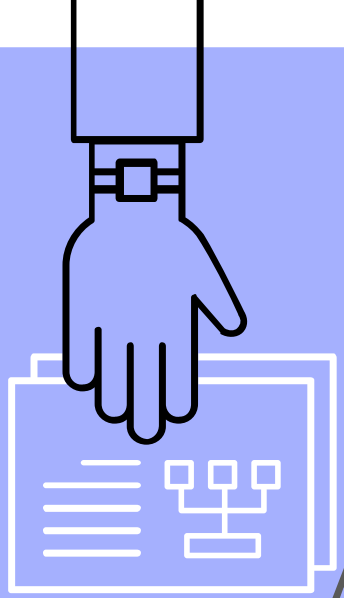
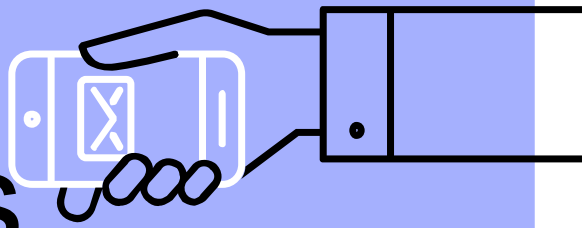
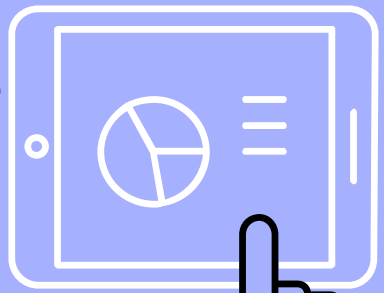
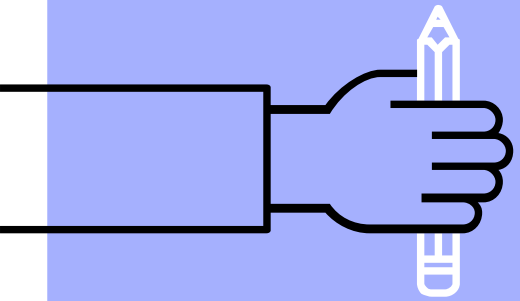




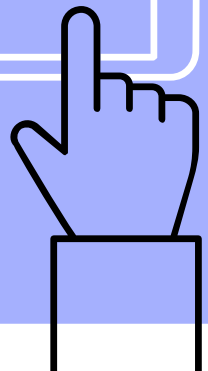
Online News Popularity

A Regression to Predict Views





W207: Applied Machine Learning
Alice Hua - Haley Farber - Derrick Xiong

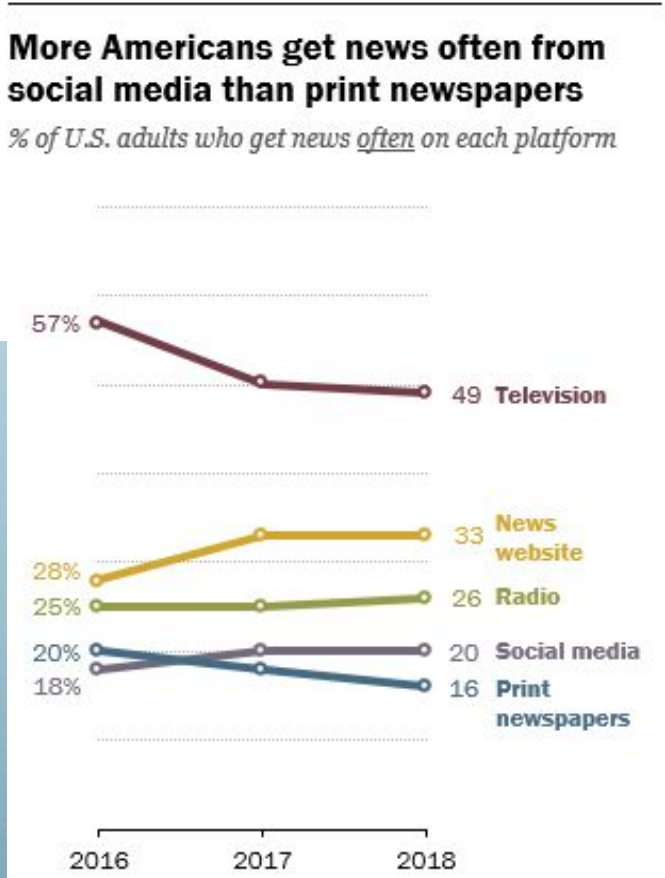


Motivation

Predict the popularity of news prior to its publication

Source: https://medium.com/journalism-trends-technologies/pay-models-for-online-news-8ea87d46a3c7

Source: https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/



Source: https://neilpatel.com/blog/social-proof/

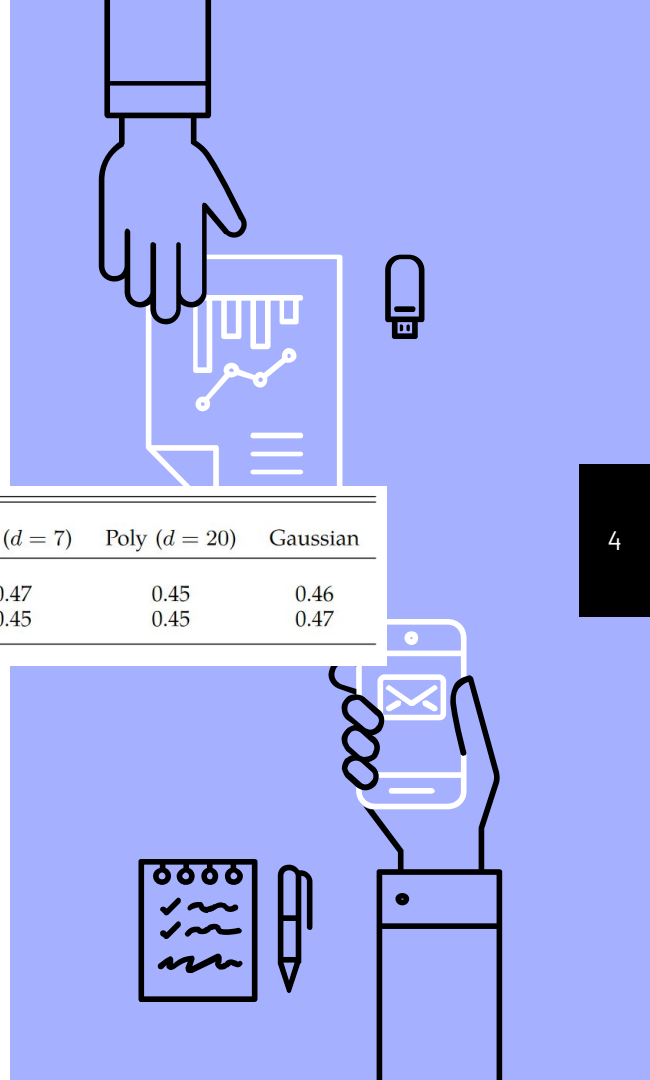
Literature

2015 [Predicting and Evaluating the Popularity of Online News]

