YouTube Daily Trending Videos Analysis

Danielle Ip & Youngshin Kim
Dec 3, 2018

Stat 154

Contents

- Project Context
- Exploratory Data Analysis (EDA)
- Our Analysis - Methods
- Our Analysis - Results & Interpretations
- Conclusion

Contents

- **Project Context**
- Exploratory Data Analysis (EDA)
- Our Analysis - Methods
- Our Analysis - Results & Interpretations
- Conclusion

Project Context

Our Kaggle dataset is of daily trending videos on YouTube. We seek to better understand why and how videos trend on YouTube.

What is going on behind the data

- The daily trending videos are determined by YouTube's own algorithm*
- The dataset was collected using the YouTube API

*This algorithm uses a combination of factors related to number of views, shares, and other user interactions with the videos

Why we chose this dataset

We are interested in gaining more insight into YouTube's formula for determining if a video trends or not

We would also like to see what general trends and common characteristics there are between these videos

Contents

- Project Context
- **Exploratory Data Analysis (EDA)**
- Our Analysis - Methods
- Our Analysis - Results & Interpretations
- Conclusion

Exploratory Data Analysis (EDA)

We focused our dataset and research to better understand YouTube's trending algorithm, specific to the US.

Scope & Granularity of dataset:

- 8 months (11/14/17 - 6/14/18), scraped for each day
- US region
- Title, channel title, publish date, trending date, number of views, likes, dislikes, comments

Assumptions? Underlying distributions?

- Data failed to meet normality assumptions for quantitative variables of views, likes, and more

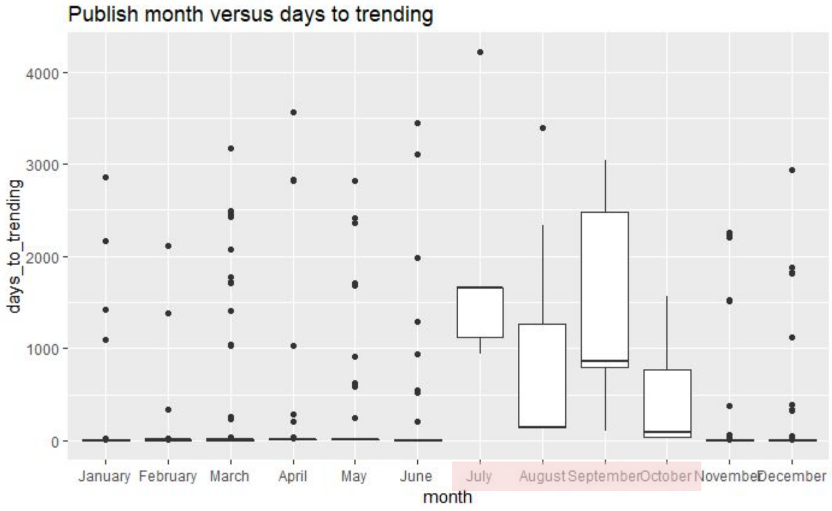
Key stats

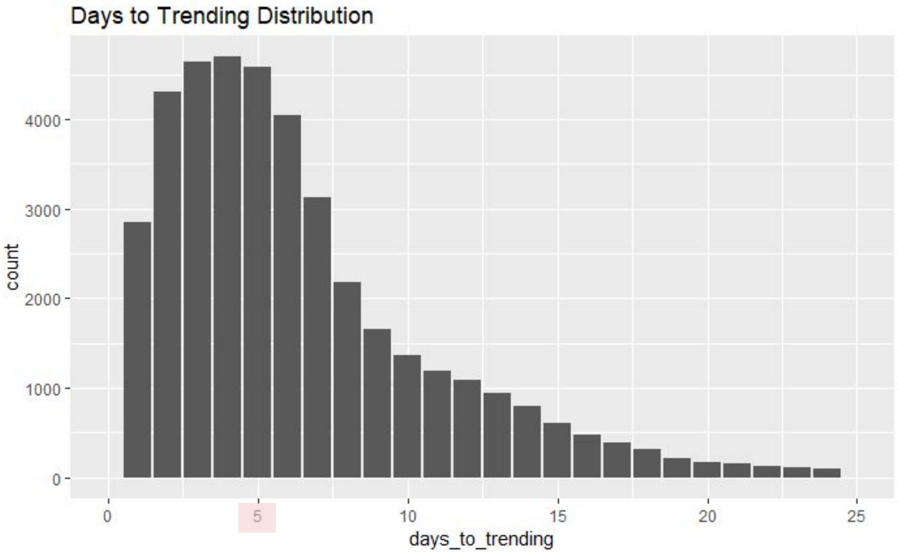
40,949 entries

Top categories: Entertainment, Music, How to and Style

Views range: 549-225,211,923

Data failed to meet normality assumptions but did show some patterns and useful distributions. This aided in selecting our variables of interest.





Contents

- Project Context
- Exploratory Data Analysis (EDA)
- **Our Analysis - Methods**
- Our Analysis - Results & Interpretations
- Conclusion

Logistic Regression shows us that views and likes are most indicative in determining if a video trends within 5 days of its publishing date.

We ran backwards selection to focus in on our variables of interest.

Views and likes are the most indicative.

