4] conv:(1.22)
16. n_non_stop_words=t  rate_positive_words=t 4647 ==>  global_subjectivity=t 4631    <conf:(1)> lift:(1) lev:(0) [4] conv:(1.22)
17. global_subjectivity=t 8711 ==>  n_non_stop_words=t 8680    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
18. num_keywords=t  global_subjectivity=t 4396 ==>  n_non_stop_words=t 4378    <conf:(1)> lift:(1) lev:(-0) [-1] conv:(0.85)
19. num_keywords=t 4411 ==>  n_non_stop_words=t 4392    <conf:(1)> lift:(1) lev:(-0) [-2] conv:(0.81)
20. n_non_stop_words=t 8718 ==>  global_subjectivity=t 8680    <conf:(1)> lift:(1) lev:(0) [0] conv:(1)
```

- According to the association rules, we got top20 rules by using APRIORI method, we know that ‘n_tokens_content’ which means Number of words in the content usually come up with ‘n_non_stop_words’, which means Rate of non-stop words in the content.
- When a news data has ‘n_tokens_content’ attribute and ‘global_subjectivity’, which means Text subjectivity, it has high probability to show up with ‘n_non_stop_words’, which means Rate of non-stop words in the content.

Exploratory Analytical Methods

Regression

Linear Regression

```
shares =

    55.8073 *  n_tokens_title +
  -593.7402 *  n_non_stop_words +
   960.0803 *  n_non_stop_unique_tokens +
    31.7819 *  num_hrefs +
   -47.5103 *  num_self_hrefs +
    17.8001 *  num_imgs +
  -548.3959 *  average_token_length +
    79.8234 *  num_keywords +
    -0.4242 *  kw_min_avg +
    -0.1929 *  kw_max_avg +
     1.6963 *  kw_avg_avg +
     0.0264 *  self_reference_min_shares +
     0.0052 *  self_reference_max_shares +
    -0.0052 *  self_reference_avg_sharess +
   307.7646 *  is_weekend +
  2732.4499 *  global_subjectivity +
 -7578.1641 *  global_rate_positive_words +
  1203.5637 *  rate_positive_words +
   716.3292 *  rate_negative_words +
 -1769.2139 *  min_positive_polarity +
 -1307.9448 *  avg_negative_polarity +
   669.5334 *  abs_title_subjectivity +
   658.3899 *  abs_title_sentiment_polarity +
-1968.166
```

- Videos doesn’t matter. In the function, there are no attribute called ‘num_videos’, which is put into the dataset. It tells us that the number of video does not affect the number of shares, so that it does not show up in the equation.

Linear Regression

- The words in title do matter. In the model function, the coefficient value before the attribute ‘n_token_title’ is 55.8073, which means that it adds 55 new shares per ‘n_token_title’ increase one
- ‘average_token_length’ do harm to the shares. This attribute represents the average length of the words in the content. According to the equation above, the coefficient before this attribute is -548, which means that there are 548 shares lost per average length of the words in the content be added.

PCA Regression

```
## lm(formula = shares ~ ., data = Dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27963  -2277  -1204    -71 837227
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.954e+05  6.129e+06   0.032 0.974570
