Analysis of YouTube Videos

Xianqi Gu

12/17/2020

Introduction

With radical and continuous development in media technology, entertainment and media have become more accessible and interactive for common people. There are more pathways and outlets available for people to share their thoughts, create original content, and become popular and successful in this digital media era. YouTube is an excellent example of these outlets. This project will focus on analyzing YouTube videos specifically. It aims to explore the factors that are predictive of the popularity on the YouTube platform. It also aims to predict whether a YouTube video is successful by examining whether the video can bring profit more than the median weekly income in the U.S.

This report is divided into five sections. In the first section, I would like to specify the statement of problem and present a brief literature review. In the second section, I would like to outline the source of the data used for this analysis. Additionally, the second section will cover the data wrangling for this project. I would like to explain the steps and tools I have utilized to compose the main dataset. In the third section, I would like to describe and justify the methods and tools I have deployed for the exploratory analysis and the statistical learning process. I would like to provide the rationale for using each method and how it contributes to the overall analysis. In the fourth section, this report will present a summary of the exploratory data analysis and the statistical learning model. I will discuss the significance of the top 5 features and the predictive power of the model. Finally, I would like to evaluate the entire project. Plus, this section will cover how this project can be expanded for a more in-depth and comprehensive analysis.

Background

Since 2005, YouTube has been growing rapidly. Currently, YouTube is the second most visited website in the world. It is a website that is based on the original user-generated content. There are more than 2 billion users worldwide. YouTube users watch more than four billion videos every day globally (O' Connell, 2018,). As the demand for original user-generated content surges, more people join the YouTuber family, creating and sharing original videos through their channel. The obsessive enthusiasm for making YouTube video can be justified by the lucrative profits a single video can generate. According to O'Connell, a YouTube video creator can earn between 3 dollars to 10 dollars for every 1000 viewer engagements (2018). YouTube relies on likes, comment, and views to determine which video ranks first (Barnhart, 2020).

However, as the number of videos proliferates, it becomes harder to have viewer engagements. Intuitions tell us that a more popular YouTube video is more likely to be exposed to more views, likes, and comments. As Arthurs, Drakopoulou, and Gandini explained, emerging channels can attract a lot of viewers with videos in the comedy, entertainment, and gaming category. In 2008, the most popular category was news and politics. In 2016, it changed to people and blogs (2018). So, video category is one of the factors that are predictive of the popularity of a YouTube video. But the most popular category has been changing all the time. Thus, the research question of this project is what factors, such as category, make a YouTube video popular and successful in 2020

Besides analyzing the factors correlated with popularity, this project also aims to examine the profitability of a YouTube video. I would like to answer the question that if a person makes a new video weekly (which

is the most common case for a full-time YouTuber), what is the probability that the person can earn more than the median weekly salary in the U.S.

Data

This project uses three main datasets to compose the final dataset. YouTube publishes the trendy video list daily. The first data is downloaded from Kaggle, where it used the YouTube API to scrape the information of the trendy videos. The CSV file contains the trending videos in the U.S. each day from Aug 03, 2020, to Dec 01, 2020, 120 days in total. Some of the important features include video id, publish date, trending date, category id, number of views, etc. The second data file is a JSON file downloaded from Kaggle that contains information such as channel id, category id, and category title. I used the JSON file because it is crucial to associate each category id for every video in the CSV file with the category title contained in the JSON file. It is the category title instead of category id that reflects to which category the video belongs. The third data file is a table containing daily Covid cases and deaths. The rationale for including the Covid data is based on the assumption that people tend to stay at home and spending more time on YouTube videos as the daily cases and deaths increase. I scraped the data from the COVID Tracking Project. This project complies daily number of tests, cases, and death in the U.S. It provides these data to Johns Hopkins for its COVID-19 Insights Initiative. So, the data sources are reliable.