TABLE IV. PERFORMANCE OF DIFFERENT ALGORITHMS

Algorithms	Accuracy	Recall
Linear Regression	0.66	0.67
Logistic Regression	0.66	0.70
SVM ($d = 9$ Poly Kernel)	0.55	0.45
Random Forest (500 Trees)	**0.69**	**0.71**
k-Nearest Neighbors ($k = 5$)	0.56	0.47
SVR (Linear Kernel)	0.52	0.59
REPTree	0.67	0.62
Kernel Partial Least Square	0.58	0.60
Kernel Perceptron (Max loop 100)	0.45	0.99
C4.5 Algorithm	0.58	0.59

Kernel	Linear	Poly ($d = 7$)	Poly ($d = 20$)	Gaussian
Test Error	0.48	0.47	0.45	0.46
Training Error	0.45	0.45	0.45	0.47



Literature

2017 Genetic Algorithm Based Correlation Enhanced Prediction of Online News Popularity

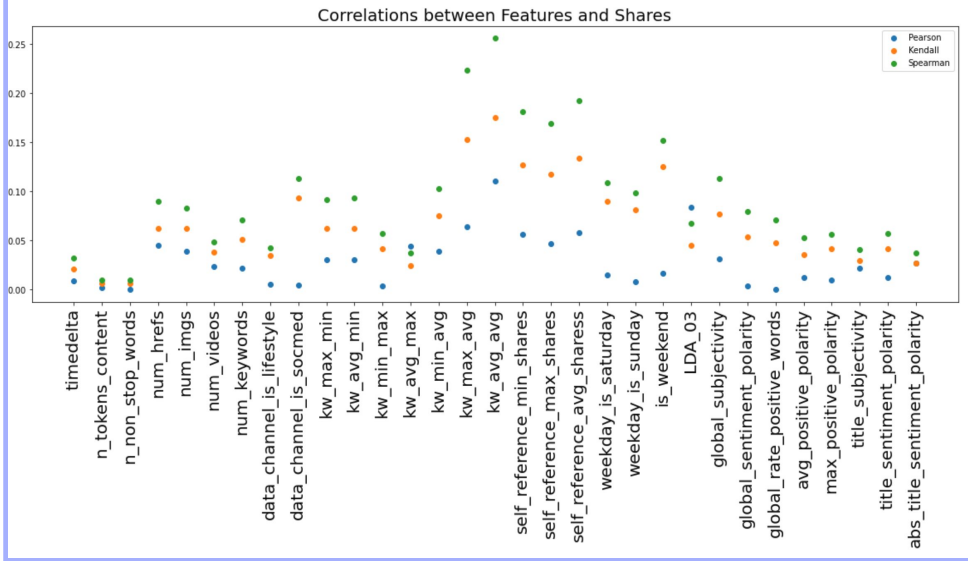


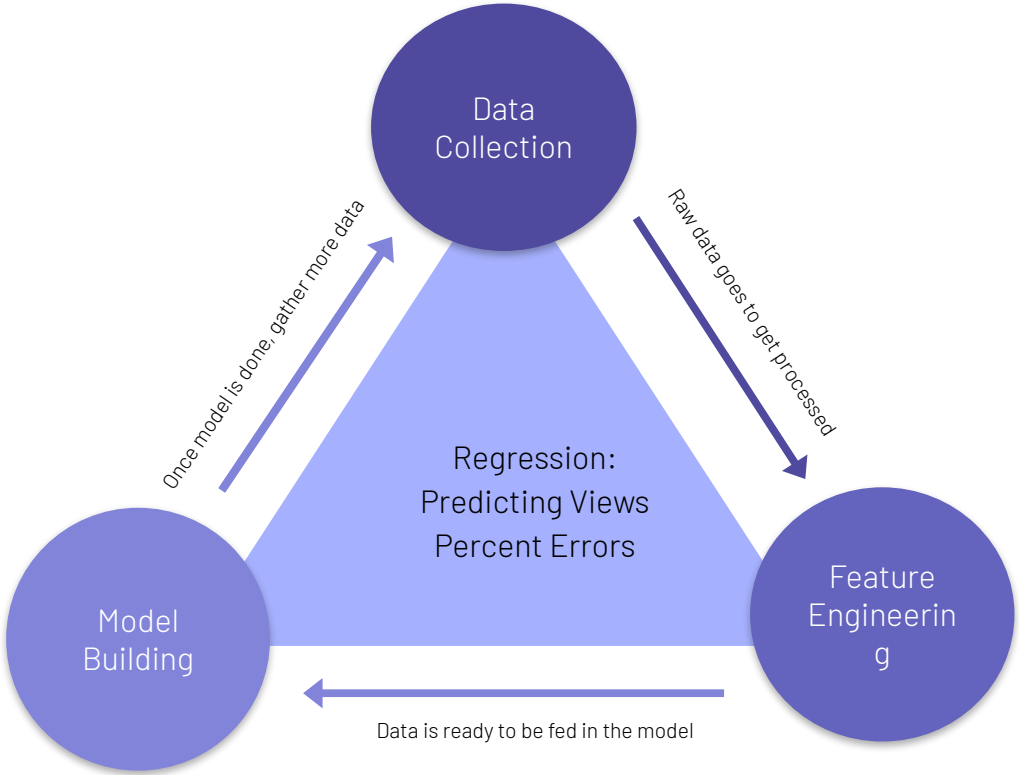
Table 2 Genetic results

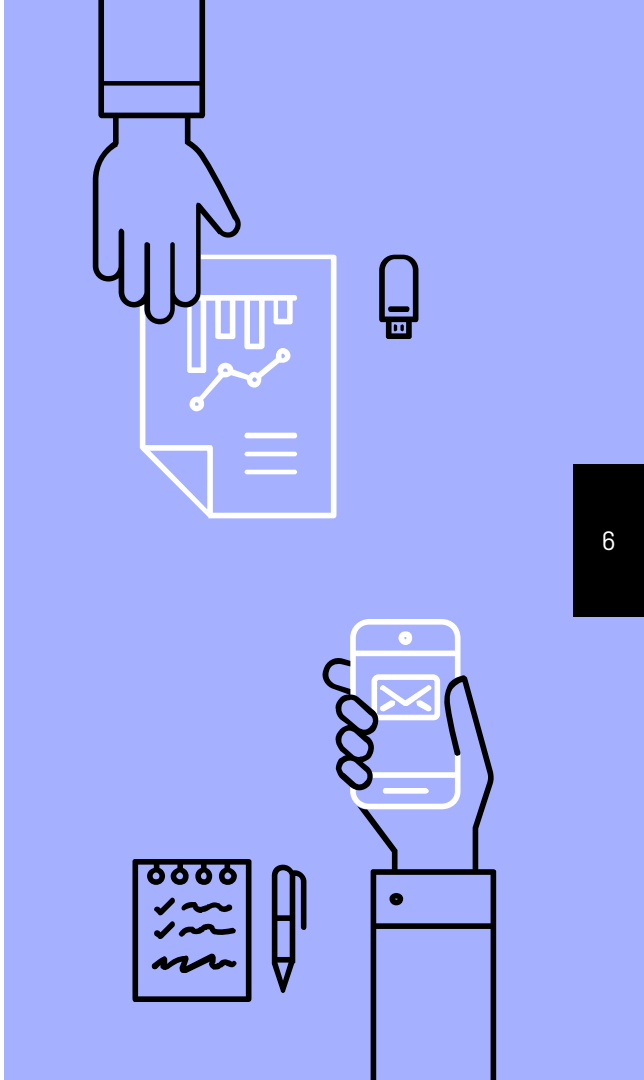
Attributes		
timedelta	num_imgs	num_videos
num_keywords	data_channel_is_tech	kw_min_min
kw_min_avg	self_reference_min_shares	self_reference_max_shares
self_reference_avg_shares	is_weekend	LDA_03
global_sentiment_polarity	global_rate_positive_words	avg_positive_polarity
