Midterm Report: Youtube Trending

Kyle Wadell[kaw282], Brandon Wang[bhw47]

October 2019

1 Introduction

The main goal of our project has evolved a little from our project proposal. Instead of trying to determine what criteria Youtube uses to create the trending list, we are going to analyze the performance of videos once they have been added to trending, and attempt to predict the duration that they will remain on the list. The popularity of Youtube as a platform has been rapidly increasing and many people have even created entire careers out of promoting their own personal brand or channel. Our goal is to analyze youtube video data in order to find the makeup and factors of different videos that makes them successful or popular on Youtube. Digging even deeper, we would like to find the relationship between different factors and utilize that in order to help Youtubers enhance their videos to reach a broader audience, gain more subscribers, and overall clout on this video platform.

2 Data Collection

2.1 Moving Away from the Kaggle Data Set

During our initial thought process, we looked at a data set from Kaggle on all trending Youtube videos which included a set of performance statistics for each video which had been added to the trending. A description of these features included in the figure below. This list excludes the "video error or removed" parameter because we are only considering videos which were removed organically and have therefore filtered out any observations where this feature was true. It is our intention to create embedding off of the text based features such as channel name and description, although at present we are simply treating them as categories whenever possible, such as former, or excluding them from analysis when not, such as the latter.

Table 1: Original Kaggle Data Set

Feature Id	**Description**
video id	Unique String
trending date	Timestamp
title	String
channel title	Categorical
category id	Categorical
publish time	Timestamp
tags	Subset of Categories
likes	Continuous
dislikes	Continuous
comment count	Continuous
thumbnail link	Image File
description	String

We believed we would be able to find some sort of insight for the cause on what made all these videos trending and kept them trending. Our first thoughts were to group the videos by different features in order to see if it was possible to isolate and quantify video qualities such as popularity, controversy, or audience reach. However, the initial data set only included videos final performance after being on trending and lacked any information about the way that performance had varied across time.

2.2 Web Scrapping

In order to introduce a temporal axis to our data, we repurposed the web scraper that was used to generate the original kaggle data set to repeatedly generate samples from the trending page overtime. In order to use the same scraper on Youtube, we applied for an API key through Google/Youtube which now allows us to scrape the same data from the trending page of Youtube. Utilizing this web scraper, we were able to observe how long videos were remaining on the trending page and how they performance was varying over this time. The scraper was run at 15 minute intervals on the trending page of Youtube for set timeframes in order to find the time at which a new video entered the trending page and the time at which it was removed from the trending page. Our new set includes all features that were present in the original data set as well as several additional parameter which are displayed in figure 2.

Table 2: Additional Features

Feature Id	**Description**
time observed	Timestamp of observation
Duration	Time video has been trending, so far
Δ Likes	Change in Likes since video added to Trending
Δ Dislikes	Change in Dislikes since video added to Trending
Δ Comments	Change in Comments since video added to Trending
Δ Views	Change in Views since video added to Trending

2.3 Server Procurement

Initially, we assumed that videos would only stay on the trending page for a short duration of time, around 10-20 minutes and built our dataset over a set of five hour time windows. We have since realized that videos are staying on the trending page much longer than anticipated. In fact, the data set we captured actually included no instances of videos being shifted on or off the trending list. As a result, much of our current analysis is instead centered around predicting a video's performance while on the trending list, as opposed to its duration. Moving forward, we have set up a server to run our scraper repeatedly over the course of several weeks, which should provide a long enough time frame to begin to analyze videos' trending duration.

3 Preliminary Analysis

As discussed in the previous discussion, our current analysis clearly points to needing a larger data set. Plots of our current set are contained in this section, as well as some naive models that were built off of them.

