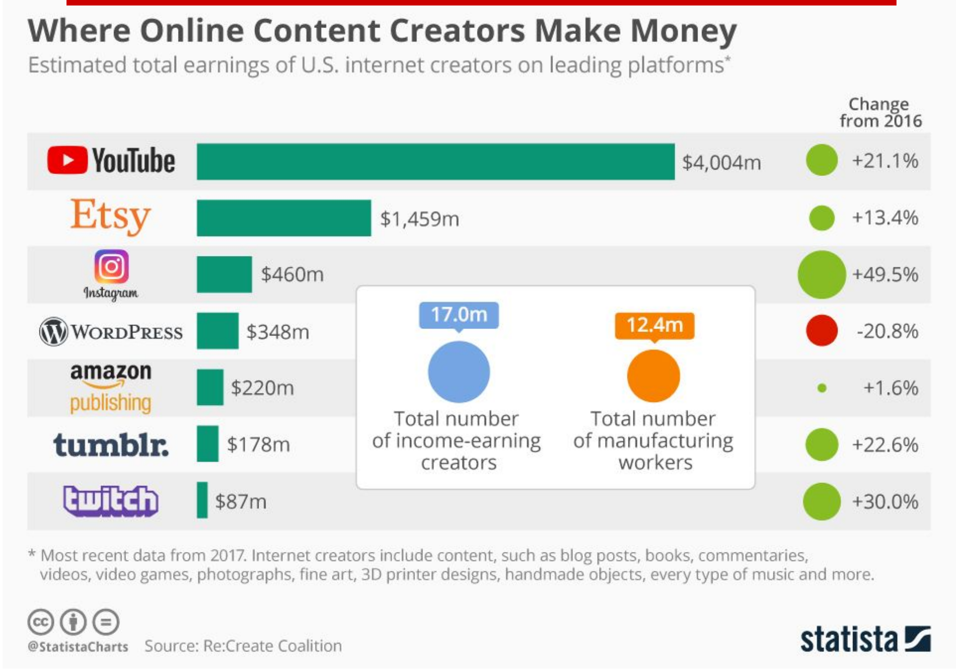
Analysis of Trending YouTube Videos

Presented by
Kuang (Steven) Li
Linxia Liu
Teng-Yun (Jacob) Chung
Vivian Kang
Yuan Liu
Zelong Qian

December 2019

Introduction & Key Question

Publishing videos on Youtube is highly profitable



Trending Videos

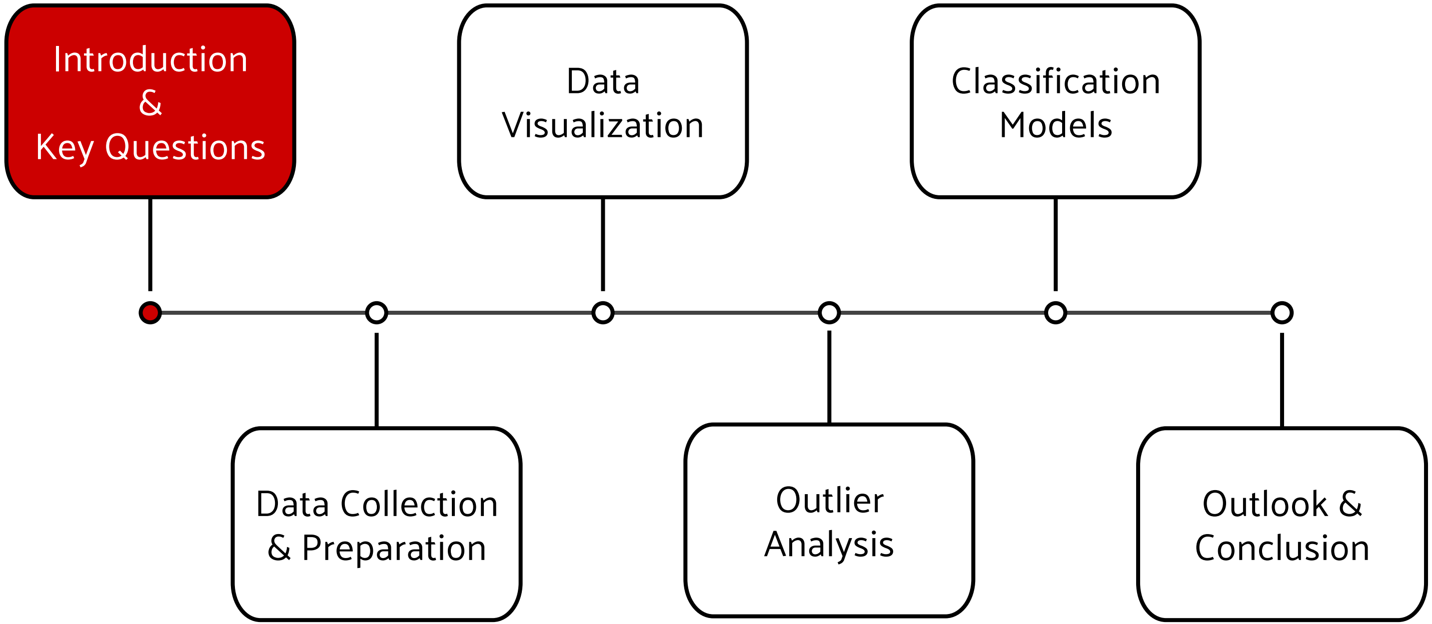
Trending aims to surface videos that:

- Are appealing to a wide range of viewers
- Are not misleading, clickbaity, or sensational
- Capture the breadth of what's happening on YouTube and in the world
- Showcase a diversity of creators
- Ideally, are surprising or novel

Key question:

- What factors make an immediately trending video?

Presentation Outline



Prepare the Dataset

Attributes	Missing Values
video_id	0
trending_date	0
title	0
channel_title	0
category_id	0
publish_time	0
tags	0
views	0
likes	0
dislikes	0
comment_count	0
thumbnail_link	0
comments_disabled	0
ratings_disabled	0
video_error_or_removed	0
description	570

16 attributes, 40949 observations

Key step: constructing dependent variable

1. Create a new variable “pre-trending time ”

pre-trending time = trending date - publish time

2. Keeps the first pre-trending time of each video

3. Remove the upper outliers of pre-trending time via IQR

count	6334.000000
mean	22.041048
std	205.918572
min	0.000000
25%	1.000000
50%	2.000000
75%	3.000000
max	4215.000000

```
def only_outlier(cleaned, col_name):
    q1 = cleaned['pretrending_time'].quantile(0.25)
    q3 = cleaned['pretrending_time'].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    only_out = cleaned.loc[(cleaned['pretrending_time'] > fence_high)]
    return only_out
```

```
def remove_outlier(cleaned, col_name):
    q1 = cleaned['pretrending_time'].quantile(0.25)
    q3 = cleaned['pretrending_time'].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    cleaned_out = cleaned.loc[(cleaned['pretrending_time'] < fence_high)]
    return cleaned_out
```

Source: Trending YouTube Video Statistics. Retrieved from https://www.kaggle.com/datasnaek/youtube-new

Prepare the Dataset

Attributes	Missing Values
video_id	0
trending_date	0
title	0
channel_title	0
category_id	0
publish_time	0
tags	0
views	0
likes	0
dislikes	0
comment_count	0
thumbnail_link	0
comments_disabled	0
ratings_disabled	0
video_error_or_removed	0
description	570

