YouTube Analysis Report

A. Motivation

YouTube is the world-famous video-sharing website. As a student who consumes YouTube videos every day, whether it is for educational or entertainment purposes, we want to understand how the YouTube algorithm work. In this paper, we will analyze the YouTube trending videos for coming into conclusion for the following questions:

1. Can we predict the number of comments that a video will have?
2. Can we classify YouTube video categories based on the video's statistics?
3. Are music videos less interactive compared to non-music videos?
4. How many views does it take for a YouTube video to go on trending?

B. Our Dataset and Data Cleaning

YouTube maintains a list of the top trending videos on the platform. The YouTube trending dataset that we have gathered is from Kaggle. It consists of statistics such as *video_id* (unique identifier for each video), *trending_date* (the date when the video went on trending), *title*, *channel_title*, *category_id* (the id which corresponds to a list of category types), *publish_time*, *tag*, *views* (the number of views when the video went on trending), *likes*, *dislikes*, *comment_count*, *thumbnail_link*, *comment_disabled*, *ratings_disabled*, *video_error_or_removed*, *description*. Furthermore, these statistics are captured from 2017 to 2018 for various countries. It is a record for daily trending videos which means that the statistics are captured from the day that the videos went on trending.

The dataset includes statistics for different countries, but they were separated into different CSV files. Each CSV file was also associated with JSON file that maps the category_id to a category type. The separation of JSON files may be because the category_id was different for each country. We decided to combine all of the files into one file so that we would not have to refilter later on for our analysis. First of all, we associated the category_id to the category type for each country. Then, we combined the country records into one. Because the dataset contained some undesired records such as NaN for certain columns, we had to filter all those records.

Since our dataset was already relatively clean, it simplified the data cleaning process. One of the problems that we encountered was having special line terminator characters when we were saving the combined data. We discovered this when we examined particular columns and noticed many NaN values in the DataFrame. After rigorous investigation, we found out that some of the columns were shifted, which caused other columns to be NaN. To fix this, we explicitly set the line terminator to be "\n" when saving and reading the combined data.

1. Predicting the number of comments that a video will have

Our initial guess is that there is a linear relationship between the number of comments and the other statistics of a video. For example, the number of comments can be predicted based on the number of views, likes, and dislikes.

First of all, we used linear regression on our entire data set to predict the number of comments. The features that we used were "views", "likes" and "dislikes". As you can see from the prediction score,

score_on_train_data=0.83018, score_on_valid_data=0.82979

We got a fairly accurate prediction for the views. However, we realized that viewers in certain video categories may have different behavior. For example, videos in the categories of Comedy and HowTo may be targeting audiences of different ages. The videos that mainly target the younger audiences (age 3-6) would likely have less interaction (fewer comments, fewer likes/dislikes).

After splitting and training the model for different categories, the prediction scores that we obtained were quite different. For examples:

category: Comedy, score_on_train_data=0.51276, score_on_valid_data=0.52986
category: Music, score_on_train_data=0.821, score_on_valid_data=0.82039
category: Entertainment, score_on_train_data=0.94268,
score_on_valid_data=0.94121
category: Howto & Style, score_on_train_data=0.35925,
score_on_valid_data=0.39935
category: Education, score_on_train_data=0.74956,
score_on_valid_data=0.69287
...

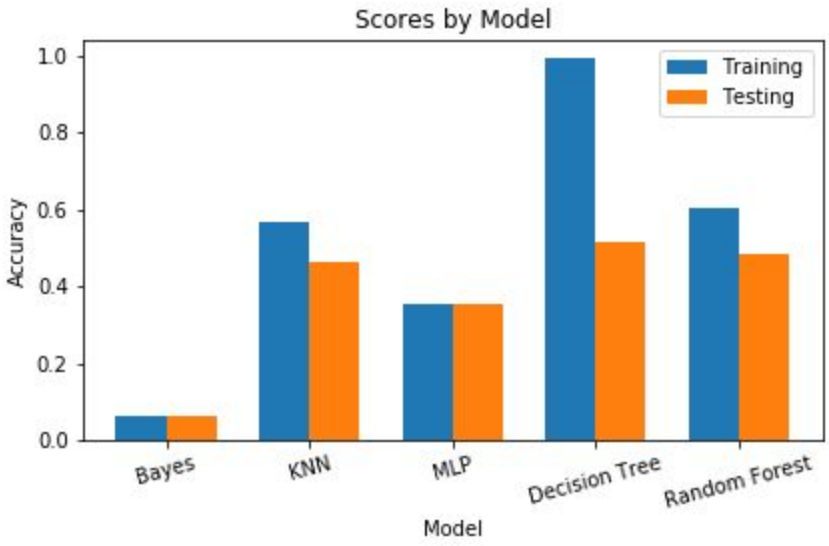
We realized that the linear regression model was not suitable to have a good prediction for some categories.