## timedelta                     1.678e+00  3.924e-01   4.275 1.91e-05 ***
## n_tokens_title                1.125e+02  2.915e+01   3.858 0.000114 ***
## n_tokens_content              5.896e-01  2.235e-01   2.638 0.008332 **
## n_unique_tokens               3.354e+03  1.924e+03   1.744 0.081228 .
## n_non_stop_words             -1.583e+03  5.910e+03  -0.268 0.788858
## n_non_stop_unique_tokens     -1.375e+03  1.630e+03  -0.844 0.398852
## num_hrefs                     2.619e+01  6.705e+00   3.906 9.38e-05 ***
## num_self_hrefs               -6.159e+01  1.784e+01  -3.453 0.000556 ***
## num_imgs                      1.148e+01  8.941e+00   1.284 0.199124
## num_videos                    4.883e+00  1.575e+01   0.259 0.795447
## average_token_length         -5.439e+02  2.430e+02  -2.238 0.025219 *
## num_keywords                  5.507e+01  3.715e+01   1.482 0.138246
## data_channel_is_lifestyle    -9.580e+02  3.952e+02  -2.424 0.015336 *
## data_channel_is_entertainment -1.076e+03  2.563e+02  -4.198 2.70e-05 ***
## data_channel_is_bus          -7.752e+02  3.827e+02  -2.026 0.042798 *
## data_channel_is_socmed       -5.249e+02  3.727e+02  -1.408 0.159039
## data_channel_is_tech         -4.774e+02  3.717e+02  -1.284 0.199026
## data_channel_is_world        -3.136e+02  3.784e+02  -0.829 0.407289
## kw_min_min                    1.592e+00  1.629e+00   0.977 0.328425
## kw_max_min                    1.879e-01  5.835e-02   2.144 0.032034 *
## kw_avg_min                   -4.935e-01  3.097e-01  -1.594 0.110971
## kw_min_max                   -2.487e-03  1.177e-03  -2.112 0.034674 *
## kw_max_max                   -2.459e-05  5.898e-04  -0.042 0.966748
## kw_avg_max                    4.521e-05  8.481e-04   0.053 0.957490
## kw_min_avg                   -3.641e-01  7.565e-02  -4.813 1.49e-06 ***
## kw_max_avg                   -2.861e-01  2.530e-02  -8.143 3.95e-16 ***
## kw_avg_avg                    1.685e+00  1.439e-01  11.707  < 2e-16 ***
## self_reference_min_shares     2.659e-02  7.524e-03   3.534 0.000409 ***
## self_reference_max_shares     5.879e-03  4.083e-03   1.440 0.149849
## self_reference_avg_sharess   -6.423e-03  1.044e-02  -0.615 0.538339
## weekday_is_monday             2.655e+02  2.631e+02   1.009 0.312888
## weekday_is_tuesday           -2.805e+02  2.592e+02  -1.082 0.279135
## weekday_is_wednesday         -1.198e+02  2.591e+02  -0.462 0.643813
## weekday_is_thursday          -2.918e+02  2.597e+02  -1.124 0.261055
## weekday_is_friday            -2.520e+02  2.689e+02  -0.937 0.348752
## weekday_is_saturday           3.804e+02  3.205e+02   1.187 0.235301
## weekday_is_sunday                    NA         NA      NA       NA
## is_weekend                           NA         NA      NA       NA
## LDA_00                       -1.969e+05  6.129e+06  -0.032 0.974376
## LDA_01                       -1.977e+05  6.129e+06  -0.032 0.974272
## LDA_02                       -1.981e+05  6.129e+06  -0.032 0.974213
## LDA_03                       -1.973e+05  6.129e+06  -0.032 0.974321

## LDA_04                       -1.973e+05  6.129e+06  -0.032 0.974322
## global_subjectivity           2.497e+03  8.504e+02   2.936 0.003322 **
## global_sentiment_polarity     8.146e+02  1.668e+03   0.489 0.625174
## global_rate_positive_words   -1.392e+04  7.165e+03  -1.943 0.052036 .
## global_rate_negative_words    1.041e+02  1.368e+04   0.008 0.993927
## rate_positive_words           2.024e+03  5.775e+03   0.350 0.726025
## rate_negative_words           2.114e+03  5.821e+03   0.363 0.716520
## avg_positive_polarity        -1.685e+03  1.366e+03  -1.233 0.217400
## min_positive_polarity        -1.898e+03  1.144e+03  -1.659 0.097057 .
## max_positive_polarity         3.113e+02  4.311e+02   0.722 0.470213
## avg_negative_polarity        -1.707e+03  1.258e+03  -1.356 0.175038
## min_negative_polarity         8.207e+01  4.590e+02   0.179 0.858083
## max_negative_polarity        -1.787e+02  1.046e+03  -0.171 0.864428
## title_subjectivity           -9.160e+01  2.741e+02  -0.334 0.738249
## title_sentiment_polarity      2.041e+02  2.504e+02   0.815 0.414962
## abs_title_subjectivity        6.557e+02  3.640e+02   1.801 0.071634 .
## abs_title_sentiment_polarity  6.199e+02  3.957e+02   1.567 0.117183
## ---
```

- R^2 and Adjusted R^2 are around 0.02,which indicate that the full models can explain extremely small part of data set.

- the coefficients of weekday_is_Sunday and weekday_is_Saturday is NA.

- variable weekday_is_Sunday: 2737
  variable weekday_is Saturday: 2453
  39644 observations, the information provided by these two variables is not enough, which can cause the singularity of matrix xTx
  The singularity of matrix xTx can cause the missing of some coefficients.