To compose the final data set, first, I created a data frame that contains the matching category id and category title. Second, I merged the newly created data frame with the video data file on category id. Then I dropped the columns that I consider unnecessary. After these steps, I got a data set that contains video id, category id, category, title, publish date, trending date, views, likes, comments, dislikes, and whether the creator enables comments and ratings. Then I used the "datetime" module to convert the publish and trending date from strings to datetime objects. After a closer look at my data set, I found that there are duplicates of video ids. It turns out that they are the same video but over multiple trending days. So, I calculated the days they have been on the trending list. I composed another data set for the information during the trending period, such as the increase in views, likes, and comments. I merged the trending period data with the main data set. Then, I created a new column for the number of days for a video to be trendy because I think it might affect whether it can attract more views. Then, I dropped the duplicates. Based on the literature review, I have created a variable called engagement by aggregating the views, likes, comments, and increases in the three for each video. I included the increases because I think the number of trending days would affect viewer engagement, which is absorbed by the increase in likes, views, and comments.

For the Covid data, I used the "requests" module to scrape the tables for daily cases and daily death from the Covid Tracking Project. I changed the datatype to datetime object for date. I merged the Covid dataset with the main data set on trending date. Right now, the final data set has 4320 observations with 19 features for each observation. The unit of observation is every individual video. The variable of interest is the viewer engagement, which is the outcome in the machine learning model. Feature variables include category title, trending days, cases, deaths, number of days the video took to become trendy, and whether the creator enabled comments and rating. There is no missing value in this data set.

Analysis

For this project, there are three parts of the analysis. The first part is to create some basic visualizations for the distribution and correlations. First, I used "matplotlib" to create a frequency bar chart for the distribution of category titles. Because the videos in this dataset are already trendy, it is reasonable to look at the distribution of categories among all trendy videos and find out the category that appeared the most. Then, the plot can provide a basic understanding of what category is most likely to make the video on the trending list. Also, the rationale for creating this plot is to find the category that appeared the least. That category can be used as a reference group when it comes to machine learning model. I would like to look at the increase in the probability that a video can make more than average weekly income, compared to the reference group.

I used "matplotlib" again to generate a scatterplot, showing the correlation between trending date and the total number of videos that became trendy on that date. I assume that there might be a specific period that makes a video trendy with a higher likelihood. So, this plot is created to roughly test my assumption.

Then, I used "seaborn" to create a heatmap to visualize the variables' correlation. The heatmap is used to examine which variables should be included in the statistical learning models as features.

The second part is about data types and distributions. First, I looked at the data types and distribution of the continuous variables. I used "plotnine" to visualize the distributions. A lot of the distributions are right-skewed, such as the distribution of likes and dislikes. So, I used log transformation on the number of comments, views, increase in comments, likes, views, as well as dislikes. The log transformation is helpful to make the distribution less skewed and meet the statistical inferential assumptions.

After I plot the distributions for categorical variables, I realize there are 15 classes for category variable and two possible outcomes for whether the creator disabled comments and ratings. The two outcomes are false and true. So, I decided to recode the "comment_disabled" and "rating_disabled" variables as dummy variables that take 1 for true and 0 for false. The reason is that a machine learning model cannot differentiate true or false. We have to substitute the outcome with numerical values (1/0). I also created a dummy variable for each category because I would like to look at the differential effect of different categories.

Since my research question for machine learning purpose is whether a full-time YouTuber can make as much as the median weekly earnings in the U.S. , if the YouTuber publishes a new video every week, I would like to convert the engagement variable into a binary variable. This research question constitutes a classification problem. According to the U.S. Department of Labor, the median weekly income of full-time workers is around 994 dollars. So, I will use this number as a threshold of the recoding on engagement. Based on the literature review, 1000 viewer engagement can bring 3 dollars to 10 dollars in profits. On average, the profit per 1000 viewer engagements is 6.5. So, to provide $994 profits, a video needs to have 152923 viewer engagements. This is calculated by dividing 994 by 6.5 and multiplying the number by 1000. So, the outcome variable will equal to 1, if the engagement is above 152923, to 0 otherwise.