From the results above, we can see that the viewers' interactivity was quite different among different categories. Thus we should further investigate the relationship between category and other video statistics and investigate the interactivity of these videos.

2. Classify video category based on video statistics

In this analysis, we attempted to use a variety of machine learning classifiers to predict the category of a video. The features that we used were views, comment counts, likes and dislikes.

The best prediction we had is still far away from the significant result: testing accuracy of 52% by using the decision tree model. Keep in mind that the decision tree is overfitting a lot here.



Thus, we refined the question to "Whether we can categorize a video into music or non-music video?" Taking a look at the data, we found that around 86% of the data were non-music, which means one can easily obtain a decent score by predicting all entries to be non-music. However, when we run our analysis, we were able to obtain a

prediction score of 91% by using the Random Forest model. Therefore, the result allows us to be more confident in predicting a music video.

Limitation:

We felt that the models did not have sufficient meaningful data to make a good prediction. The accuracy may increase if we used natural language processing to analyze the video title, channel title, tags, and description of each entry. It may also help if we used image recognition to examine the thumbnail of the videos.

Moreover, the data we captured is only at the moment the videos went on trending. We believe the change in the number of views, likes, and dislikes will be different for different categories after they go on trending. For example:

Music Video	Time	Views	Likes	Dislikes	Comment_count
Stampede - Alexander Jean Ft. Lindsey Stirling	Date on trending	296615	38.6k	0.4k	2348
	Today	6471733	153k	2.1k	4200
Change rate %		**21x**	3.9x	5.25x	1.8x

Non-music Video	Time	Views	Likes	Dislikes	Comment_count
HOW2: How to Solve a Mystery	Date on trending	80685	1.7k	0.1k	1312
	Today	660905	5.6k	0.4k	2388
Change rate %		**8.2x**	3.3x	4x	1.8x

The examples indicate that the prediction may be better if we use a dataset of all YouTube videos rather than just using the on trending videos. Also, we would use more complex models such as Keras to obtain a better prediction.

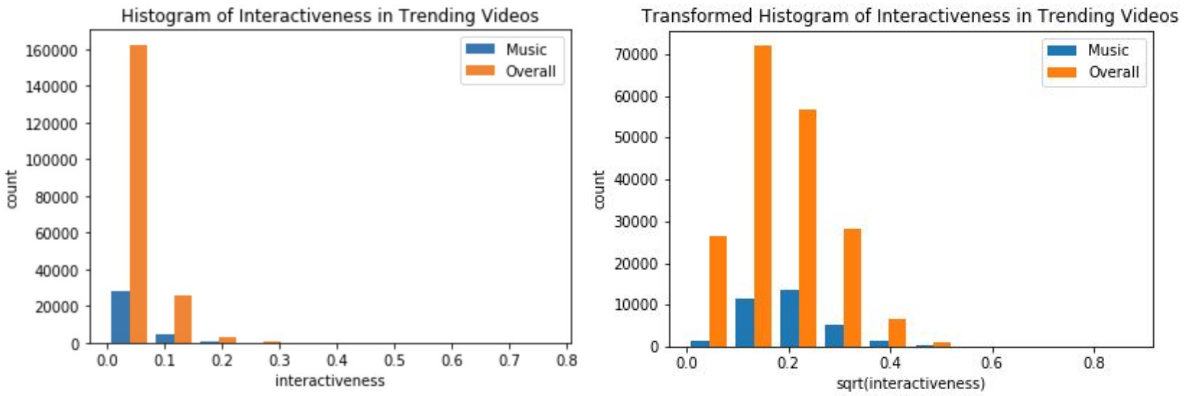
3. Music videos are less interactive than non-music videos

Since we are guessing that certain videos would have fewer viewer interactions, we predict that music videos in trending are less interactive than other videos, because content consumers may add the video on their playlist, and play the video many times without commenting or leaving alike.