Principal Components Regression

- reducing dimensionality
- decreasing computational cost

PCA Regression

```
## Importance of components:
##                             PC1       PC2       PC3       PC4       PC5
## Standard deviation     2.316e+05 1.073e+05 4.985e+04 4.748e+04 1.862e+04
## Proportion of Variance 7.628e-01 1.638e-01 3.533e-02 3.205e-02 4.930e-03
## Cumulative Proportion  7.628e-01 9.266e-01 9.620e-01 9.940e-01 9.990e-01
##                             PC6       PC7       PC8       PC9  PC10  PC11
## Standard deviation     6.582e+03 4.494e+03 2.842e+03 1.088e+03 496.1 453.2
## Proportion of Variance 6.200e-04 2.900e-04 1.100e-04 2.000e-05   0.0   0.0
## Cumulative Proportion  9.996e-01 9.999e-01 1.000e+00 1.000e+00   1.0   1.0
##                          PC12  PC13 PC14  PC15  PC16  PC17  PC18  PC19
## Standard deviation     192.5 156.2 35.5 10.21 7.331 7.067 3.912 3.342
## Proportion of Variance   0.0   0.0  0.0  0.00 0.000 0.000 0.000 0.000
## Cumulative Proportion    1.0   1.0  1.0  1.00 1.000 1.000 1.000 1.000
##                          PC20  PC21  PC22  PC23   PC24   PC25   PC26   PC27
## Standard deviation     2.039 1.625 0.813 0.502 0.4702 0.4385 0.4323 0.4302
## Proportion of Variance 0.000 0.000 0.000 0.000 0.0000 0.0000 0.0000 0.0000
## Cumulative Proportion  1.000 1.000 1.000 1.000 1.0000 1.0000 1.0000 1.0000
##                            PC28   PC29   PC30   PC31   PC32   PC33   PC34
## Standard deviation     0.4278 0.4139 0.3914 0.3856 0.2903 0.2689 0.2555
## Proportion of Variance 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## Cumulative Proportion  1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
##                            PC35   PC36   PC37  PC38   PC39   PC40   PC41
## Standard deviation     0.2449 0.2373 0.2335 0.214 0.1641 0.1607 0.1559
## Proportion of Variance 0.0000 0.0000 0.0000 0.000 0.0000 0.0000 0.0000
## Cumulative Proportion  1.0000 1.0000 1.0000 1.000 1.0000 1.0000 1.0000
##                          PC42  PC43   PC44   PC45    PC46    PC47    PC48
## Standard deviation     0.151 0.138 0.1277 0.1026 0.08898 0.07567 0.06863
## Proportion of Variance 0.000 0.000 0.0000 0.0000 0.00000 0.00000 0.00000
## Cumulative Proportion  1.000 1.000 1.0000 1.0000 1.00000 1.00000 1.00000
##                            PC49    PC50    PC51    PC52    PC53    PC54
## Standard deviation     0.06665 0.04806 0.03985 0.03559 0.02914 0.02167
## Proportion of Variance 0.00000 0.00000 0.00000 0.00000 0.00000 0.00000
## Cumulative Proportion  1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
##                             PC55     PC56      PC57      PC58      PC59
## Standard deviation     0.01155 0.003939 4.214e-06 1.744e-11 1.744e-11
## Proportion of Variance 0.00000 0.000000 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  1.00000 1.000000 1.000e+00 1.000e+00 1.000e+00
```

seven principal components
- explain most variance.