Then, I created a machine learning pipeline that contains Naïve Bayes Classifier, K-nearest Neighbors Classifier, Decision Tree Classifier, and Random Forest Classifier. I incorporated these four classifiers to compare them to find the best one with the highest roc_auc score. The AUC-ROC curve is a measure of the performance of the model for a classification problem. It evaluates whether the model can classify the outcomes correctly instead of taking a random guess. The higher the score, the better. So, I used the roc_auc score to sort the model performance the find the best model. For all the classifiers, I specified the fold equal to 5 for the cross-validation process. Cross validation is used to randomly shuffle the data and split it into 5 groups. Each group will be used as a test set, and the remaining groups are the training set. This is useful because we can use all the data for training and testing.

I included the Naïve Bayes Classifier because it is a simple conditional probability model that estimates the probability given a vector of features. The mechanism of K-nearest Neighbors Classifier is different. It counts the average of the K-nearest data points' outcome values to estimate a new data point's outcome. I have specified the K to be 5, 10, 25, 50, respectively. The model will use the 5, 10, 25, and 50 nearest data points, respectively. The decision tree distinguishes the outcome by testing whether the data can pass the nodes in the model. The nodes are the thresholds or set by the model to determine how to classify the data. I have set the nodes to be 2, 3, 4, respectively. So, the number of qualifiers changes. Random forest is used because it can correct the overfitting problem of the decision tree models. It consists of multiple decision trees. It will classify the data in the same way that most of the decision trees classify. So, it can reduce the error made by individual decision tree. I have set the number of nodes for each tree to be 2, 3, 4, respectively. I set the number of trees in the random forest to be 500, 1000, 1500, respectively. I set the max features the models can use from 1 to 18.

After building the pipeline and running the model, I identified the best model that has the highest roc_auc score. I searched for the parameters for that best model. I looked at the performance on the in-sample data and the accuracy score. I examined the importance of each variable regarding the reduction in the roc_auc score if they were to be removed from the model. The reduction indicates how less predictive the model

would be without a feature. Finally, I plotted the partial dependency for the five most important variables to analyze the change in probability of the outcome variable associated with an increase in these features.