First, we define interactiveness as the total number of likes, dislikes, and comments, divided by the number of views. The formula is shown below:

$$interactiveness = (likes + dislikes + comments) / views$$

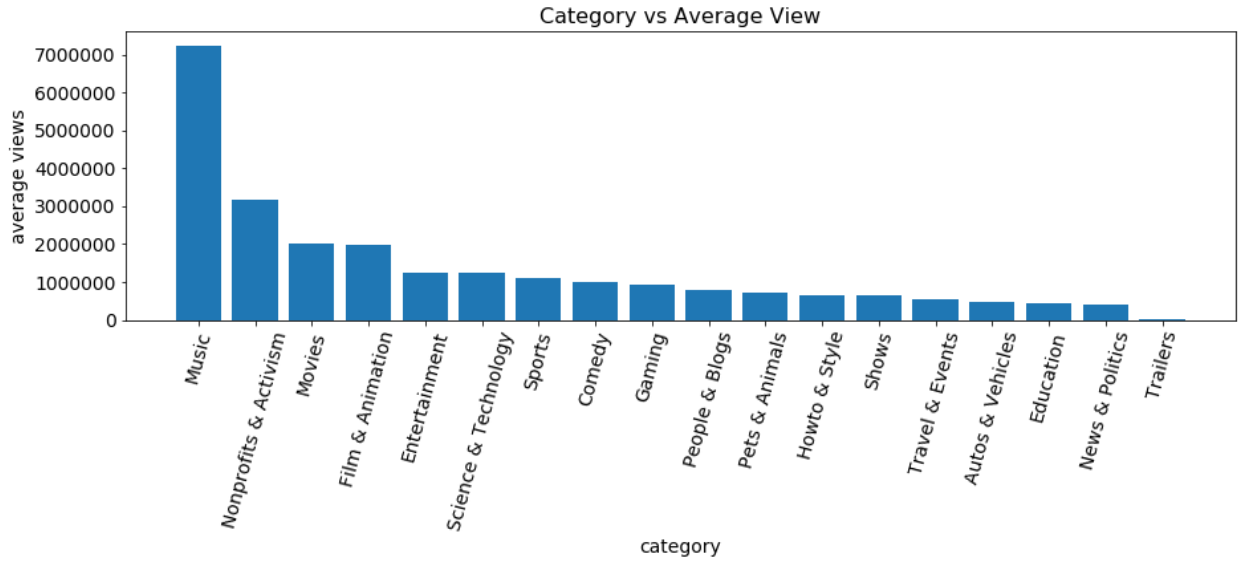
Next, we split the data by category into music and non-music entries. Now we can create the histogram, which looks very skewed initially. So we transform the data by taking the square root. The results are still skewed, but since we have many data points, and the histogram looks reasonably normal, we can assume normality.