title_subjectivity	title_sentiment_polarity	abs_title_sentiment_polarity

Table 3 Prediction test

Techniques	32 attributes (in %)	18 attributes (in %)
Naíve bayes	**93.46**	76.06
Random forest	79.63	79.62
Neural networks	91.89	**91.96**
KNN (k = 5)	90.27	74.49
MLR	61.11	60.47
C4.5 tree	88.8281	84.7947
C5.0 tree	79.92	79.62

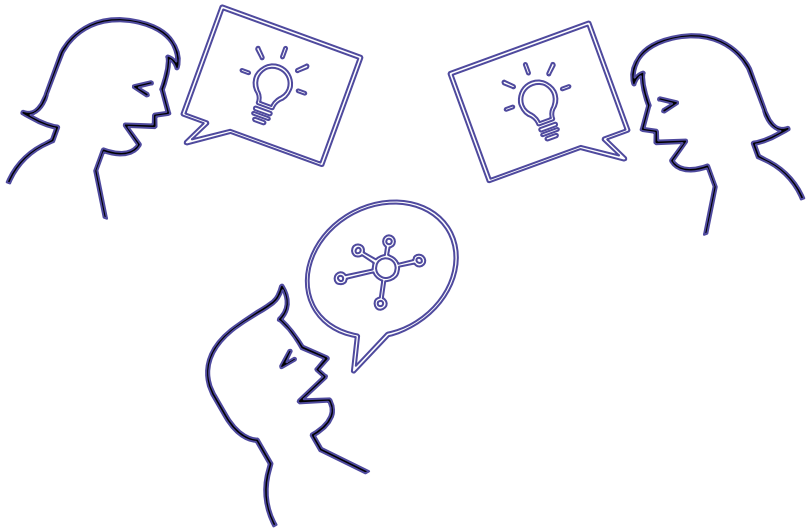
Project workflow



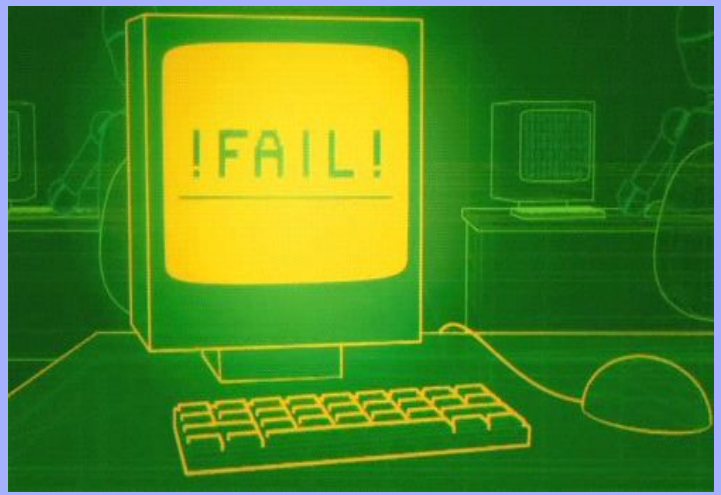


Initial Idea & Final Decision

Mashable → Forbes



Data Collection



Total articles scraped: 7k

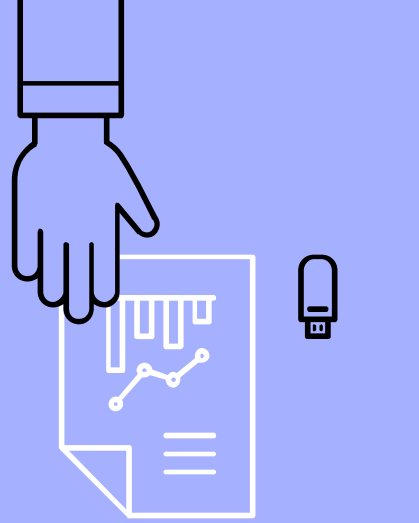
Feature Engineering

Leadership	...	avg_negative_polarity	min_negative_polarity	max_negative_polarity	LDA_00	LDA_01	LDA_02	LDA_03	LDA_04	timedelta	num_keywords
0	...	-0.252778	-0.900000	-0.100	0.997484	0.000629	0.000629	0.000631	0.000628	4	316
0	...	-0.220726	-0.600000	-0.050	0.000195	0.760765	0.000196	0.238648	0.000196	193	1069
0	...	-0.220483	-								
0	...	-0.293182	-								
0	...	-0.207833	-								