Results

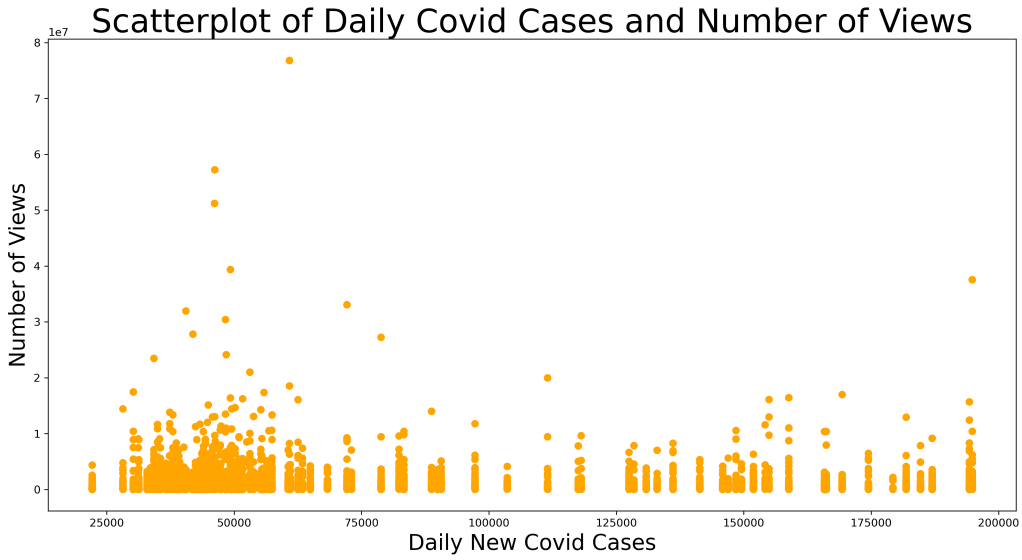

Figure 1: Daily cases and views

Figure 1 shows the distribution of the category variable. From this figure, music is the category that appears the most, while nonprofit and activism is the least. So, we can make a rough statement that it is more likely for a video on music to be trendy than videos on nonprofit and activism. Also, it shows that category is a factor that influences the popularity of a video.

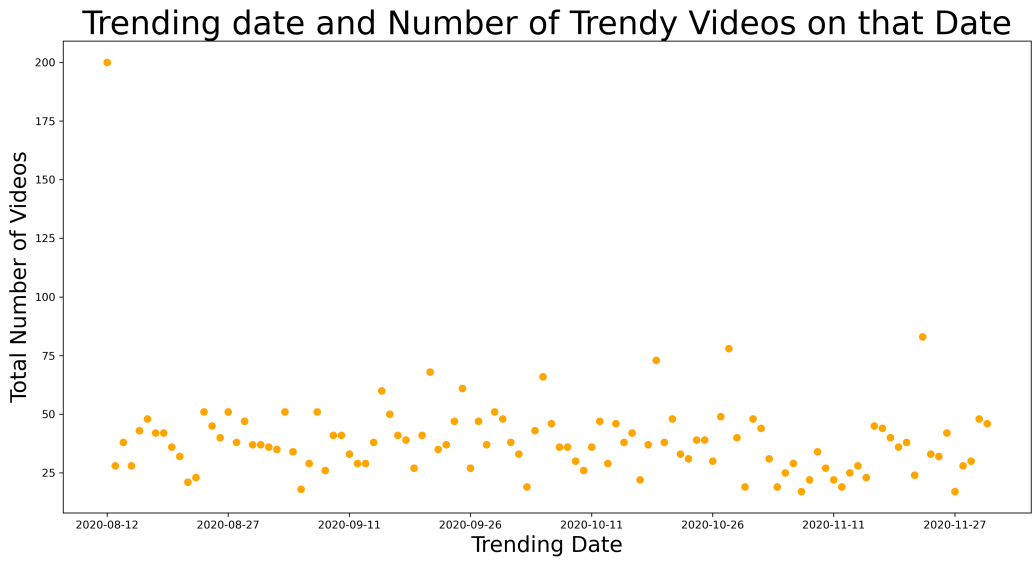

Figure 2: Trending date and Number of Trendy Videos on that Date

Figure 2 is a scatterplot showing the correlation between the trending date and the total number of trendy videos on that date. If I had observed any trend from this graph, it would mean that a specific period can make the videos trendy with higher probability. But there is no trend, meaning there is barely a correlation between date and popularity.

The best model identified has a roc_auc score of 0.78, which means that the model is pretty valid. It is greater than 0.5, meaning that this model performs much better than a random guess. The best model identified is a random forest model that has a depth of 2, uses 18 features, and has 500 decision trees in the model. The in-sample data performance is 0.86, which is very solid. The model's accuracy is 0.98, which means that the model can classify the outcome correctly 98% of the time, although the accuracy is highly dependent on the data set. Overall, the model is very predictive.

Figure 3 shows the AUC reduction if the variable was going to be taken out from the model. The top 5 important variables are the difference in days between the publish and the trending date, the science & technology dummy, the number of trending days, the number of daily Covid deaths, and the people and blog dummy. If the difference in publish and trending date is not in the model, in AUC would reduce by 0.000309. The AUC would reduce by 0.000278 and 0.000257 if without the science and technology dummy and the trending days respectively. The reduction in AUC for the daily Covid deaths is 0.000226 and 0.000021 for the people and blog dummy is 0.000021.