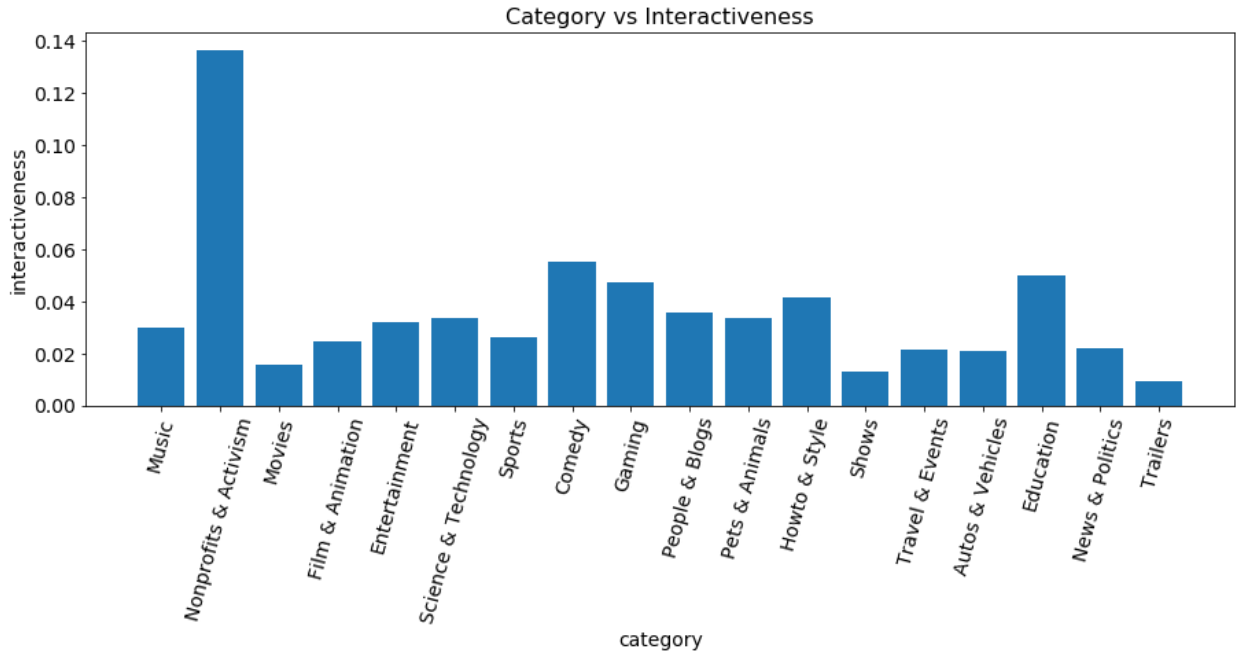
To find out if music videos and non-music videos are equally interactive, we can perform a t-test on the transformed results. However, the test assumes equal variance by default. Thus, we must first perform an equal variance test, which returns a p-value of 4.36e-233. This means that we must set equal_var to false in the t-test. Now, the t-test returns a p-value of 0.0, stating that the two are not equally interactive. This makes sense because the average value for music videos is about 20% higher than the average value for non-music videos. Therefore, we conclude that music videos are (surprisingly) more interactive than non-music videos.

4. How many views does it take for a video to be on trending?

In order to find the number of views to take a video to YouTube trending, our first approach was to compute the average views for each video when they went on trending. We sorted the results in descending order and created a bar plot:



The plot shows that videos in categories need a different amount of views to be selected for trending. The result also indicates that the views may not be the only feature that decides when a video can go on trending. We may also need to consider the other features such as comment count, likes and dislikes. Thus, we continued the analysis by comparing the interactiveness for each category.



Again, the plot shows no discernable pattern to suggest what makes the video go on trending. However, for YouTubers making videos about nonprofits and activism, it may be a good idea to encourage viewers to interact more. By reminding the viewers to

leave a like and comment, the YouTuber may increase the probability of the video to go on trending.

Limitation:

One of the features that we could not include in the *interactiveness* formula was the number of likes for each comment in the videos. It may give us a better understanding of how a video is selected to be on trending.

C. Final discussion

In conclusion, we have cleaned our data and analyzed YouTube trending in different ways. Firstly, we were able to predict the number of comments that a video will have based on the video's statistics with fairly high accuracy (~0.83). Furthermore, we failed to classify all of the video categories. Therefore, we refined our scope to simply classify music and non-music categories only, and we were able to obtain a prediction score of 91%. Additionally, we used the t-test to verify that the music videos were surprisingly more interactive than non-music videos. Lastly, we realized that different video categories might have different views threshold to go on trending. Although we went further to analyze the interactiveness threshold, we could not see any discernible pattern to suggest that the interactiveness would make a video go on trending.

D. Project Experience

[Tu Dat Nguyen]

- Used Pandas to combine and clean dataset to create valid data for statistical analysis
- Conducted statistical analysis using proper machine learning models to predict the number of comments based on the video's statistic with significant accuracy
- Produced the report with graphs and plots generated by matplotlib to present the meaningful statistical analysis

[Jun Wei (Jason) Li]

- Used Pandas to combine and clean dataset to create valid data for statistical analysis
- Defined and refined problem scopes based on analytic results and research
- Produced the report with graphs and plots generated by matplotlib to present the meaningful statistical analysis

[Jacky Lee]

- Identified problems and challenges in combining and cleaning the dataset
- Defined a metric to measure the interactiveness of YouTube videos
- Explored and compared various machine learning models for classification

E. References

Dataset:
Kaggle Trending Youtube Video Statistics | Kaggle

Other YouTube analysis:
What 40,000 Videos Tell Us About The Trending Tab