16 attributes, 40949 observations

Key step: constructing dependent variable

1. Create a new variable "pre-trending time "

   pre-trending time = trending date - publish time

2. Keeps the first pre-trending time of each video

3. Remove the upper outliers of pre-trending time via IQR

4. Create a new variable "immediate trending"

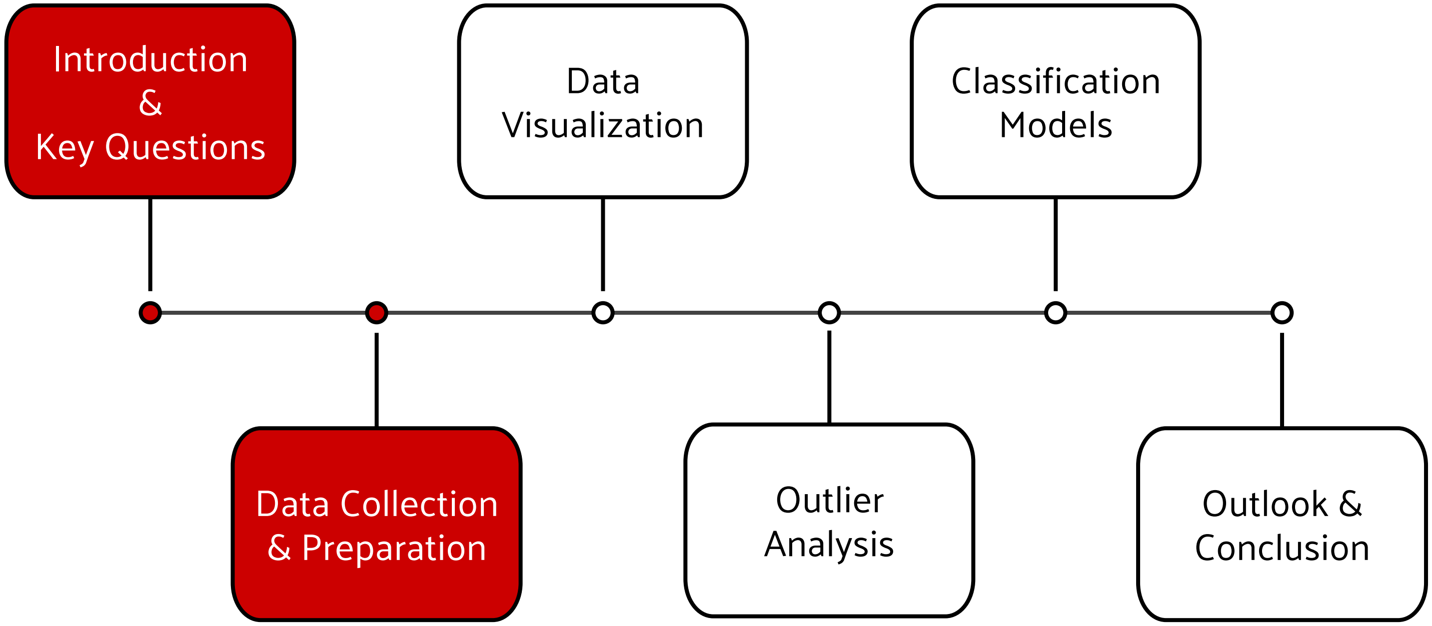
   pre-trending time average = 1.86

   pre-trending time <= 1.86 ⟶ immediate trending = 1

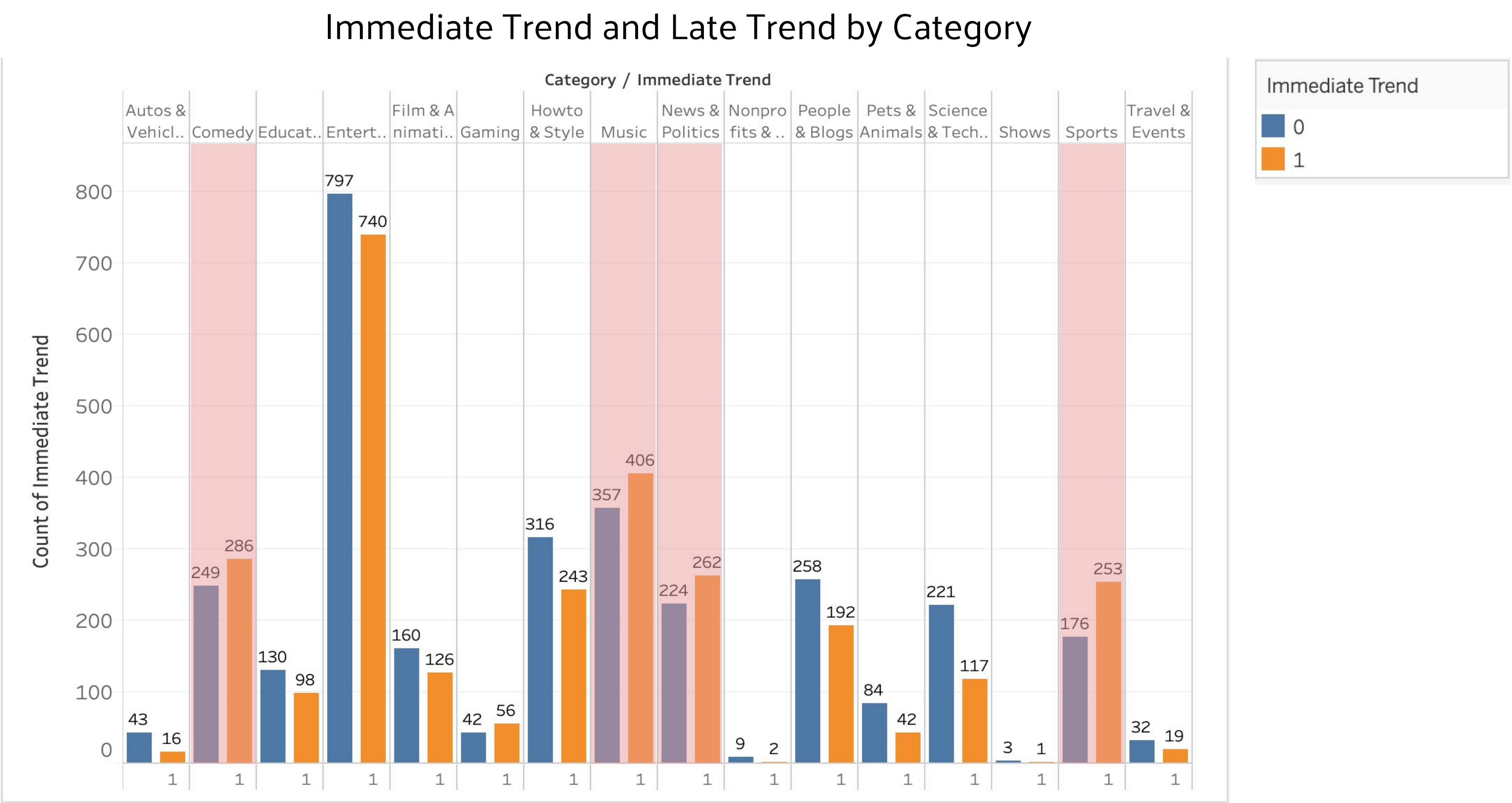
   pre-trending time > 1.86 ⟶ immediate trending = 0

Source: Trending YouTube Video Statistics. Retrieved from https://www.kaggle.com/datasnaek/youtube-new