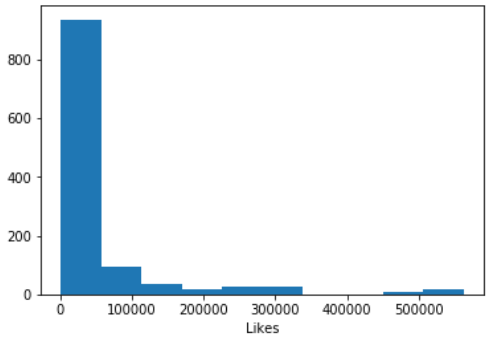

Figure 1: Likes of Trending Videos **??**).

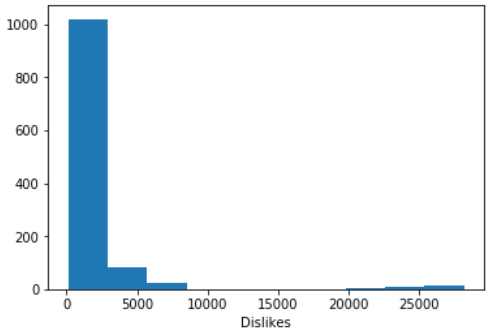

Figure 2: Dislikes of Trending Videos

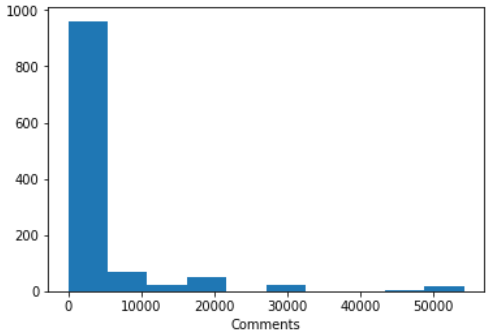

Figure 3: Comment Count of Trending Videos

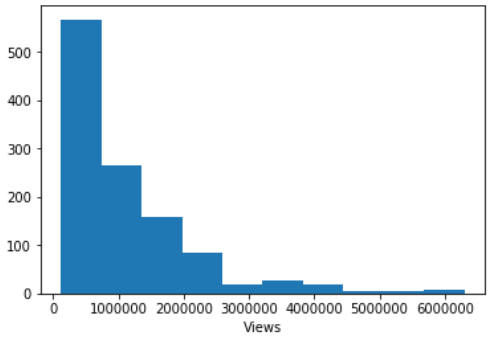

Figure 4: Views of Trending Videos

3.1 Naive Model

As a first step, we've built a handful of naive linear model's using our processed features to predict different measurements of video performance. One of the more simple versions of this model, which uses just the initial change in likes to predict final views is included in the figure below. We found that removing an outlier observation, Apple's Launch Video, greatly improved our out of sample error across of these naive models. The final version will likely have to include some formulaic method to detect and remove such outliers, although it is possible this will no longer be necessary once we have access to a broader data set.

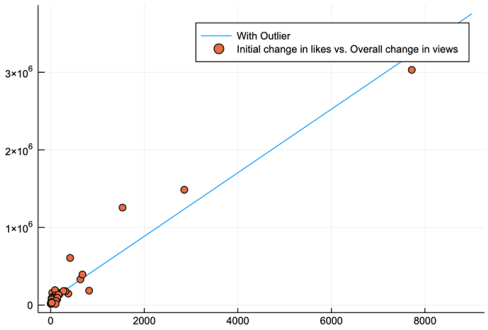

Figure 5: Model Predicting Final View Count with Initial Change in Likes, With Outlier

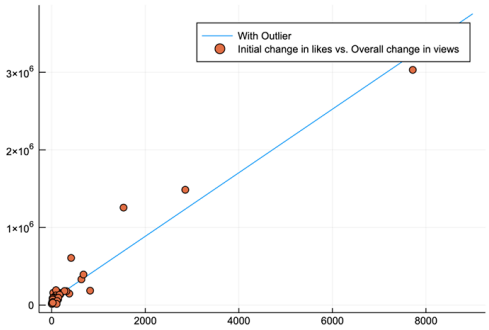

Figure 6: Model Predicting Final View Count with Initial Change in Likes, Outlier Removed