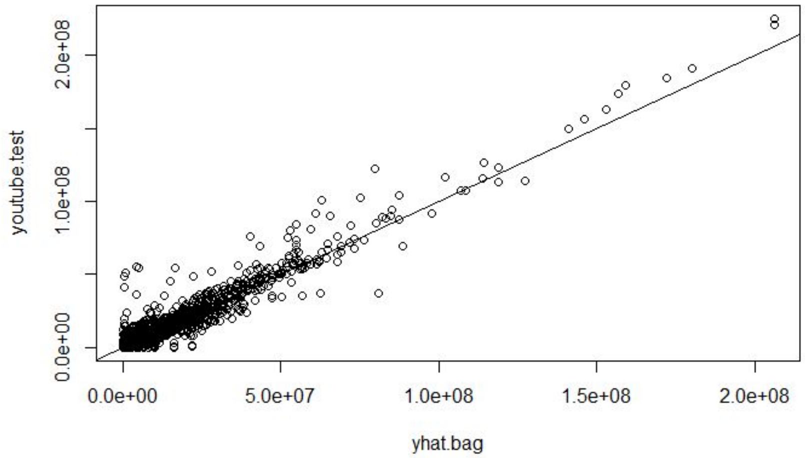
```{r}
glm.selected <- glm(fivedays ~ views + likes + dislikes + waitmonth, data=trainset, family=binomial)
summary(glm.selected)
```

```
Call:
glm(formula = fivedays ~ views + likes + dislikes + waitmonth,
    family = binomial, data = trainset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6441  -1.2341   0.9857   1.0795   2.7979

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.699e-01  8.358e-02   3.229  0.00124 **
views         -3.216e-07  7.579e-08  -4.243  2.2e-05 ***
likes          3.522e-06  1.214e-06   2.901  0.00372 **
dislikes       5.257e-05  2.783e-05   1.889  0.05890 .
waitmonthTRUE -1.382e+01  3.786e+02  -0.036  0.97089
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1133.5  on 817  degrees of freedom
Residual deviance: 1082.9  on 813  degrees of freedom
AIC: 1092.9

Number of Fisher Scoring iterations: 12
```

Random Forest Bagging showed us which of our quantitative variables were most indicative of high viewership

We centered our focus on viewership to determine what factors influenced a "viral" video

Likes and dislikes were the most correlated to views




```
[1] 3.543079e+12
                    %IncMSE IncNodePurity
likes             113.70400  8.065164e+17
dislikes          103.66571  1.471272e+17
comment_count      62.45669  3.320877e+16
days_to_trending   52.81602  7.231065e+16
```

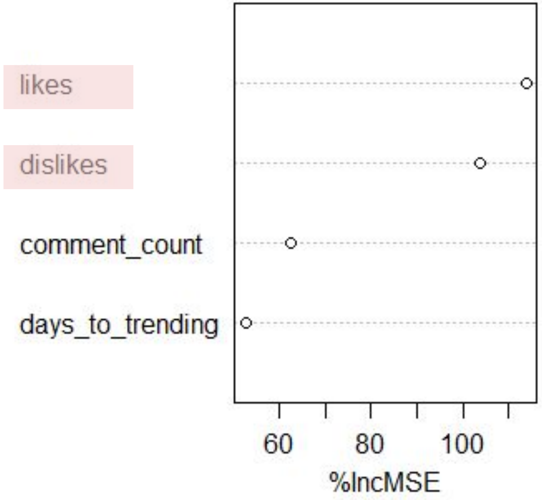
Contents

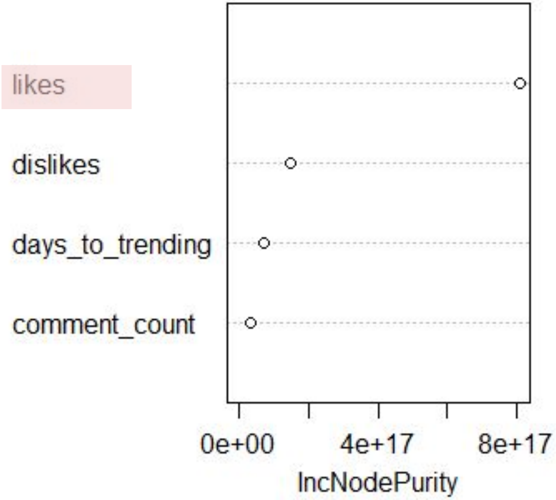
- Project Context
- Exploratory Data Analysis (EDA)
- Our Analysis - Methods
- **Our Analysis - Results & Interpretations**
- Conclusion

Likes and dislikes are important indicators of the number of views a video will have; they have large %IncMSE and different order of magnitude compared to the other variables

IncMSE is the MSE with the original dataset and the permuted dataset. The variables with *high IncMSE would be the predictors that matter the most in the dataset.*

IncNodePurity is the increase in node purity or decrease in node impurity, so the *variables with high node purity values are the important predictors in the dataset.*





Contents

- Project Context
- Exploratory Data Analysis (EDA)
- Our Analysis - Methods
- Our Analysis - Results & Interpretations
- **Conclusion**

Conclusion

What makes a video trend *quickly*? Views, likes, and dislikes
What makes a video *viral*? Likes and dislikes

Quickly = 5 days or less to trending

Backwards selection logistic regression

- Average days to trend: 16.81
- Median days to trend: 5

Determinants of quick trending videos?

- Views, likes, dislikes
- Why? Most indicative of engaged users that willingly and opinionatedly interact with it

Viral = High amount of views

Random forest model

- Average views: 2,360,785
- Median views: 681,861

Determinants of quick trending videos?

- Likes and dislikes
- Why? Video creators concerned with virality want high viewership and polarizing reactions

Thank you