Presentation Outline

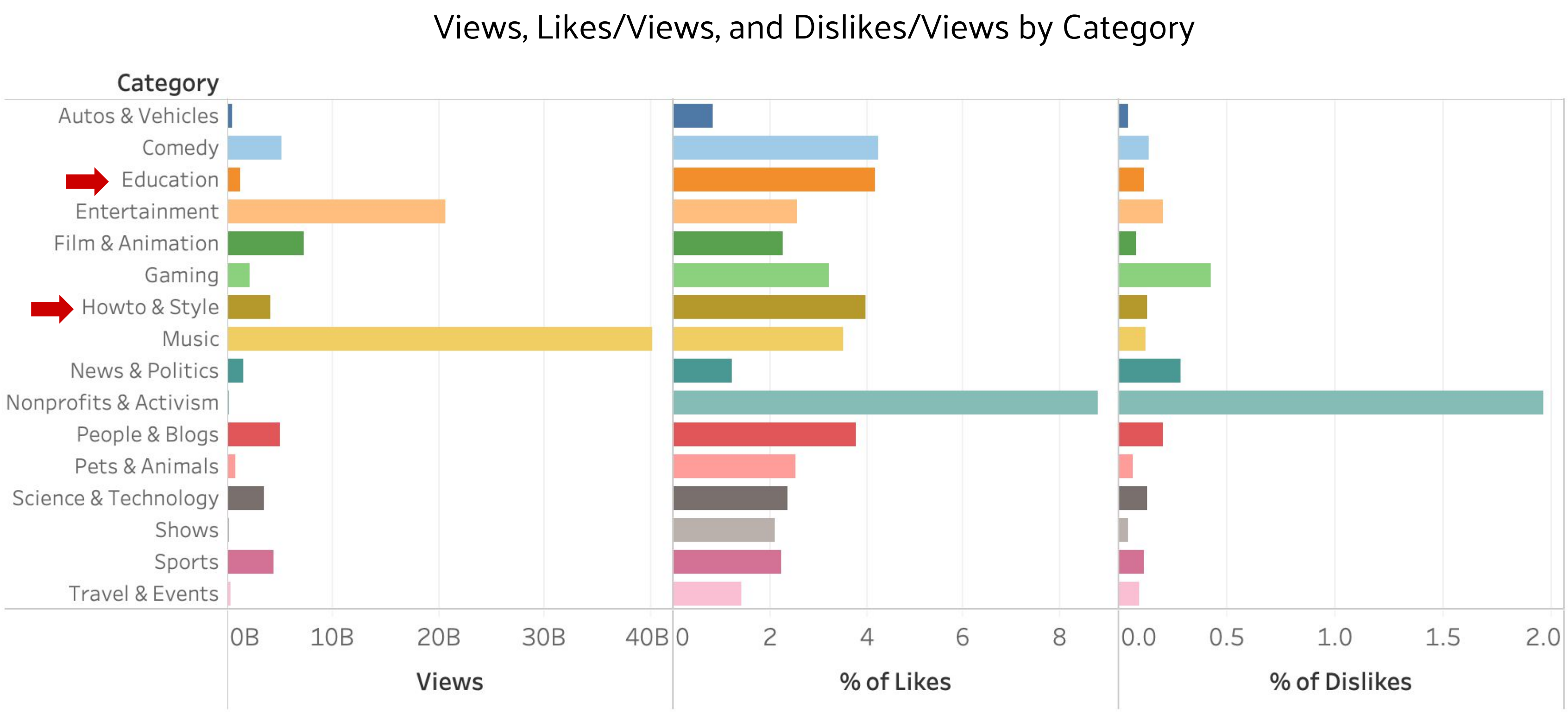


Compare Immediate Trend and Late Trend by Category



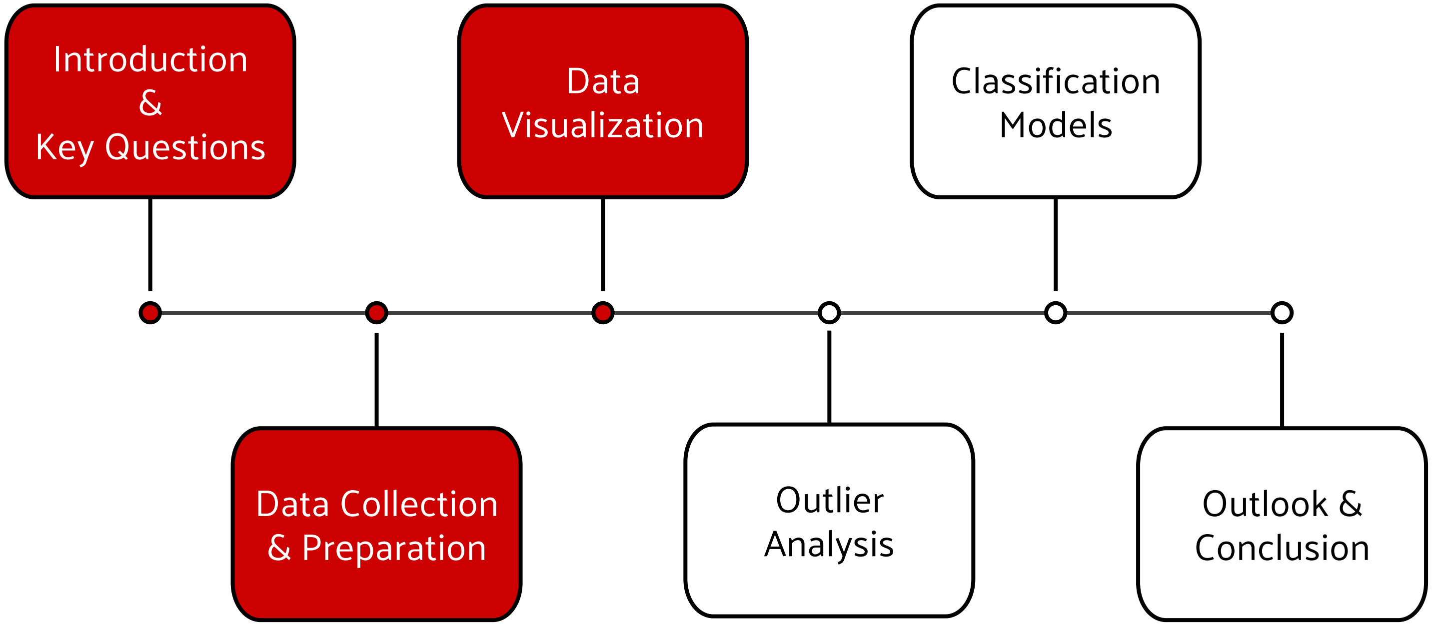
Takeaway: publish current events videos to be immediately trending

Compare Views, Likes, Dislikes by Category

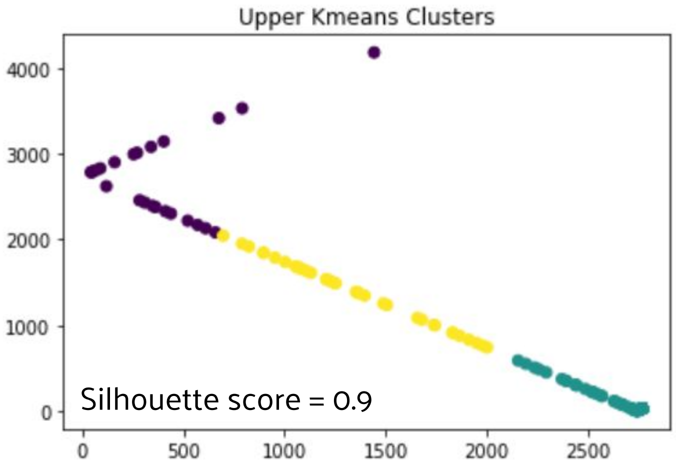


Takeaway: explore opportunities in Education and Howto & Style categories

Presentation Outline



Study the Upper Outliers via Clustering



Cluster 0

video_id	views	likes	dislikes	comment_count	pretrending_time	cluster_labels
-37nlo_tLnk	2863.0	2.0	0.0	0.0	2933.0	0
2vQ_fnlvvr8	45096.0	287.0	9.0	59.0	3113.0	0
5WUDfviiKRE	16823.0	93.0	275.0	172.0	2463.0	0
76c_nxhuVdM	163780.0	826.0	10.0	167.0	2839.0	0
9L4-DV1nVek	25796.0	64.0	4.0	22.0	2823.0	0
K9kVsnTQh-g	73685.0	260.0	55.0	96.0	2816.0	0
MJO3FmmFuh4	258506.0	459.0	152.0	82.0	4215.0	0
P2I7hQHOqNI	2896.0	30.0	0.0	3.0	2489.0	0
Tn5OBFgiExQ	51984.0	215.0	8.0	31.0	2163.0	0
UQtt9I6c-YM	49942.0	46.0	6.0	26.0	3563.0	0

Cluster 1

video_id	views	likes	dislikes	comment_count	pretrending_time	cluster_labels
-BQJo3vK8O8	48431654.0	609101.0	52259.0	29172.0	6.0	1
-CS84oCtjvc	994662.0	21094.0	714.0	3212.0	6.0	1
-QR-TB_k20M	23877.0	93.0	30.0	20.0	6.0	1
-hg_VRwS5RI	110470.0	7366.0	69.0	1247.0	7.0	1
-t1q78GYNww	127481.0	4865.0	234.0	1117.0	6.0	1
...	...	...	...	...	...	...
yFRPhi0jhGc	635806.0	35790.0	864.0	2857.0	6.0	1
z5JQMBcVFns	161532.0	1272.0	45.0	54.0	7.0	1
zYWt2mnalP8	160477.0	8388.0	691.0	950.0	6.0	1
zbV1zyg_4qU	5965.0	186.0	8.0	52.0	26.0	1
zzQsGL_F9_c	154206.0	1180.0	107.0	55.0	6.0	1