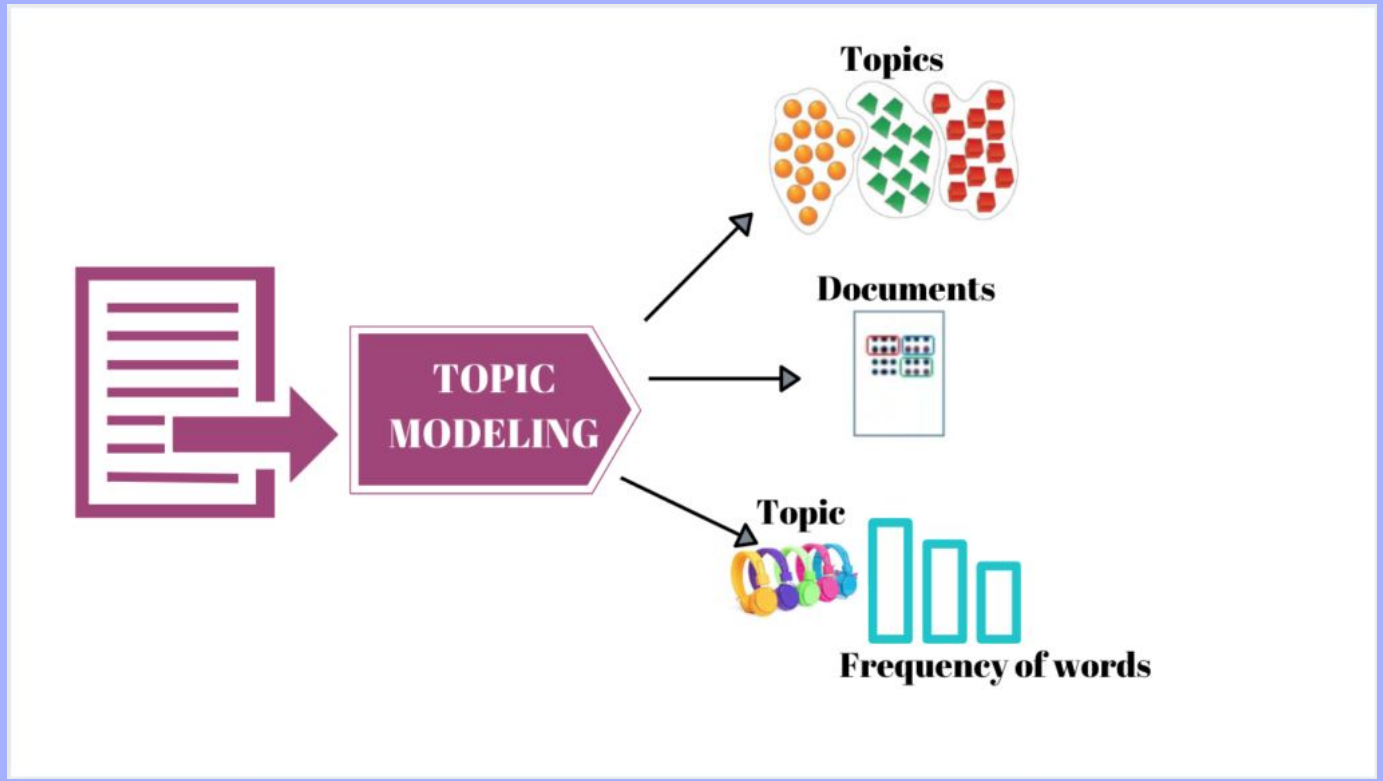
65

Features Added:

1. Innovation: dummy variable for articles in the innovation topic
2. Leadership: dummy variable for articles in leadership topic
3. Lifestyle: dummy variable for article in lifestyle topic
4. Money: dummy variable for article in money topic
5. month: month article was published
6. Month dummies (11 total) - Jan, Feb, Mar, Apr, May, Jun, July, Aug, Sep, Oct, Nov
7. n_tokens_title: Number of words in the title
8. n_tokens_content: Number of words in the article
9. n_unique_tokens: Percent of unique words in the article
10. average_token_length: Average length of the words in the content
11. n_non_stop_words: Percent of non-stop words in the article
12. n_non_stop_unique_tokens: Percent of unique non-stop words in the article
13. day_of_week: day of the week the article was published



LDA Analysis: Latent Dirichlet Algorithm



LDA Explanation

1. Preprocessing

```
0    [bitcoin, ethereum, rippl, bitcoin, cash, lite...
1    [sourdough, loav, plus, creativ, use, discard,...
2    [addit, offer, deliveri, curbsid, pick, regula...
3    [hyundai, luxuri, brand, genesi, strong, start...
4    [meet, genesi, automot, market, go, past, year...
5    [meet, lexus, lexus, engin, makeov, lineup, ca...
6    [meet, toyota, sienna, hybrid, brand, vehicl, ...
7    [adjust, menus, smaller, gather, drop, dish, l...
8    [startup, fail, dont, viabl, nich, reli, wrong...
9    [local, busi, go, duti, serv, healthcar, hero,...
```

2. Master Dictionary

```
0 add
1 altern
2 announc
3 asia
4 asset
5 assist
6 attent
7 author
8 ayyar
9 bank
10 barrier
```