seven four components to perform regression

PCA Regression

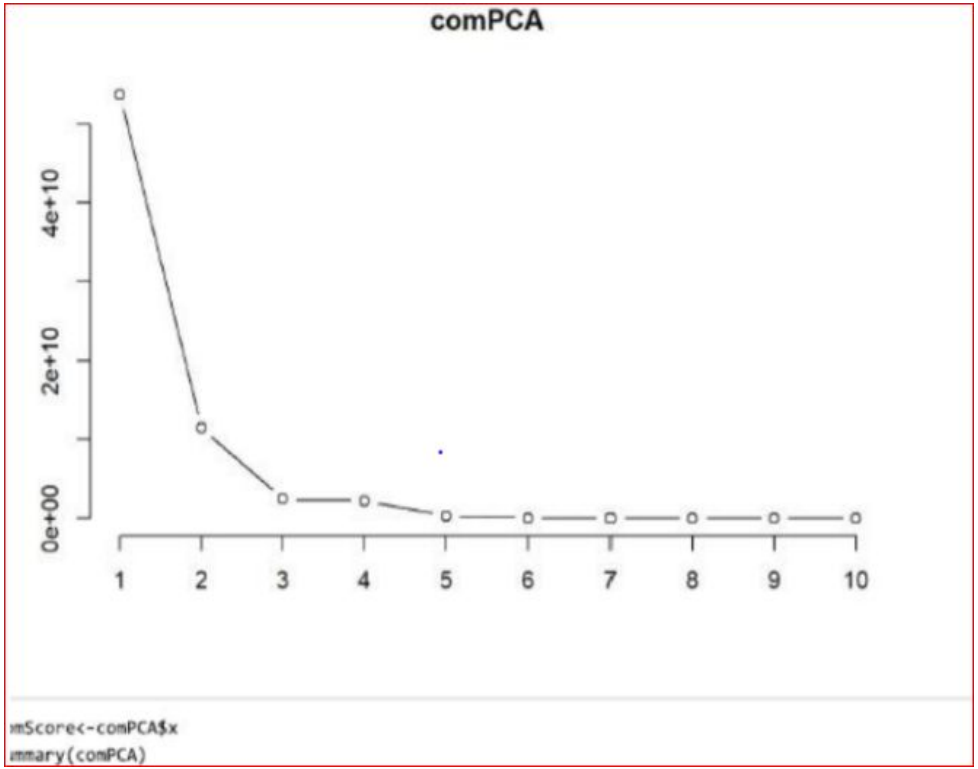
Mean square error : 123200737.
Bad !
Is not better than simple linear regression.
(R^2 is only 0.007)

```
##
## Call:
## lm(formula = V8 ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40724  -2284  -1654   -480 838749
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.358e+03  6.805e+01  49.350  < 2e-16 ***
## PC1          7.744e-04  2.929e-04   2.644  0.00819 **
## PC2         -4.804e-03  6.332e-04  -7.588 3.34e-14 ***
## PC3         -8.746e-03  1.369e-03  -6.390 1.68e-10 ***
## PC4          7.969e-03  1.448e-03   5.503 3.77e-08 ***
## PC5         -6.456e-03  3.710e-03  -1.740  0.08183 .
## PC6         -8.859e-02  9.836e-03  -9.007  < 2e-16 ***
## PC7          1.352e-02  1.517e-02   0.891  0.37288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11730 on 29725 degrees of freedom
## Multiple R-squared:  0.00731,    Adjusted R-squared:  0.007076
## F-statistic: 31.27 on 7 and 29725 DF,  p-value: < 2.2e-16
```

```
pca_pre<-predict(pca_fit,test)
MSE<-mean((pca_pre-test[,8])^2)
MSE # Used to compare different models
```

```
## [1] 123200737
```

PCA Regression



Bad results
- Less information
- Collinearity

Ridge Regression

```
## 60 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                   -1.793861e+03
## timedelta                      1.802823e+00
## n_tokens_title                 1.110597e+02
## n_tokens_content               4.317714e-01
## n_unique_tokens                6.584633e+00
## n_non_stop_words              -3.515134e+00
## n_non_stop_unique_tokens       5.720222e+00
## num_hrefs                      2.660563e+01
## num_self_hrefs                -6.335614e+01
## num_imgs                       7.238677e+00
## num_videos                     1.157819e+01
## average_token_length          -1.589680e+02
## num_keywords                   8.539575e+01
## data_channel_is_lifestyle     -1.090749e+03
## data_channel_is_entertainment -1.237649e+03
## data_channel_is_bus           -1.164383e+03
## data_channel_is_socmed        -5.030139e+02
## data_channel_is_tech          -4.754283e+02
## data_channel_is_world         -5.887924e+02
## kw_min_min                     1.603253e-01
## kw_max_min                     2.971068e-03
## kw_avg_min                     1.912590e-01
## kw_min_max                    -2.865725e-03
## kw_max_max                    -3.793539e-04
## kw_avg_max                     1.279919e-03
## kw_min_avg                    -1.458753e-01
## kw_max_avg                    -1.201892e-01
## kw_avg_avg                     1.169409e+00
## self_reference_min_shares      1.046404e-02
## self_reference_max_shares      5.439722e-03
## self_reference_avg_sharess    -1.336894e-03
## weekday_is_monday              3.051687e+02
## weekday_is_tuesday            -1.469474e+02
## weekday_is_wednesday           6.803334e+01
## weekday_is_thursday           -2.740554e+02
## weekday_is_friday             -8.383964e+01
## weekday_is_saturday            4.162125e+02
## weekday_is_sunday             -4.634198e+01
## is_weekend                     1.864841e+02
## LDA_00                         6.704297e+02
## LDA_01                        -1.523147e+02
## LDA_02                        -7.475333e+02
## LDA_03                         3.785326e+02
## LDA_04                        -1.157827e+02
## global_subjectivity            2.732672e+03
```

-The magnitudes of the coefficients are like that from simple linear regression.

