Project Proposal

Problem to Solve

The Google-owned video-hosting site YouTube provides user-generated content such as original clips, animation shorts, music videos, movie trailers, live streams, and educational lessons. Ranked as the second-most popular site in the world, YouTube claims that one billion hours of content is watched on the broadcast site each day.

With such a strong influence, exposure on the platform has made previously unheard of people and products an instant household name. This phenomenon is thanks in large part to the video site's recommendation capabilities. Some trends are predictable, like when a new song is uploaded from a popular artist or when a new superhero movie trailer is released. Others are more surprising, like a viral cat video that transpires into an overnight sensation. YouTube's Trending tab features videos algorithmically, using a combination of factors, including view count, rate of growth in viewership, the age of the video and other user interactions such as shares, comments, and likes. It is important to understand that trending videos are not necessarily the most-viewed videos within the calendar year.

Videos that are marked as trending are then given a high-profile spot on YouTube's home page and can rack up millions or even billions of views in a short period of time. YouTube's Trending tab is country-specific, which means that a video trending in one region may or may not trend in another.

There are a number of factors to consider when predicting what makes a video popular to one audience and not another. With this context in mind, this project will attempt to solve the following problem:

What factors affect how popular a YouTube video will be in different countries, and using such factors can we predict popularity or trending?

Client

Given that both private individuals and large production companies have used YouTube to grow audiences, there are several clients who would care about knowing the answer to this project's question. In particular, entertainment content creators such as movie studios or musical artists would be interested in knowing how to ensure that their videos reach a wide audience who would then purchase a movie ticket or buy a single to download. Likewise, advertisers who use music in their commercials to sell a product such as a luxury car would want to know which artists are trending to increase the likelihood of viewership for their sponsored video ads.

Thus, based on the findings of this project, the client would decide which factors to either include or exclude from their video posts, depending on the country the client wanted to *trend* in.

Data

The data that will be used for this project include statistics for trending YouTube videos. This data comes from an open-source dataset containing a daily record of the top YouTube videos over a period of several months, with up to 200 listed trending videos per day. More specifically, 16 variables were collected across 10 different regions (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India). The dataset was released on Kaggle and collected using the YouTube API. Click this **link** to access the dataset.

The dataset variables include: a unique video ID, the trending date, the published time, the video title, the channel title, the category (which varies based on region), tags, the video's description, view count, the number of likes and dislikes, whether the ratings were disabled, comment count, whether the comments were disabled, thumbnail link, and whether there was a video error or if the video was removed for violating YouTube's terms of service.

Approach to Solving the Problem

To solve the problem, **factors** and **popular** will have to be defined. Based on the available variables collected in the dataset, "tags", "category_id", and "title" will be **factors**. Note that keyword modeling may have to be done on the "title" variable and the "category_id" variable may have to be recoded as there seems to be some overlap in category types, which vary across region.

"Views", "likes", and "comment_count" will count toward what makes a video **popular**. Multiple time points will need to be analyzed by comparing the "publish_time" to the "trending_date" to see how fast the growth rate occurred. For this project, the "description" variable will not be used due to the long character strings and that doing so could bias the results based on the particular style the content creators used to write them. During the data exploration phase of the project, all 10 regions will be analyzed to see which tags, categories, and titles are best to gain viewers, likes, and comments in a particular country.

Ultimately, the purpose of the project will be to build a model that predicts the number of views, likes, or comments a YouTube video will receive based on these identified factors ("tags", "category_id", "title") when played in each of the 10 regions included in the dataset (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India). For example, if the client uses _____ tags, they will gain _____ views in _____ country.

Deliverables

The deliverables for this project will include code, paper documentation, and a visual slide deck presentation.