3. Individual Dictionaries

```
Word 0 ("add") appears 2 time.
Word 1 ("altern") appears 1 time.
Word 2 ("announc") appears 1 time.
Word 3 ("asia") appears 1 time.
Word 4 ("asset") appears 1 time.
Word 5 ("assist") appears 1 time.
Word 6 ("attent") appears 1 time.
Word 7 ("author") appears 1 time.
Word 8 ("bank") appears 1 time.
Word 9 ("barrier") appears 1 time.
```

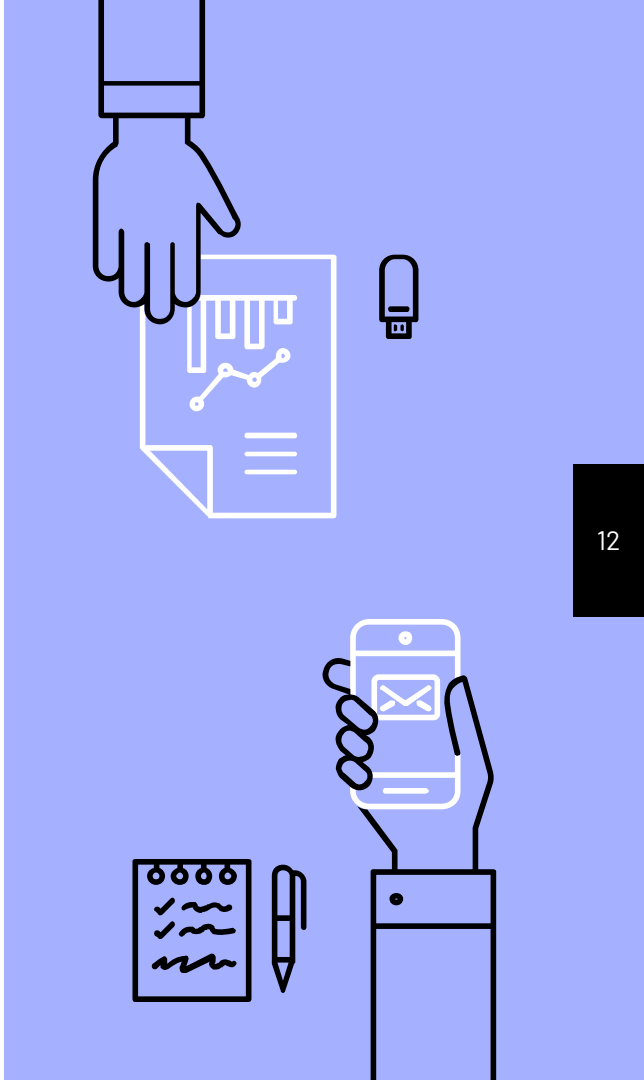
```
[(0, 2),
 (1, 1),
 (2, 1),
 (3, 1),
 (4, 1),
 (5, 1),
 (6, 1),
 (7, 1),
 (8, 1),
 (9, 1),
```

4. Probabilities for Each Topic

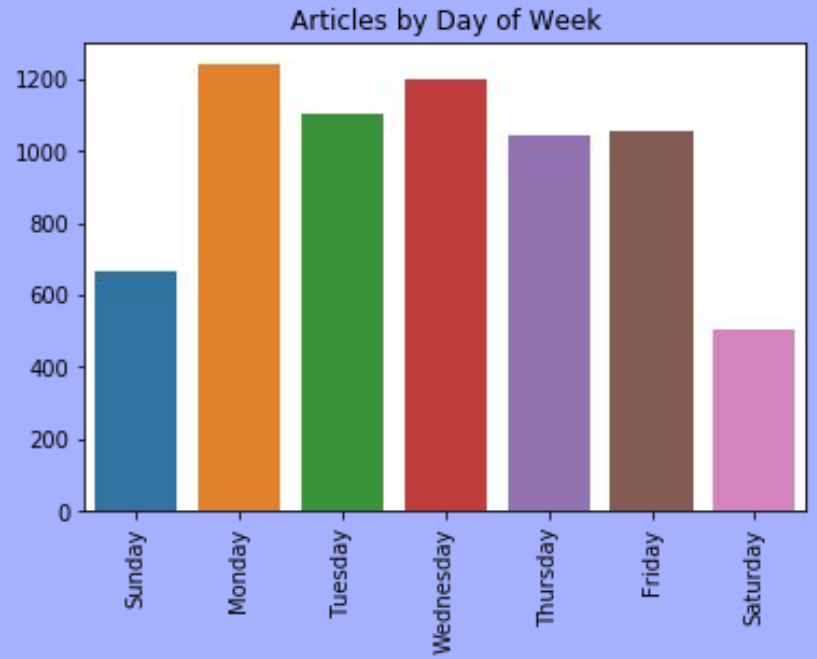
LDA_00	LDA_01	LDA_02	LDA_03	LDA_04
0.043497	0.000628	0.000637	0.059345	0.895893
0.999218	0.000196	0.000195	0.000196	0.000195
0.999199	0.000200	0.000200	0.000200	0.000201