Cluster 2

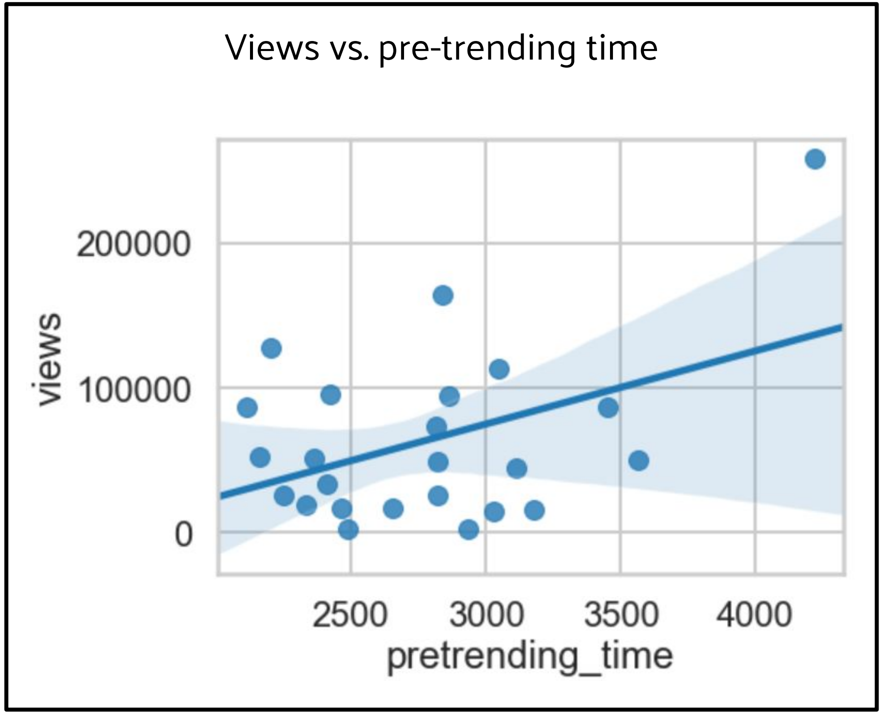
video_id	views	likes	dislikes	comment_count	pretrending_time	cluster_labels
0f7CuSU_huU	6491.0	15.0	0.0	8.0	1770.0	2
1x77e4XvqZ4	8502.0	42.0	0.0	4.0	1094.0	2
4Ek6UCI0YQs	2302.0	2.0	0.0	0.0	862.0	2
6A3cHzFQsql	112310.0	612.0	7.0	95.0	1878.0	2
6nJw-jPQYVI	192609.0	1345.0	24.0	126.0	908.0	2
7sEKooUZi7I	2992.0	28.0	0.0	1.0	1520.0	2
9o2FXVhjLyY	18264.0	315.0	29.0	122.0	1267.0	2
9vIKOFd73XM	10283.0	49.0	23.0	25.0	1719.0	2
A6owSHYjOiE	28954.0	74.0	9.0	11.0	1661.0	2
GDUncuEErzQ	7188.0	29.0	2.0	2.0	1820.0	2

Cluster 0: Significant Impact From Views on Pre-trending Time

Cluster 0

pretrending_time ~ views + likes + dislikes + comment_count

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       pretrending_time   R-squared:                       0.486
Model:                            OLS   Adj. R-squared:                  0.378
Method:                 Least Squares   F-statistic:                     4.499
Date:                Sat, 30 Nov 2019   Prob (F-statistic):             0.0100
Time:                        11:33:10   Log-Likelihood:                -174.87
No. Observations:                  24   AIC:                             359.7
Df Residuals:                      19   BIC:                             365.6
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      2700.8934    131.134     20.596      0.000    2426.428    2975.359
views             0.0058      0.002      2.840      0.010       0.002       0.010
likes            -0.0479      0.732     -0.065      0.949      -1.580       1.484
dislikes          1.8776      1.707      1.100      0.285      -1.694       5.450
comment_count    -4.5664      2.686     -1.700      0.105     -10.188       1.055
=================================================================================
Omnibus:                        0.203   Durbin-Watson:                   2.057
Prob(Omnibus):                  0.904   Jarque-Bera (JB):                0.086
Skew:                           0.122   Prob(JB):                        0.958
Kurtosis:                       2.835   Cond. No.                     1.39e+05
==============================================================================
```



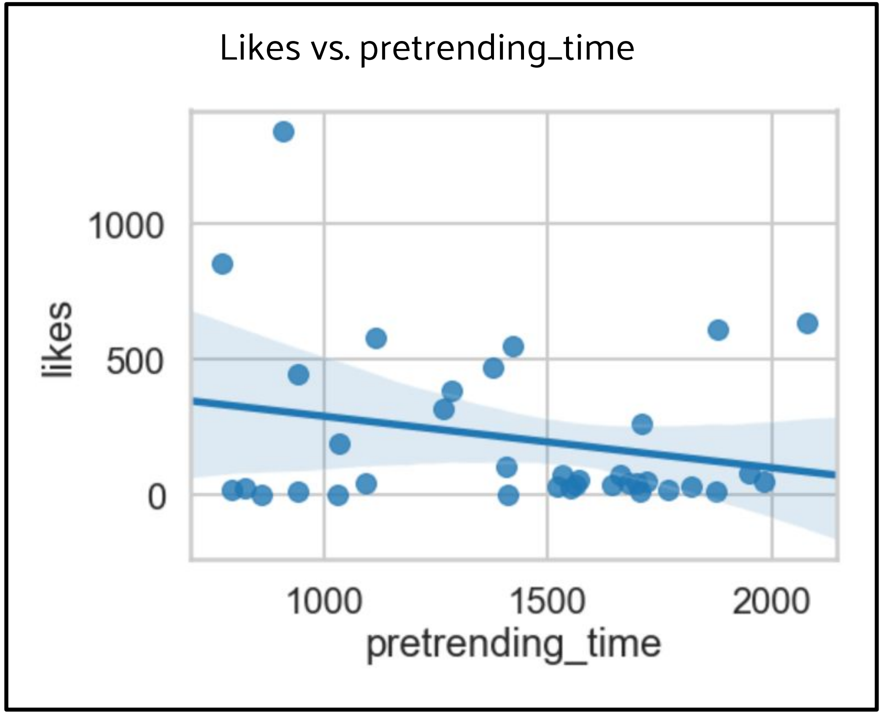
Observation: videos cumulate views before they actually get on the trending list

Cluster 2: Significant Impact From Likes on Pre-trending Time

Cluster 2

pretrending_time ~ views + likes + dislikes + comment_count

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        pretrending_time   R-squared:                       0.155
Model:                             OLS   Adj. R-squared:                  0.046
Method:                  Least Squares   F-statistic:                     1.425
Date:                 Sat, 30 Nov 2019   Prob (F-statistic):              0.249
Time:                         11:34:39   Log-Likelihood:                -261.36
No. Observations:                   36   AIC:                             532.7
Df Residuals:                       31   BIC:                             540.6
Df Model:                            4
Covariance Type:             nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept      1457.4035     79.373     18.361      0.000    1295.521    1619.286
views             0.0027      0.002      1.435      0.161      -0.001       0.007
likes            -0.8888      0.471     -1.886      0.069      -1.850       0.072
dislikes          1.4120      9.097      0.155      0.878     -17.142      19.966
comment_count     1.5655      3.199      0.489      0.628      -4.959       8.090
==============================================================================
Omnibus:                         3.158   Durbin-Watson:                   2.489
Prob(Omnibus):                   0.206   Jarque-Bera (JB):                1.999
Skew:                           -0.357   Prob(JB):                        0.368
Kurtosis:                        2.093   Cond. No.                     9.07e+04
==============================================================================
```