-To achieve best models, we need to use stepwise methods to perform variable selections.

-However, using stepwise methods may not solve the problem of multicollinearity and have high computation cost.

Therefore, LASSO can be applied.

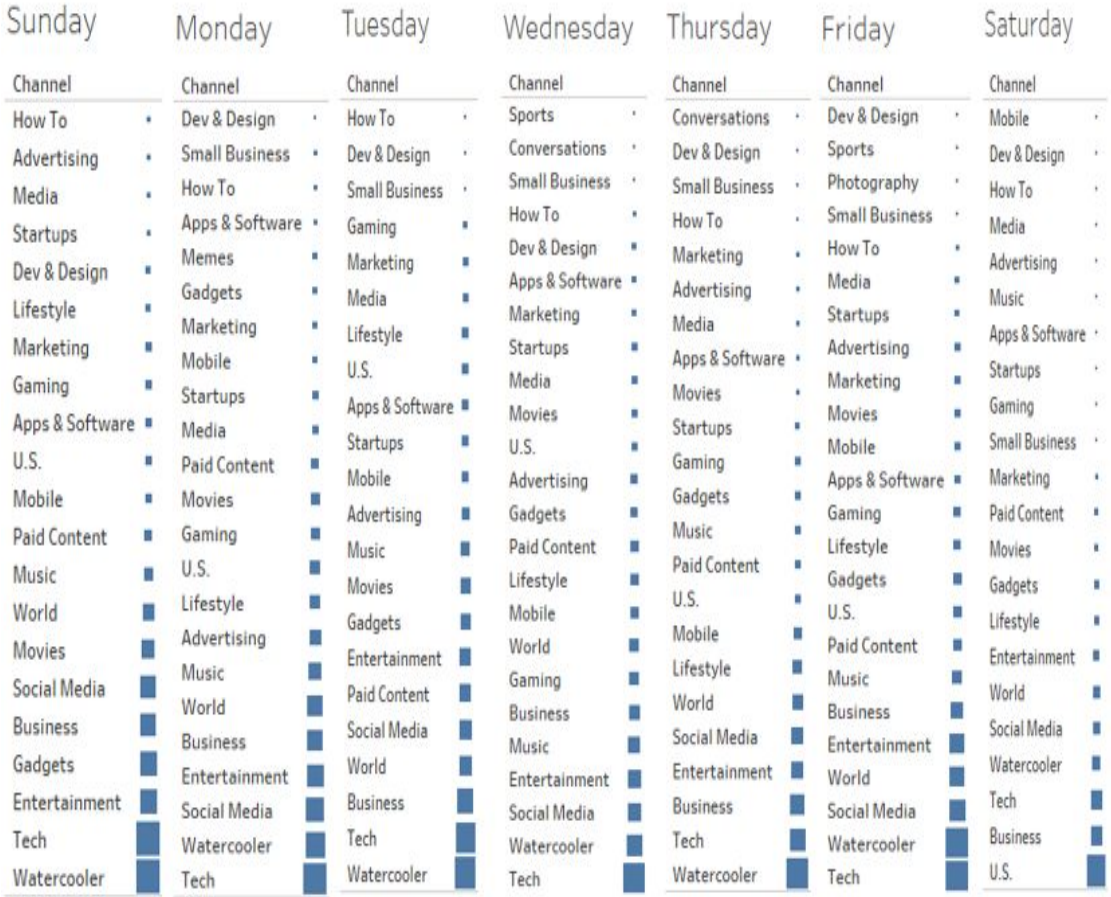
LASSO Regression