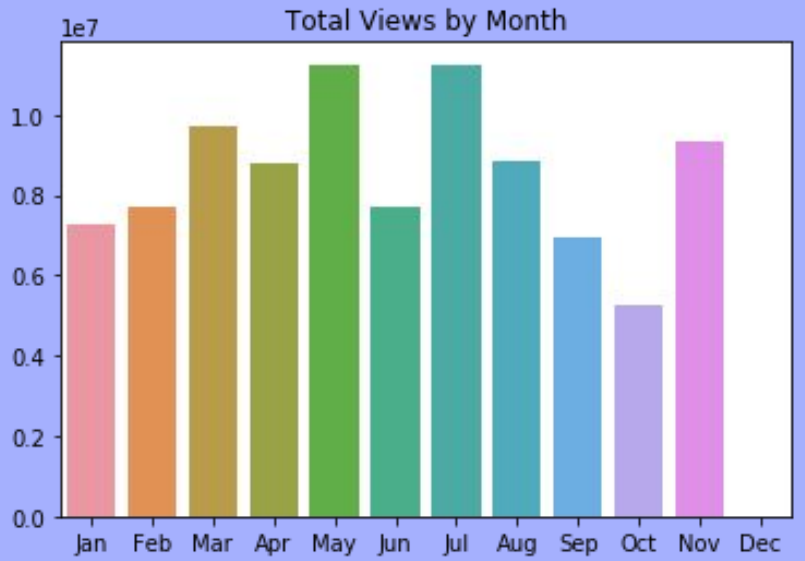
Keyword feature

	kw_min_min	kw_min_max	kw_min_avg	kw_max_min	kw_max_max	kw_max_avg	kw_avg_min	kw_avg_max	kw_avg_avg
0	33.0	47503.0	7025.750000	33.0	11524072.0	9.725148e+05	41.0	1859311.0	27658.245020
1	60.0	2826.0	964.800000	30.0	12490200.0	2.226071e+05	63.0	41856.0	5524.631579
2	137.0	4064.0	1397.833333	84.0	8418600.0	1.778267e+05	162.0	70403.0	7310.030303
3	738.0	4264.0	1983.333333	30.0	12490200.0	1.054787e+05	2926.0	19966.0	8524.333333
4	738.0	738.0	738.000000	738.0	1859311.0	3.048802e+05	24.0	551511.0	11766.080460

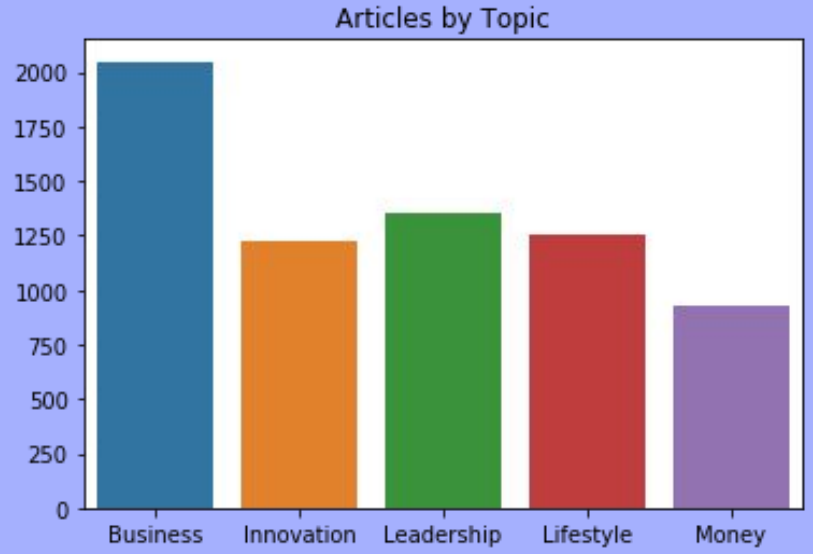


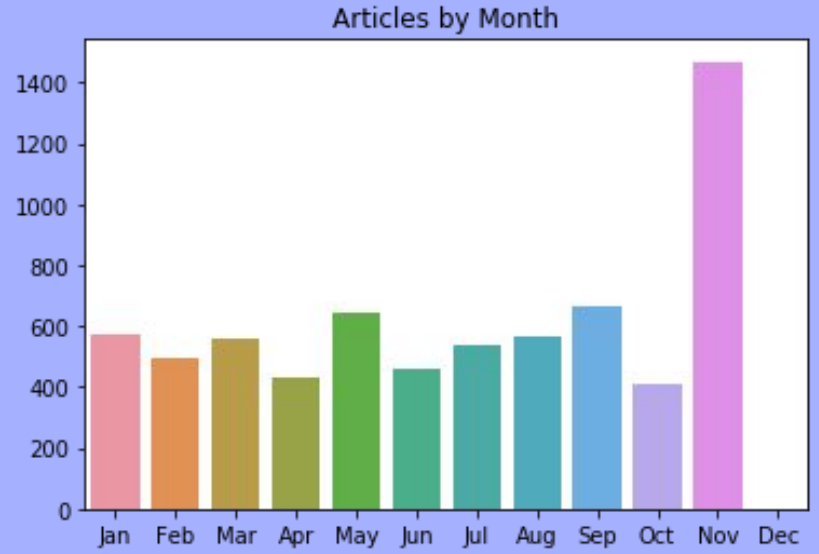
EDA


Articles by Day of Week
1200
1000
800
600
400
200
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