Observation: the higher the like count, the shorter pre-trending time will be

Cluster 1: No Significant Relationship Observed

Cluster 1

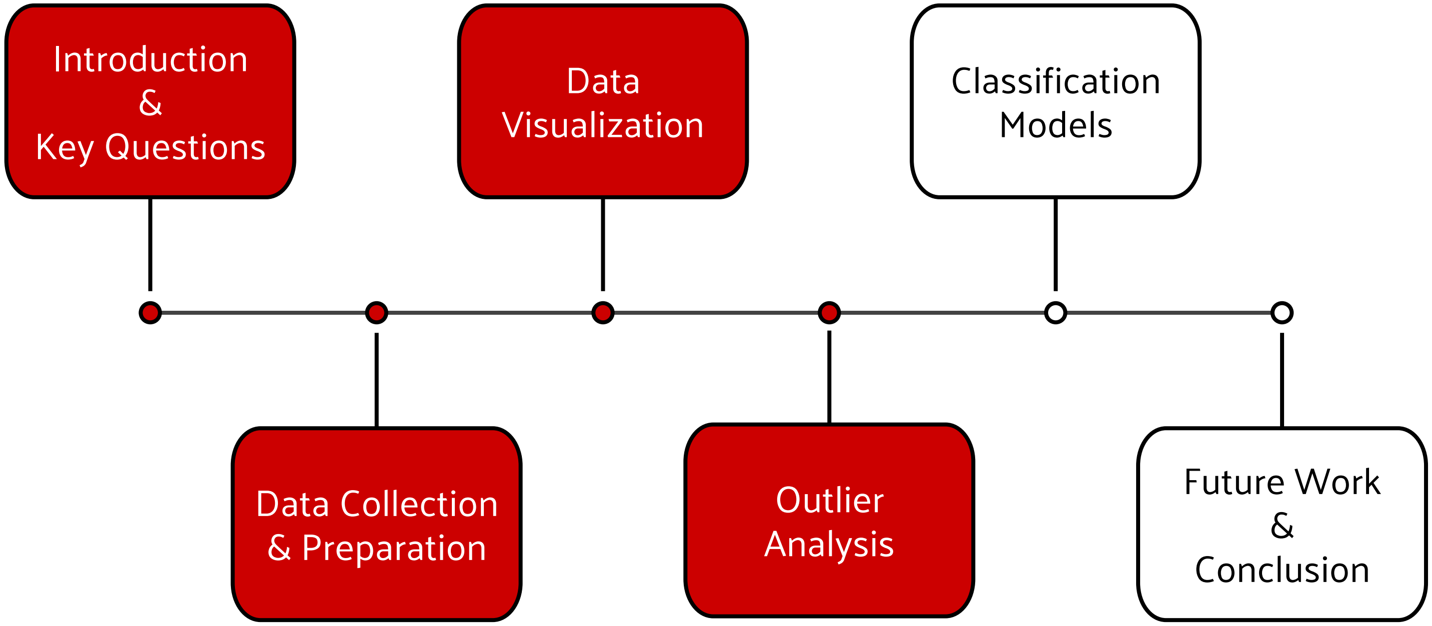
pretrending_time ~ views + likes + dislikes + comment_count