```
## 60 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                   -2.328615e+03
## timedelta                      1.796363e+00
## n_tokens_title                 1.063043e+02
## n_tokens_content               3.517763e-01
## n_unique_tokens                3.069667e+00
## n_non_stop_words               .
## n_non_stop_unique_tokens       .
## num_hrefs                      2.595584e+01
## num_self_hrefs                -5.603344e+01
## num_imgs                       5.839297e+00
## num_videos                     9.453074e+00
## average_token_length          -1.450603e+02
## num_keywords                   7.032808e+01
## data_channel_is_lifestyle     -7.007533e+02
## data_channel_is_entertainment -9.873508e+02
## data_channel_is_bus           -6.061569e+02
## data_channel_is_socmed        -9.425572e+01
## data_channel_is_tech          -4.779642e+01
## data_channel_is_world         -1.826990e+02
## kw_min_min                     .
## kw_max_min                     1.492264e-02
```

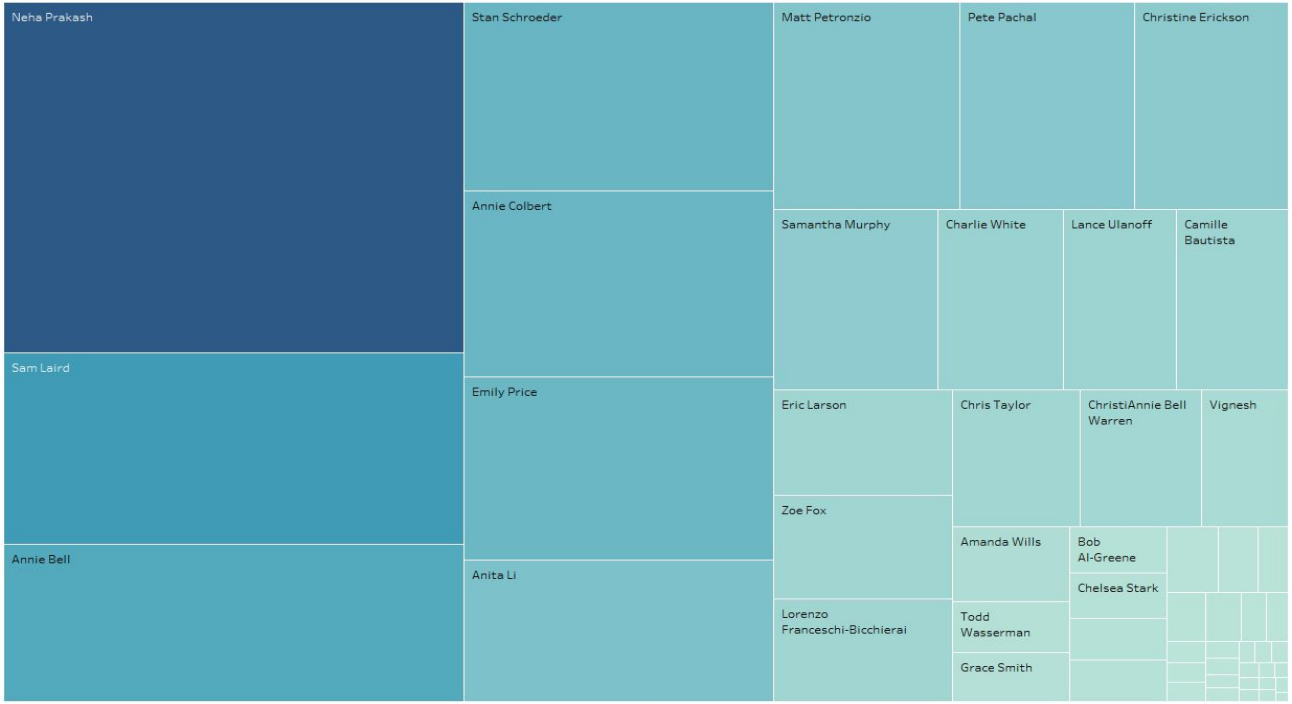
-The mean square error of the prediction is 121612240, which is smaller than that from ridge regression.

-So far, lasso performs the best prediction. From the result, we can see LASSO performs variable selections and disregard 12 variables.

