# Model Story

Phase 1: Learning API

Phase 2: Tuning

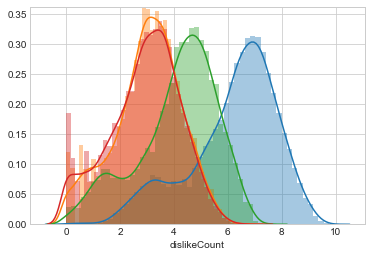
Phase 3: EDA

Phase 4: First Features & Models

Phase 5: New Structure

# Descriptive

## Like / Dislike / Comment / View Count Interaction



Key:

* Blue - views
  + bump at 3.25
* Green - Likes
  + bump at 1.6
* Red - Comments
* Orange - Dislikes

If nothing else comes out of this analysis, the above graph should prove interesting enough.

Table 1: Ratio of Log of Counts at fit Mode (row divided by column)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Views | Likes | Dislikes | Comments |
| Views |  | 1.58 | 2.37 | 2.22 |
| Likes | 0.63 |  | 1.50 | 1.41 |
| Dislikes | 0.42 | 0.67 |  | 0.94 |
| Comments | 0.45 | 0.71 | 1.06 |  |

Dislikes and comments are almost perfectly matched, within 5% in the log (50% actual) of eachother. Dislikes and comments would seem to reflect similar levels of engagement with a video. If you’re wondering at the effort level of comments on youtube, I invite you to scroll through and find something that looks like a fully formed thought anywhere in the first dozen results.

NOW the really interesting thing about this shape similarity is that it shows in log-space, meaning there is a diminishing degree of engagement for high view count videos. This is going to make analysis very difficult. Interestingly, the relationship falls off even in log space, although not more than as a peculiarity of the math (6/4 is a bigger number than 6+2 / 4+2).

Another peculiarity is the left-most bump in each of the curves. The bump ratio holds to the same description as the mode ratio. I strongly suspect that this is a feature of the YouTube search algorithm, which touches on how this data was collected. I searched for each individual letter of the alphabet, then returned results by relevance. Nothing is relevant to the letter ‘d’, So my guess is that view count is about 5,000 times more important than everything else is the search algorithm (ratio of mode views to bump height views).

I also suspect that what this means is there are actually two different sets of data in this set which could possibly throw off the analysis. This strongly suggests setting some threshold - e.g. 4.1 - at which to perform the analysis.

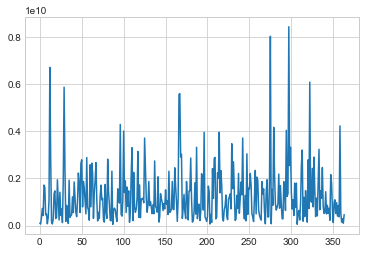
I would love to know if the engagement features described here are time dependent, but for now I’m just going to have to assume that people are disinterested in wildly popular videos from the outset.

## Publish Date Features

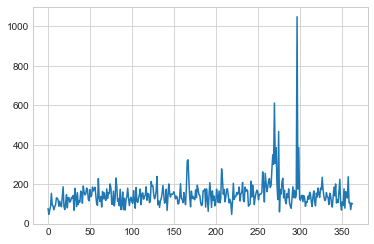
There is no reason there should be any signal here, and yet…

The importance of examining the right variable:

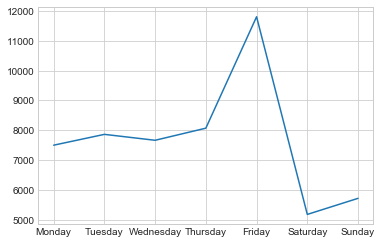
Noise:



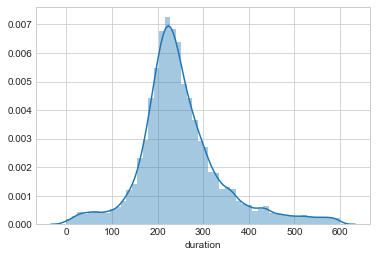
Signal:



Release on a Friday:



## Duration features, or lack thereof…



## Text Analysis

Sentiment analysis and wordcount done for each of these.

### Title



Sentiment score:

Discussion:

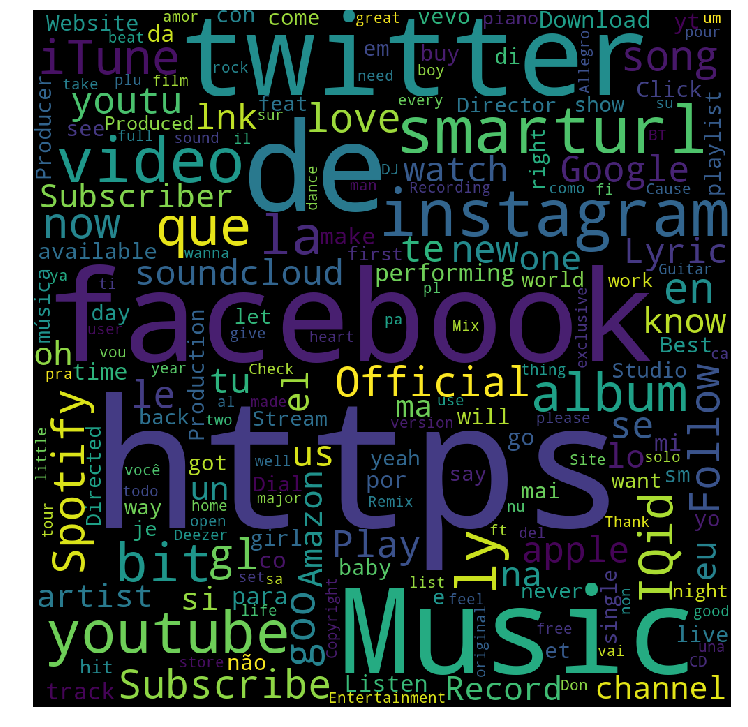
### Tags



Sentiment score:

Discussion:

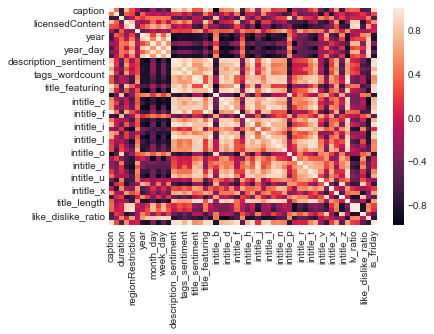
### Description



Sentiment score:

Discussion:

## After feature engineering



# Predictive

## Issues

A key

## Classification

## Regression

Vectorizer alone was .25 regression, but adding to existing models generally made them worse (except for GradientBoost)