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       pretrending_time   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                    0.8117
Date:                Sat, 30 Nov 2019   Prob (F-statistic):              0.518
Time:                        11:34:05   Log-Likelihood:                -1942.1
No. Observations:                 330   AIC:                             3894.
Df Residuals:                     325   BIC:                             3913.
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept        35.3264      5.125      6.893      0.000      25.245      45.408
views         -5.644e-07   2.43e-06     -0.233      0.816   -5.34e-06    4.21e-06
likes         -4.426e-05      0.000     -0.361      0.718      -0.000       0.000
dislikes      -5.587e-05      0.000     -0.138      0.890      -0.001       0.001
comment_count    -0.0004      0.001     -0.540      0.589      -0.002       0.001
==============================================================================
Omnibus:                      327.036   Durbin-Watson:                   2.090
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6884.267
Skew:                           4.427   Prob(JB):                         0.00
Kurtosis:                      23.550   Cond. No.                     3.49e+06
==============================================================================
```

Rationale

- Skewness for Cluster 1 linear regression is high compared to other two groups of clusters, which means containing data that has various of pre-trending time
- Cluster 1 has the highest average views, likes and dislikes, despite they are not significant for pre-trending time
- Cluster 1 has similar characteristics as “inner” dataset

Presentation Outline



Feature Engineering

Variable Transformation

Tags (object)	→	Tag number (integer)
Title (object)	→	Title length (integer)
Comments_disabled (bool)	→	Comments_no (float)
Video_error_or_removed (bool)	→	Video_error_or_removed_no (float)

New Variable Creation

Like ratio = likes/(likes + dislikes)

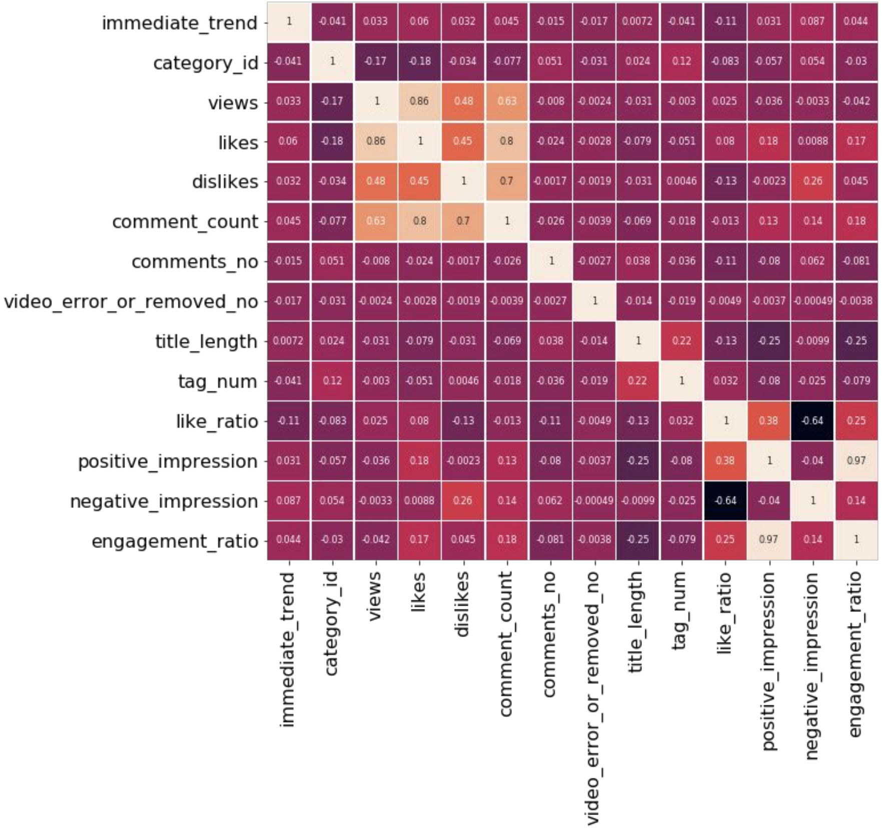
Positive impression = likes/views

Negative impression = dislikes/views

Engagement ratio = (comment count + likes + dislikes)/views

14 attributes, 39281 observations

Feature Selection via Correlation Analysis

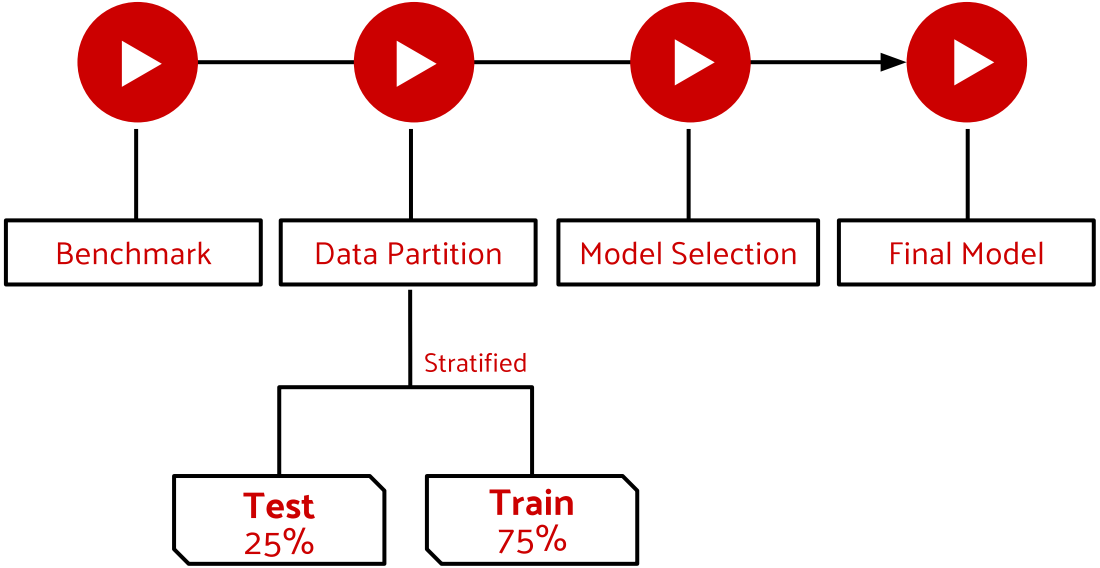


Correlation with *Immediate_trend* attribute

Attributes	Corr. Coeff.	p-Value
category_id	−0.041042	4.033523e−16
views	0.033477	3.208751e−11