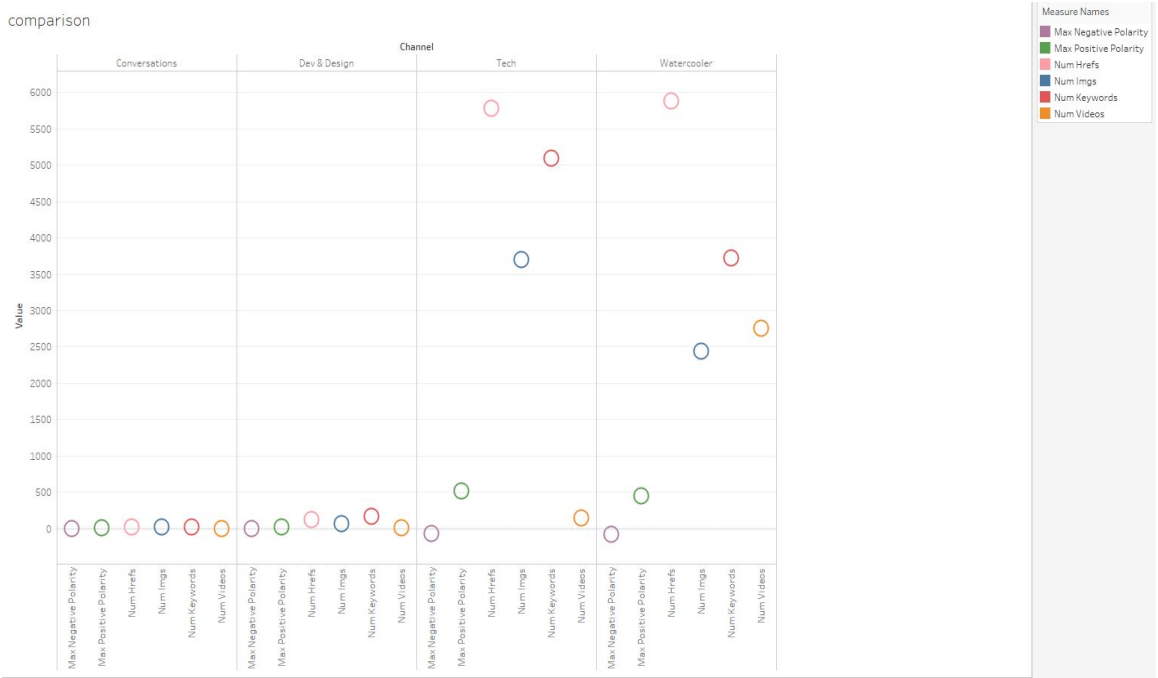
Data visualization Analysis



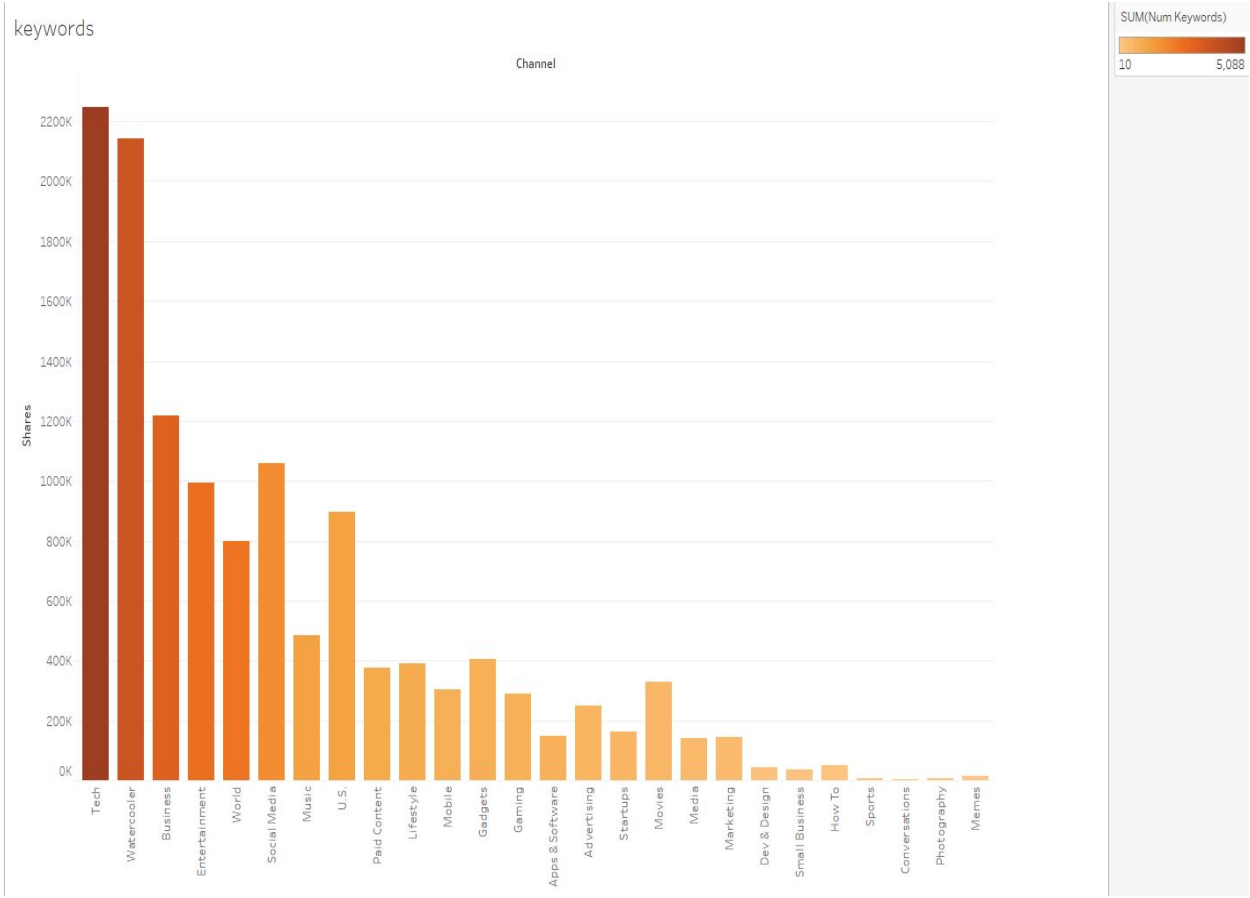
This graph describes every topic's popularity sorted by the date, we can see that Tech is the most popular topic in Monday, Wednesday, Friday, and the Watercooler is popular in Sunday, Tuesday, Thursday. the lowest 2 topics is conversations and Dev&Design topics.



Then, we find that the top 2 topics in a week expect Saturday is Tech and WaterCooler, so we want to find why they are popular in most of days. Because those are all news, so we can analyze the author according to these news to find out the reason.



We also find that the top2 news topic have more number of links and keywords in the news. Because more links will give readers more choice, readers can read more related article following their own hearts. Putting more keywords rather than meaningful words will save readers time.



Here, we sort all the topics in the dataset by the number of keywords, To increase the number of shares, companies need add more keywords into the Tech, watercooler, and Business topics' articles

Conclusion

Classification:

- Designed a model using C5.0 algorithm with 80:20 random sampling split for training and testing respectively.
- Handled feature-engineering methodology to convert the continuous to category target label, which was challenging and interesting.
- Even though, model's overall accuracy is less (41 %), other performance measures show that the model is performing better. Model is performing above the diagonal line in ROC curve. However, in future studies, proper pruning will improve the model's performance even better.

Regression:

- Employed three ways of regression, including linear regression, PCA regression, and Ridge and LASSO Regression to find a more precise model for predicting online news popularity.
- The results indicate that our best model is coming from Ridge and Lasso Regression since the mean square error of the prediction is 121612240.

Future work:

Based on our results, we recommend that future studies focus on refining the model by including more independent variables, extending the time interval, currently we only collected data for 2 years. We also suggest that future studies explore what factors influence news with particular topic.

Reference

Anon, (2017). [online] Available at: https://www.linkedin.com/pulse/online-news-popularity-trend-analysis-krunal-khatri/.

Archive.ics.uci.edu. (2017). UCI Machine Learning Repository: Online News Popularity Data Set. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity# .

He Ren and Quan Yang, ‘Predicting and Evaluating the Popularity of Online News’, Standford University Machine Learning Report, 2015.

K. Fernandes, P. Vinagre and P. Cortez. A Proactive In-telligent Decision Support System for Predicting the Pop-ularity of Online News. Proceedings of the 17th EPIA 2015- Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

Thanks!