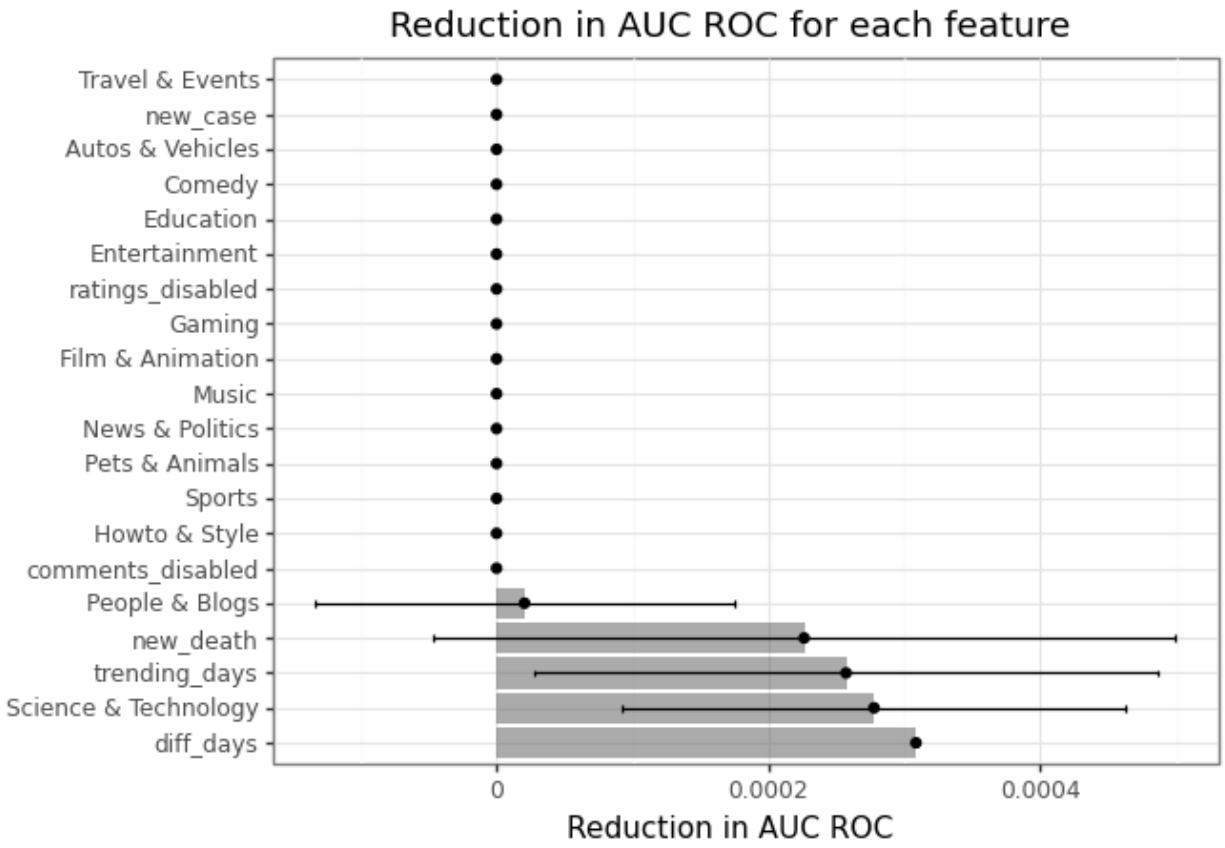

Figure 3: Reduction in the AUC ROC

Figure 4 shows the partial dependency for the top 5 significant variables. For the difference between publish and trending date, if the difference increases from 0 to 15, the probability that the video can generate more earnings than the median weekly income, an outcome I will refer to as probability for following interpretation, would decrease by 15%. But if the days increase from 15 to 25, the probability would increase by 5%. If the video is about science, the probability would decrease by 3%, compared to the nonprofit category. If it is a blog, the probability would not change. The trending-days would increase the probability by 4%, but the impact becomes zero when the trending-days is above 10 days. The increase in probability is quite steady but small as the new deaths increase.