likes	0.060450	3.940152e−33
dislikes	0.032090	1.997401e−10
comment_count	0.044811	6.367884e−19
comments_no	−0.015116	2.735510e−03
video_error_or_removed_no	−0.017156	6.729631e−04
title_length	0.007221	1.524000e−01
tag_num	−0.041054	3.959438e−16
like_ratio	−0.113157	4.285877e−112
positive_impression	0.030669	1.203739e−09
negative_impression	0.086773	1.586447e−66
engagement_ratio	0.043715	4.402762e−18

11 attributes, 39281 observations
Label: *Immediate_trend*
Features: the other 10 attributes

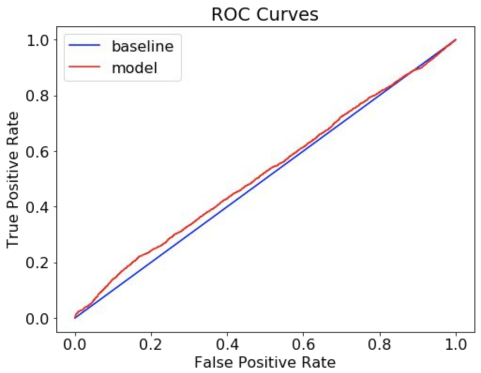
Modeling Process Overview



Modeling Optimization

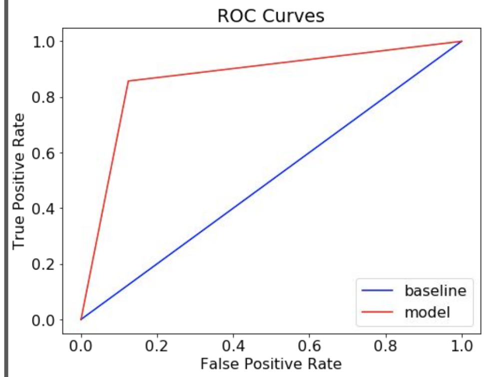
Benchmark model (GNB)
Recall: 0.05 Precision: 0.54 ROC: 0.50 Accuracy: 0.52

Naive Bayes



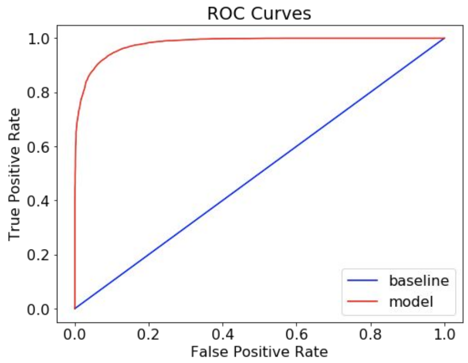
Recall: 0.05
Precision: 0.53
Roc: 0.52
Accuracy: 0.52

Decision Tree



Recall: 0.86
Precision: 0.87
Roc: 0.87
Accuracy: 0.52

Random Forest



Recall: 0.91
Precision: 0.93
Roc: 0.98
Accuracy: 0.92

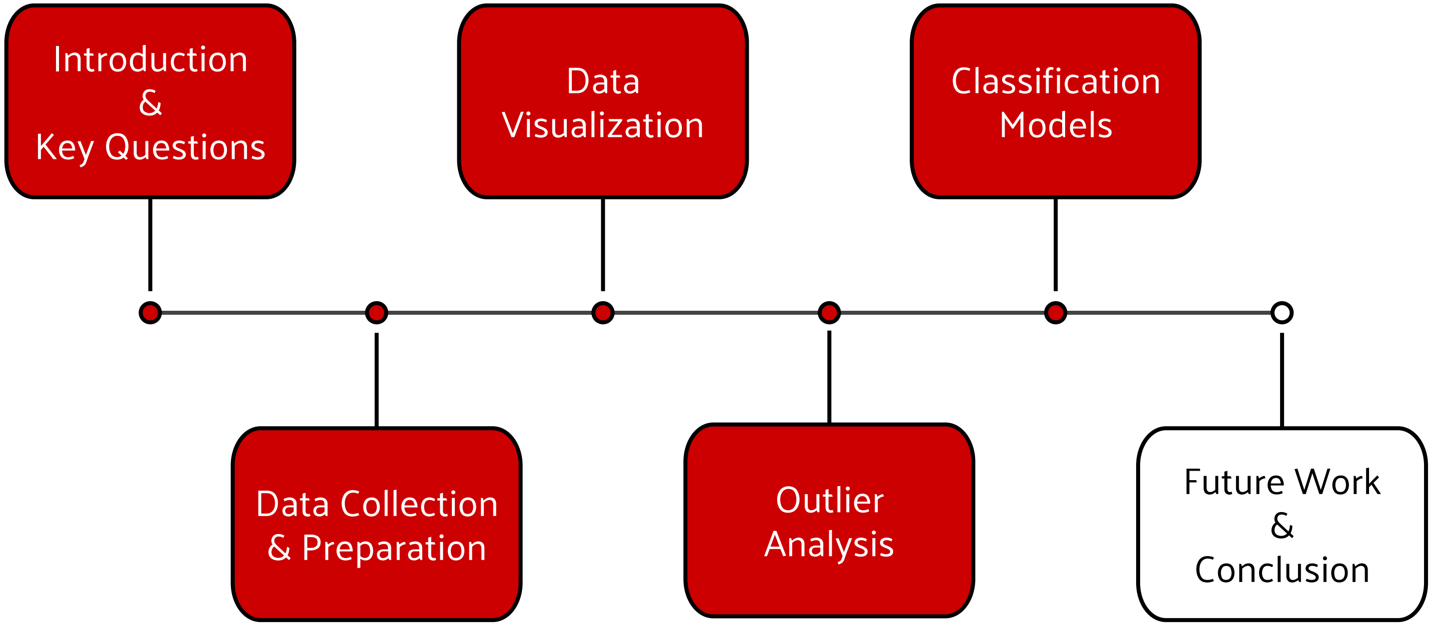
Tag Number is the Most Important Factor

Results from random forest model		
Attributes	**Importance**	**Impact Direction**
tag_num	0.133807	Negative
comment_count	0.119388	Positive
like_ratio	0.115624	Negative
negative_impression	0.114163	Positive
likes	0.098008	Positive
dislikes	0.096474	Positive
postive_impression	0.087713	Positive
engagement_ratio	0.086578	Positive
views	0.080964	Positive
category_id	0.067280	Negative

Takeaways

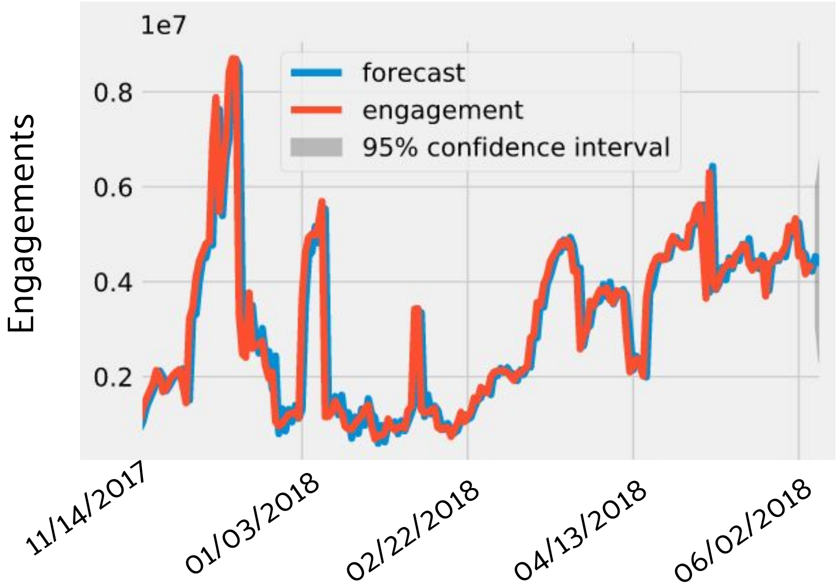
- Reduce tag number to make the video more focused
- More comments, more likely to be immediately trending
- Both like and dislike are important to make immediate trending

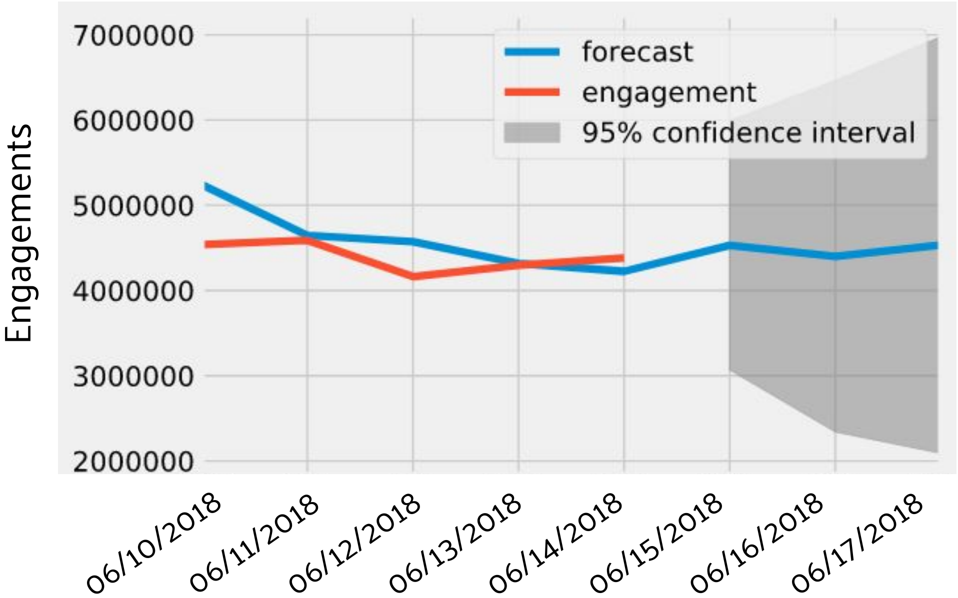
Presentation Outline



Future Work: Time Series Analysis - Preliminary Results

Time series analysis for trending entertainment videos





Recommendation: not publish video in next 3 days due to slightly decrease in predicted engagement

Summary

What we did

- Conducted EDA to understand the difference between immediate trend and late trend by category, and difference in views, positive impression and negative impression between each category
- Performed clustering and linear regression analysis to study the videos with long pre-trending time (outliers)
- Constructed classification models to predict immediate trend videos and to figure out important features

What we found

To make an immediately trending video

- Publish current events videos
- Make the video more focused by reducing tag number