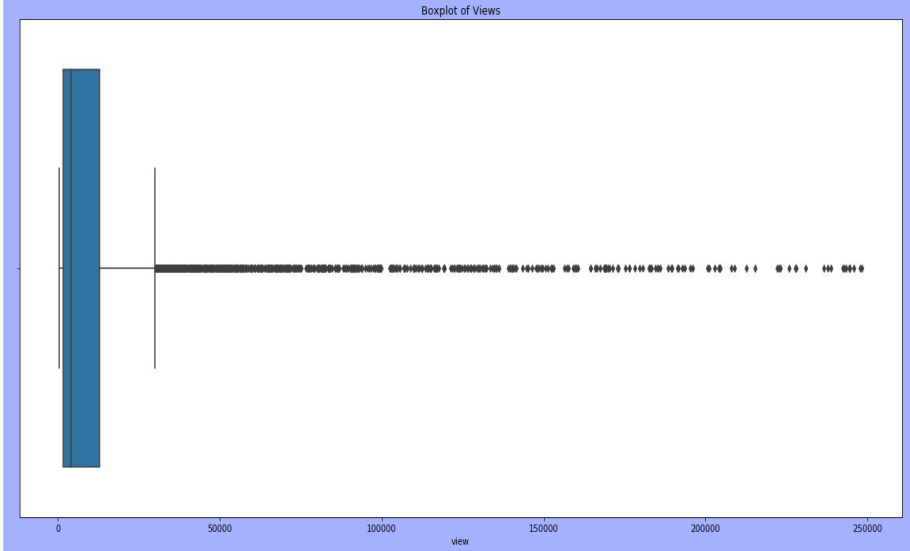

Total Views by Month
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

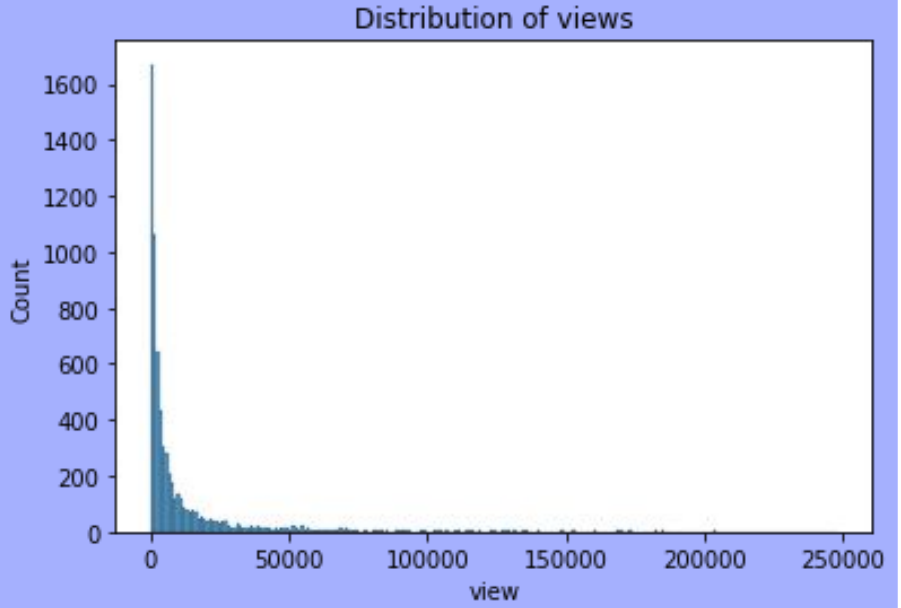
EDA


Articles by Topic
2000
1750
1500
1250
1000
750
500
250
0
Business
Innovation
Leadership
Lifestyle
Money


Articles by Month
1400
1200
1000
800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

EDA


Boxplot of Views
0
50000
100000
150000
200000
250000
view


Distribution of views
Count
1600
1400
1200
1000
800
600
400
200
0
0
50000
100000
150000
200000
250000
view

Model

OLS Model trained on articles from Jan to March and predicted on April

```
In [333]: pred
```

Out[333]:

	true_view	predict_view	pct_error
2	4064	-6.719910e+08	-1.653531e+07
22	3540	-6.720025e+08	-1.898322e+07
82	1645	-6.719964e+08	-4.085095e+07
114	662	-6.720035e+08	-1.015112e+08
148	1684	-6.719961e+08	-3.990486e+07
...	...	...	...
5694	512	-6.719787e+08	-1.312459e+08
5695	3152	-6.719843e+08	-2.131940e+07
5697	3097	-6.719789e+08	-2.169784e+07
5720	6817	-6.719607e+08	-9.857232e+06
5738	2054	-6.719736e+08	-3.271547e+07

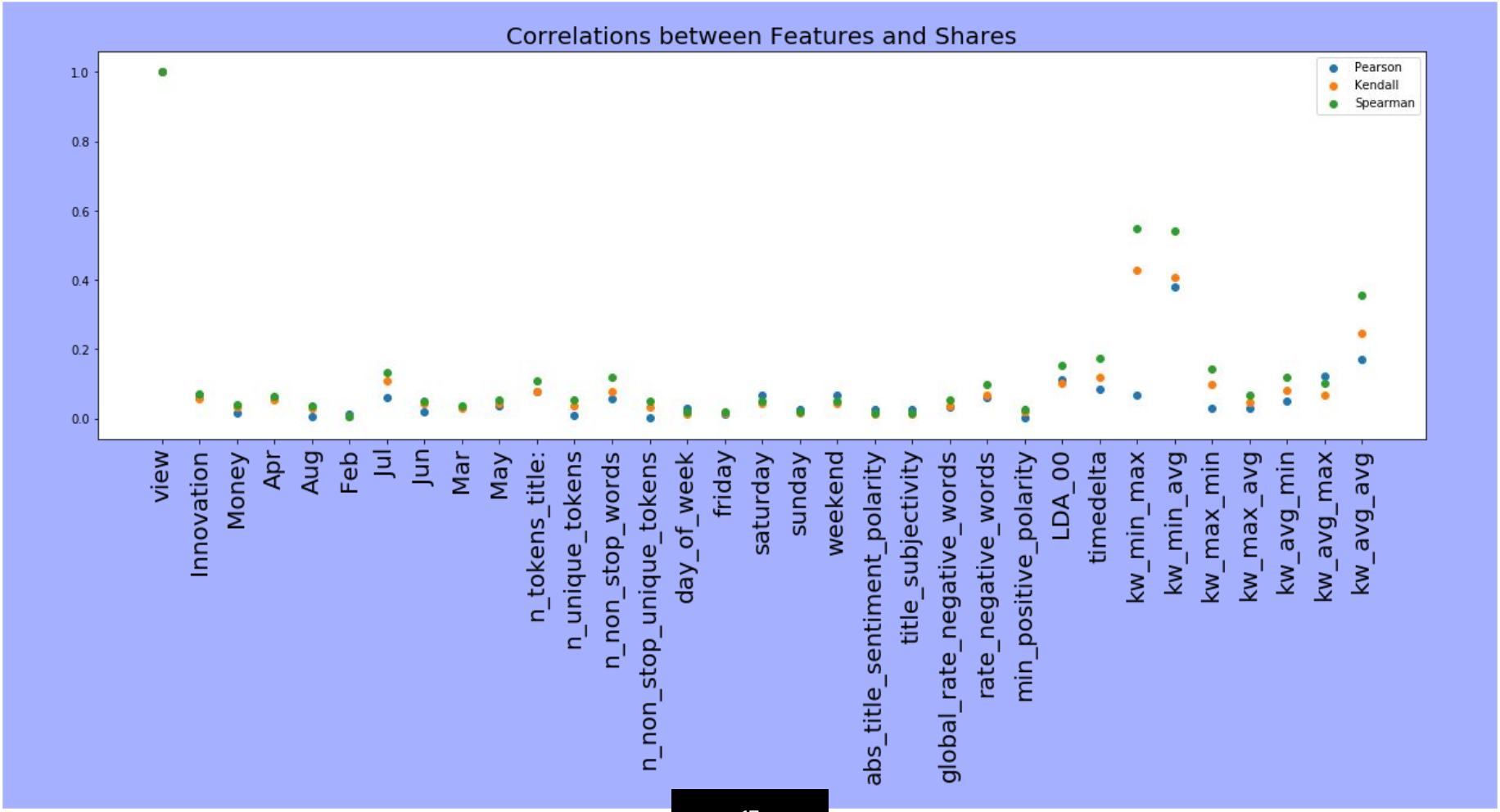
408 rows × 3 columns

OLS Model trained on 7 different rolling windows of 3 months period

```
In [397]: result
```