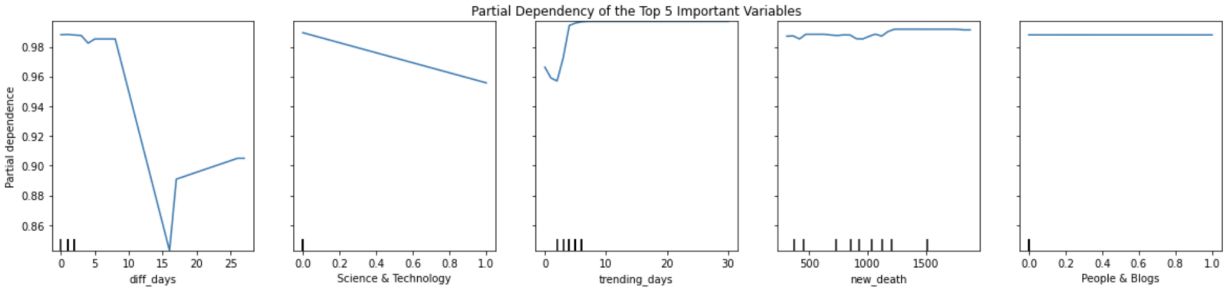

Figure 4: Partial Dependency of the Top 5 Important Variables

Overall, category matters when it comes to the popularity of a video, as shown in the distribution plot and

the partial dependency plot. Time period seems to have no impact on popularity. The model that can best classify whether a video can generate more profits than the median weekly earnings is a random forest model that has 2 qualification nodes, 500 decision trees, and 18 explanatory variables. The model is very accurate and predictive with a high AUC score of 0.86. The most important variable in this model is the difference in days between the publish date and the trending date.

Discussion

I think the project is quite successful. First, it has a clear structure and the analysis proceeds smoothly and reasonably. Every method deployed contributes to the overall question. Every plot provides insights. Second, the final data set is tidy and clean. The final data set can be used for reproductible analysis. Third, the visualizations are of publishable qualities. Fourth, the machine learning model is pretty accurate and predictive.

If more time were given, I would include more data about the videos that are not trendy to include more variation in my data set. The reason why I got such a high accuracy score might be that the videos in my data set are all trendy. My model may not perform well on non-trendy videos.

Word Count: 2999

Citation

1. Arthurs, J., Drakopoulou, S., & Gandini, A. (2018). Researching YouTube. Convergence, 24(1), 3–15. https://doi.org/10.1177/1354856517737222
2. O'Connell, B. (2018, October 12). How Much Do YouTubers Make? Revenue Streams and Top Performers.
3. Barnhart, B. (2020, August 4). How to promote your YouTube channel to maximize views. Sproutsocial.
4. (2020). Retrieved from Covid Tracking Project.
5. Sharma, R. (2020). YouTube Trending Video Dataset. Retrieved from https://www.kaggle.com/rsrishav/youtube-trending-video-dataset?select=US_youtube_trending_data.csv.
6. McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
7. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2. (Publisher link).
8. J. D. Hunter. (2007). Matplotlib: A 2D Graphics Environment, in Computing in Science & Engineering. (Vol. 9, no. 3, pp. 90-95).
9. Waskom, M., Botvinnik,Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, ... Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859
10. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.
11. Chandra, R. V., & Varanasi, B. S. (2015). Python requests essentials. Packt Publishing Ltd.

Note

In the proposal, I specified that I would like to include data from UK. But after careful analysis to decide the outcome variable in the machine learning model, I think it is reasonable to look at one country only because the median weekly income is not the same for two countries. It also makes no sense to compare with the average median weekly income across these two countries. So, I choose to focus on the US only.