- Encourage leaving comments
- Both like and dislike are important

Other insights for content creators

- Explore opportunities in Education and Howto & Style categories

Future work

- Construct the time series models to predict the engagement changes in trending videos for recommending video publish time

Thank You

Supplementary Information

Remove Upper Outliers via IQR

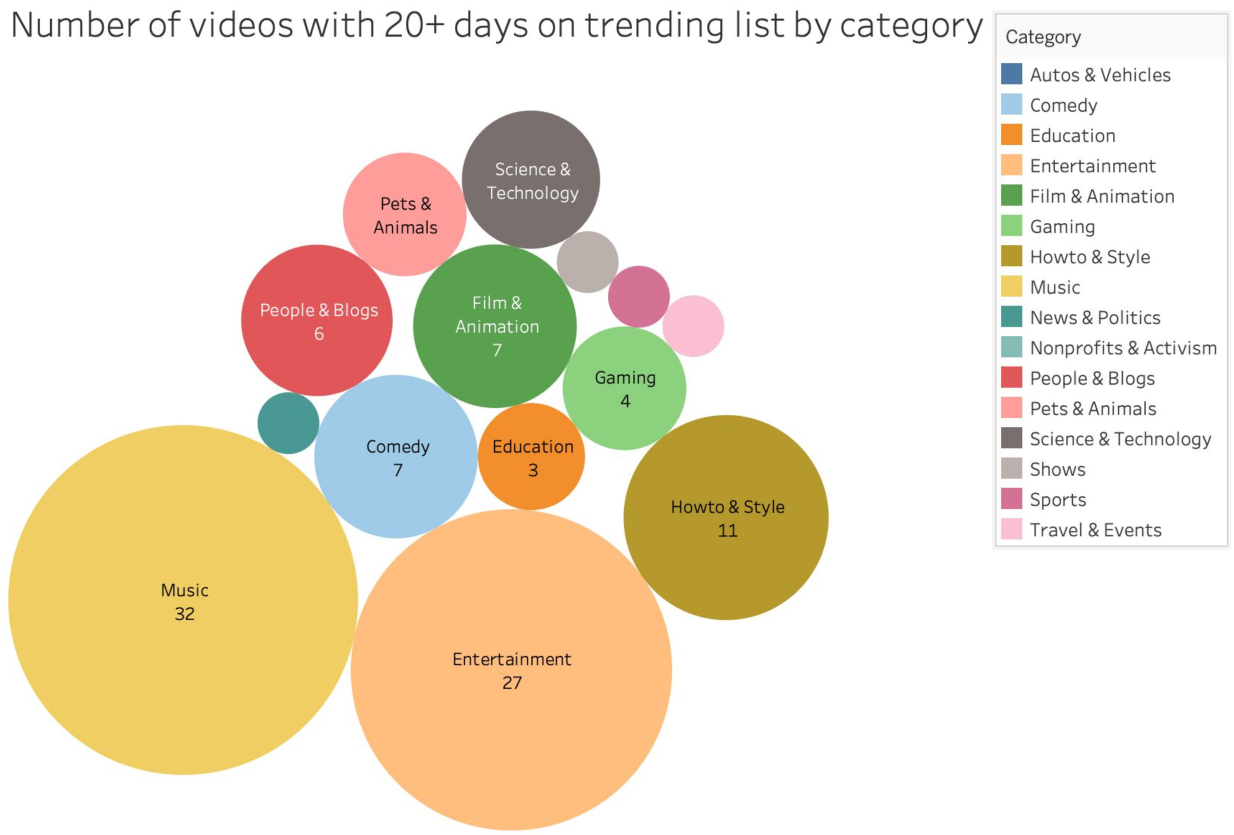
Pretreding time overview

count	6334.000000
mean	22.041048
std	205.918572
min	0.000000
25%	1.000000
50%	2.000000
75%	3.000000
max	4215.000000

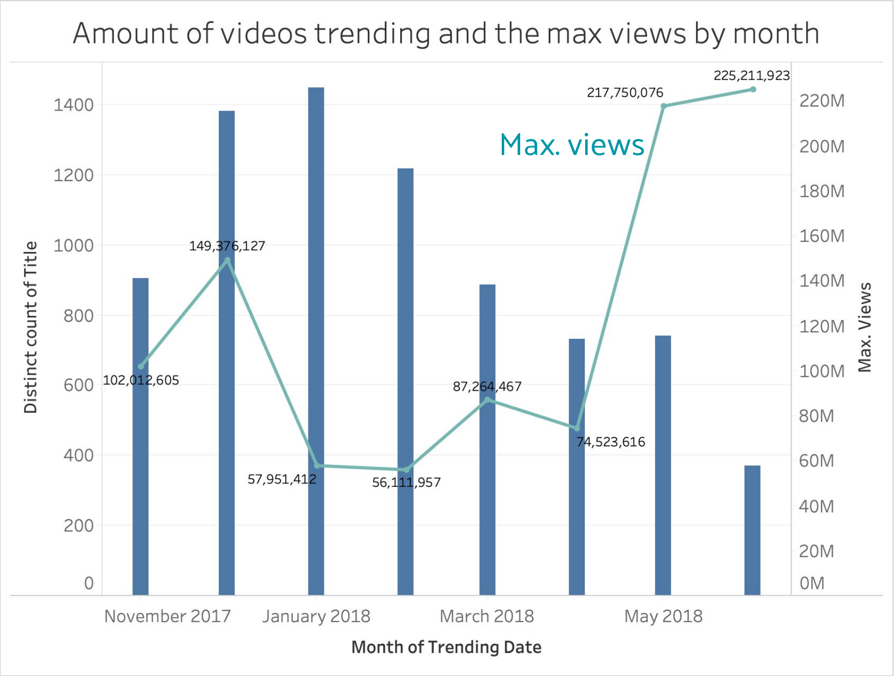
```python
def only_outlier(cleaned, col_name):
    q1 = cleaned['pretrending_time'].quantile(0.25)
    q3 = cleaned['pretrending_time'].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    only_out = cleaned.loc[(cleaned['pretrending_time'] > fence_high)]
    return only_out
```

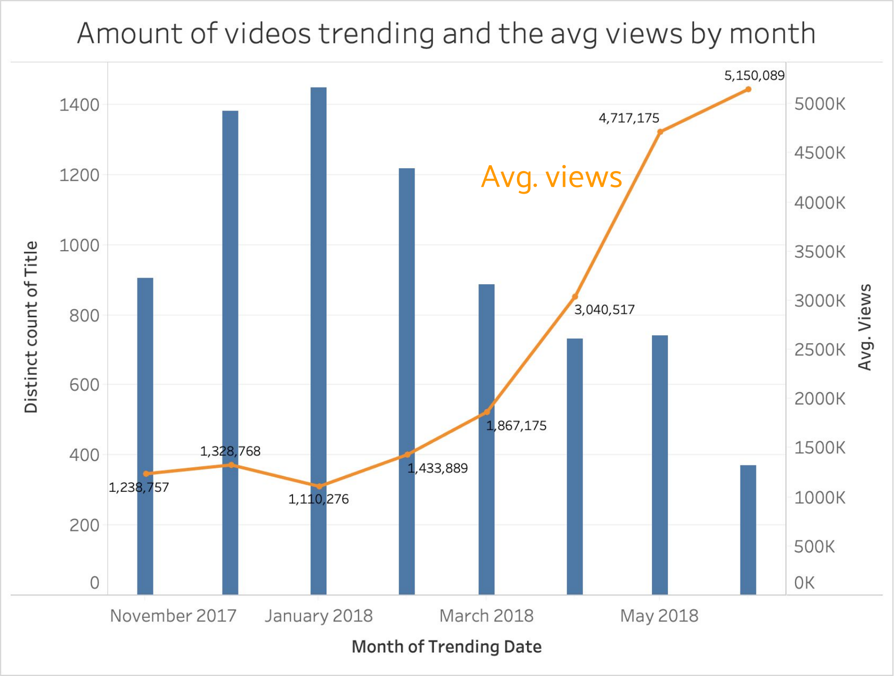
```python
def remove_outlier(cleaned, col_name):
    q1 = cleaned['pretrending_time'].quantile(0.25)
    q3 = cleaned['pretrending_time'].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low  = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    cleaned_out = cleaned.loc[(cleaned['pretrending_time'] < fence_high)]
    return cleaned_out
```

Long-Trending Videos



Dramatic Increase in Views in May 2018





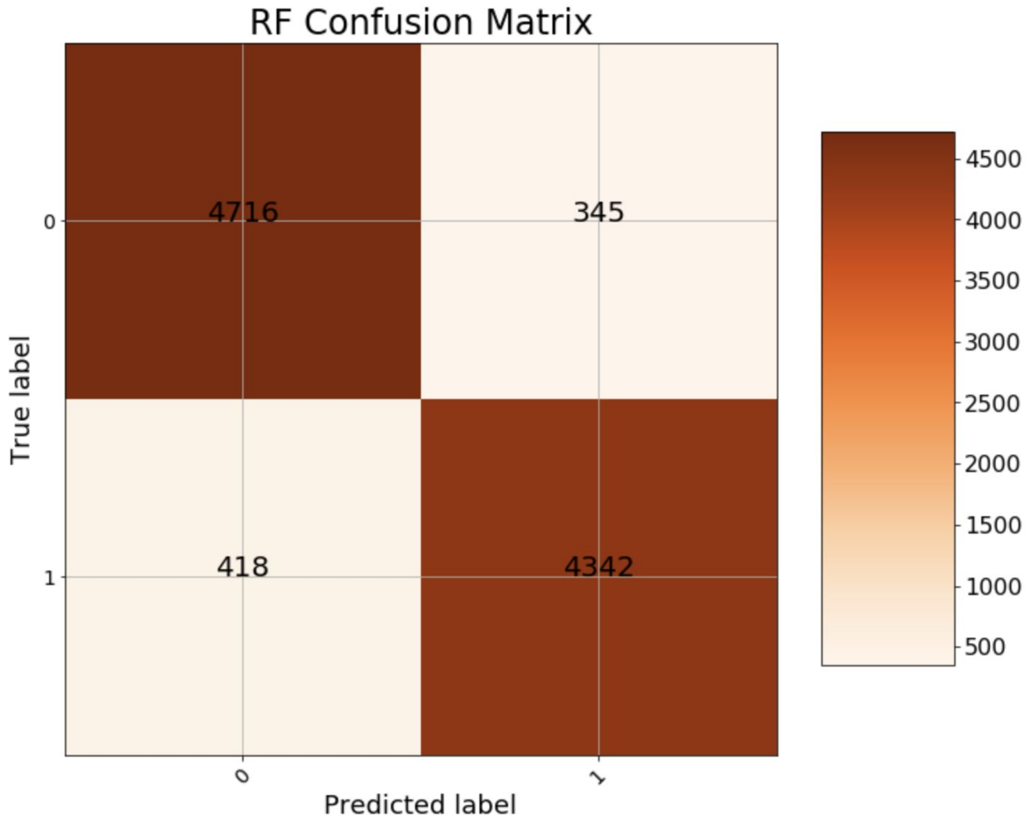
May 2018: “Childish Gambino - This Is America” was published

Identifying Cluster Number via Silhouette Score

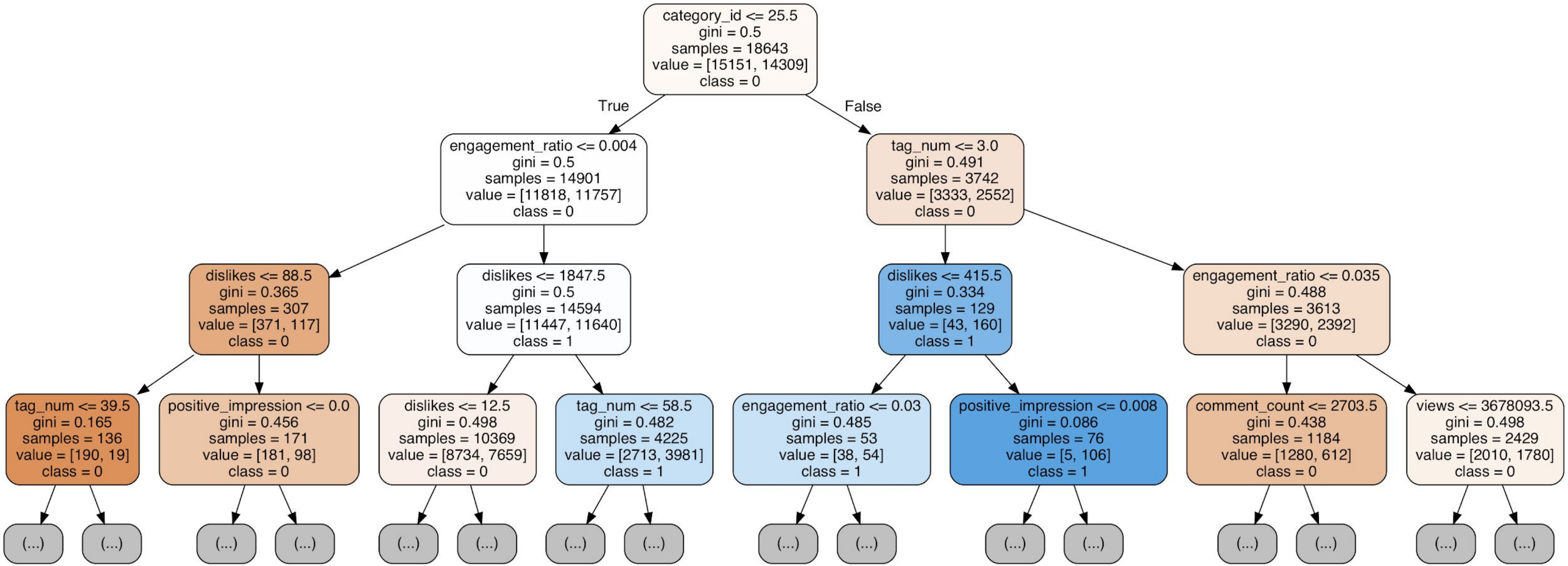
```
# Print out `s_score_dict`
print(s_score_dict)
```

```
{2: [0.9009486705900968], 3: [0.8971033378954894], 4: [0.8853091053839264], 5: [0.885842839667557], 6: [0.8887056890145145], 7:
[0.8833120752778203], 8: [0.882921629467595], 9: [0.8796317163090666], 10: [0.8768285388044179]}
```

Confusion Matrix For Random Forest



One Tree from Random Forest Model



ARIMA: AutoRegressive Integrated Moving Average

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

d: the number of non seasonal differences
y: original series
Y: stationaized (differenced) series

Forecasting Equation

$$\hat{y}_t = \overbrace{\mu}^{\text{constant}} + \overbrace{\phi_1 y_{t-1} + \ldots + \phi_p y_{t-p}}^{\text{AR terms (lagged values of } y)}$$

$$\underbrace{-\theta_1 e_{t-1} \ldots - \theta_q e_{t-q}}_{\text{MA terms (lagged errors)}}$$

By convention, the AR terms are + and the MA terms are –

Source: https://towardsdatascience.com/unboxing-arima-models-1dc09d2746f8