Out[397]:

	Train_month	Test_month	Average_pct_error
0	Jan, Feb, Mar	Apr	-4.078043e+07
1	Feb, Mar, Apr	May	-2.512117e+08
2	Mar, Apr, May	Jun	-2.077185e+08
3	Apr, May, Jun	Jul	-1.620805e+08
4	May, Jun, Jul	Aug	-2.158531e+08
5	Jun, Jul, Aug	Sep	2.129441e+07
6	Jul, Aug, Sep	Oct	-7.432516e+07
7	Aug, Sep, Oct	Nov	-6.090499e+08


Correlations between Features and Shares
Pearson
Kendall
Spearman
1.0
0.8
0.6
0.4
0.2
0.0
view
Innovation
Money
Apr
Aug
Feb
Jul
Jun
Mar
May
n_tokens_title:
n_unique_tokens
n_non_stop_words
n_non_stop_unique_tokens
day_of_week
friday
saturday
sunday
weekend
abs_title_sentiment_polarity
title_subjectivity
global_rate_negative_words
rate_negative_words
min_positive_polarity
LDA_00
timedelta
kw_min_max
kw_min_avg
kw_max_min
kw_max_avg
kw_avg_min
kw_avg_max
kw_avg_avg

Improved Model

OLS Model trained on articles from Jan to March and predicted on April with selected features

```
In [345]: pred
```

Out[345]:

	true_view	predict_view	pct_error
2	4064	2268.734393	-44.174843
22	3540	1361.347454	-61.543857
82	1645	-4532.256416	-375.517107
114	662	-4162.322176	-728.749573
148	1684	-707.909079	-142.037356
...	...	...	...
5694	512	17320.722842	3282.953680
5695	3152	14174.035234	349.683859
5697	3097	17563.426713	467.110969
5720	6817	33285.688613	388.274734
5738	2054	25485.580065	1140.777997

408 rows × 3 columns

OLS Model trained on 7 different rolling windows of 3 months period with selected features

```
In [353]: result
```

Out[353]:

	Train_month	Test_month	Average_pct_error
0	Jan, Feb, Mar	Apr	269.543188
1	Feb, Mar, Apr	May	380.486238
2	Mar, Apr, May	Jun	1193.665884
3	Apr, May, Jun	Jul	1125.715705
4	May, Jun, Jul	Aug	1713.486585
5	Jun, Jul, Aug	Sep	1587.295163
6	Jul, Aug, Sep	Oct	70.649050
7	Aug, Sep, Oct	Nov	316.077906

More Models

Out[354]:

	Train_month	Test_month	Average_pct_error_OLS	Average_pct_error_Ridge	Average_pct_error_RANSAC
0	Jan, Feb, Mar	Apr	319.729076	356.004239	1392.898791
1	Feb, Mar, Apr	May	401.005362	368.403006	398.307972
2	Mar, Apr, May	Jun	1193.665884	298.099140	217.026743
3	Apr, May, Jun	Jul	1125.715705	242.256164	396.768366
4	May, Jun, Jul	Aug	1713.486585	347.450707	171.833617
5	Jun, Jul, Aug	Sep	1587.295163	347.152479	107.398336
6	Jul, Aug, Sep	Oct	167.605017	127.312918	1353.532792
7	Aug, Sep, Oct	Nov	354.960497	305.893875	599.146119

More Model Improvement

	Train_month	Test_month	Average_pct_error_OLS	Average_pct_error_Ridge	Average_pct_error_RANSAC
0	Jan, Feb, Mar	Apr	477.185388	466.504964	62.594194
1	Feb, Mar, Apr	May	487.871928	454.136389	163.364450
2	Mar, Apr, May	Jun	461.134148	422.646195	20.804344
3	Apr, May, Jun	Jul	374.704348	347.975590	31.416039
4	May, Jun, Jul	Aug	439.161979	415.720109	183.568721
5	Jun, Jul, Aug	Sep	696.576601	692.576029	246.302175
6	Jul, Aug, Sep	Oct	500.632742	497.333328	66.804967
7	Aug, Sep, Oct	Nov	515.204290	535.530223	134.833123

6 Features

'LDA_00', 'n_tokens_title:', 'rate_negative_words','kw_min_avg','kw_avg_max', 'kw_avg_avg'

	Train_month	Test_month	Average_pct_error_OLS	Average_pct_error_Ridge	Average_pct_error_RANSAC
0	Jan, Feb, Mar, Apr, May, Jun	Jul	357.296020	340.065529	12.765784
1	Feb, Mar, Apr, May, Jun, Jul	Aug	409.791899	386.235687	14.712310
2	Mar, Apr, May, Jun, Jul, Aug	Sep	666.568385	634.645327	89.812642
3	Apr, May, Jun, Jul, Aug, Sep	Oct	523.779937	498.377277	47.477806
4	May, Jun, Jul, Aug, Sep, Oct	Nov	765.292144	727.014411	94.771695

Limitations

Limitations:

- Too many data points in November, and thus not enough data to train
- The topics are not evenly distributed.
- Given the distribution of the data, regression might not be the best model for the job. We found because our data has a lot of variance as well as a fat-tailed distribution, regression cannot fully capture that, a neural network with more data to train can potentially improve the results more.

BIG CONCEPT

In conclusion, we realize, after many iterations of models, that with given the distribution of our data, special care needs to be taken in handling the outlier and high variance in our data. Our regression model provides a good baseline for the prediction of views, but better models could be implemented with more data. Moreover, our data collection and feature engineering process could be generalized to collect more data from many different sources other than Forbes to create more diverse and robust datasets for future research endeavor beyond the prediction of views and popularity.

THANKS!

Any questions?