Youtube Video TrendingAnalysis Based on Machine Learning

Zhanbei Liu*

GSAS, Columbia University, New York, NY USA 11101

*Corresponding author: zl3070@columbia.edu

Abstract. Social media platforms play an important role in commerce, entertainment, marketing, education, media and communication. YouTube has a large and active user base, making it the center of corporate digital marketing efforts. There are always videos that attract a lot of attention in a short period of time and become trending videos on YouTube, and these videos can be displayed as trending videos on YouTube and updated daily. In this article, we analyzed data on YouTube trending videos in the US region and analyzed the targets that captured the attention of users in a relatively short period of time. We conduct exploratory data analysis on each aspect to gain data insights and find statistical similarities between them to understand viewing patterns across video categories. We present our analysis by measuring, mining, analyzing, and doing ANOVA comparisons of four scales of different categories of likes, dislikes, opinions, and comments. In addition, the study compared the accuracy of three machine learning methods on YouTube data, which counts the number of views and viewer reactions of popular videos on YouTube.

Keywords: machine learning, Video Trending Analysis, exploratory data analysis.

1. Introduction

Social media platforms play an important role in business, entertainment, marketing, education, media and communication. Since its inception in early 2005, YouTube has become the most successful Internet site, offering a new generation of short video-sharing services. Users can share their lives and watch content uploaded by youtube users around the world. It offers a wide variety of user-generated and corporate media videos. Available content includes entertainment videos, TV show clips, music videos, and educational videos ..... .

There are always videos that catch the attention of many people and become trendy in a short period of time. This dataset records the most popular videos on YouTube and is updated daily. The rules for determining whether a video is popular or not depend not only on the number of views but also on how fast the number of views increases, how long the video has been uploaded, user feedback, etc. Popular videos are important information for both the company and the users (i.e. YouTubers.) Youtube can monitor user engagement and usage by looking at popular videos to get a visual idea of where the company is going. When companies increase user engagement and create an environment conducive to participation, they can greatly increase their chances of business success (Kim et al., 361). Therefore, the total number of daily trending video views will reflect fluctuations in user engagement and the future of Youtube in the industry.

Users who upload videos can refer to the Trending video list to improve their content and make their videos go viral. Many YouTubers upload their game videos, edited vlogs, and self-created shows to build their unique channels. They use Youtube as one of their social media platforms to disseminate their interesting and insightful content to viewers and subscribers in the form of a variety of videos. Sharing everyday life experiences on social media creates a sense of belonging and creates a sense of community online. (McCay - Peet). In order to feel a sense of belonging on the Youtube platform, video creators are eager to discover the secret technology that makes their videos trendy so they can gain more subscribers and viewers to watch their high-quality content.

2. Data & methods

2.1. Data

The Youtube Trending dataset originates from the Kaggle dataset, which is composed of detailed information on daily trending videos from August 2020 to March 2022.This dataset contains 22 columns with 116033 observations. The columns contain the unique id of each video, the date, time, and country of a video published, the id of the video category and the id of the video channel, the date and time when the video shows on the trending page, the number of views, likes, dislikes and comments, description, tags, title and channel of a certain video, the link for thumbnails and if the video can be commented and rated by viewers.

In this project, we mainly analyze the trending video in the United States region. Since some videos are trending for more than one day, they may appear more than one time in this dataset with a unique video ID. Hence, we change the date to the date-time format so that it is more convenient to analyze and visualize data. After that, we eliminate the missing values, delete unusual cases that have more likes than view counts, and prepare two versions of data as validation to proceed with our analysis.

To understand the distribution, the dataset is right-skewed with some of the data lying far right compared to the rest of the samples, this illustrates that there are some hot videos way more popular than others based on their views, likes, dislikes, and the number of comments. Since some statistical techniques require data normalization, it is important to do the transformation of the data by applying logarithmic functions to make most of the dataset normalized.

2.2. Models

In this project, after finding the dataset, this study investigates the background and does the literature review. Then, the exploratory data analysis (EDA) is carried out for this dataset, including the ANOVA, ANCOVA, and machine learning to calculate the accuracy. Below is the formula for this project.

ANOVA is used to determine the difference between the means of two or more populations by testing the amount of within-sample variation consistent with the amount of between-sample variation. In this dataset, where each trend video has a category, we will use one-way ANOVA for the study of one factor. ANCOVA, or analysis of covariance, is an extended form of ANOVA in which the effect of one or more interval-scale uncorrelated variables is removed from the dependent variable prior to conducting the study. It is the midpoint between ANOVA and regression analysis, allowing for the comparison of one variable across two or more populations while taking into account the variability of other variables. It is used in this paper to quantify the differences between different categories of opinions, likes, dislikes, and comments.[11]

Then we use three machine learning methods, decision tree, Lasso, and ridge, to compare the best accuracy for predicting the number of views of each trending video. Decision trees are decision support tools that use tree-like decision models and their possible outcomes, including chance event outcomes, resource costs, and utilities. Lasso and ridge are two regularization techniques used in feature selection as the shrinkage method for penalization.

Below is the formula that used for this paper:

$$Z=\frac{x-min(x)}{max(x)-min(x)} \tag{1}$$

To calculate the accuracy for the machine learning models, following expression is used:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{2}$$

3. Results & Discussion

Exploratory Data analysis (EDA) is a method of manipulating data and summarizing its key characteristics by exploring and understanding relationships and trends between variables. We'll use the visualization package in Python to study the distribution and properties of the data.

If people want to be a video creators and look for a head start on Youtube, it would be better to check which category is most popular among all with respect to the view count.

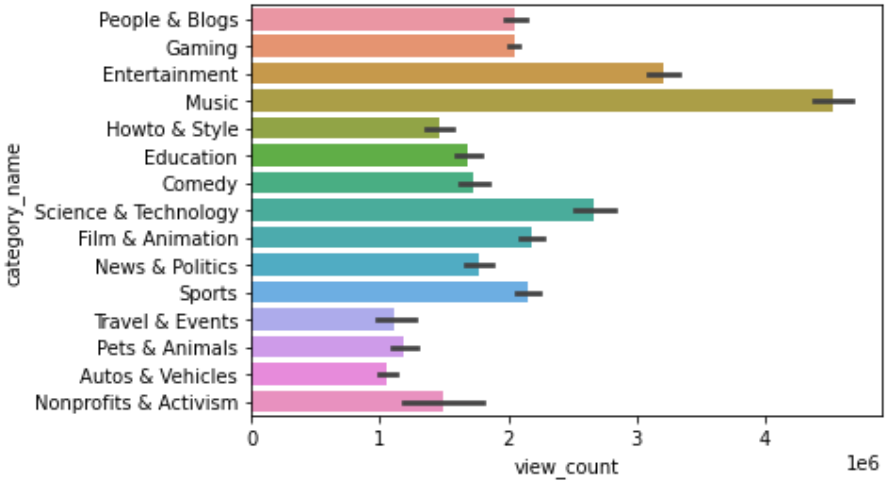

Figure 1. Number of Views in each Category

The graph shows the total view counts for each category. It is easy to find from figure 1 that the music and entertainment categories have the highest number of views and autos & vehicles have the smallest number of views.

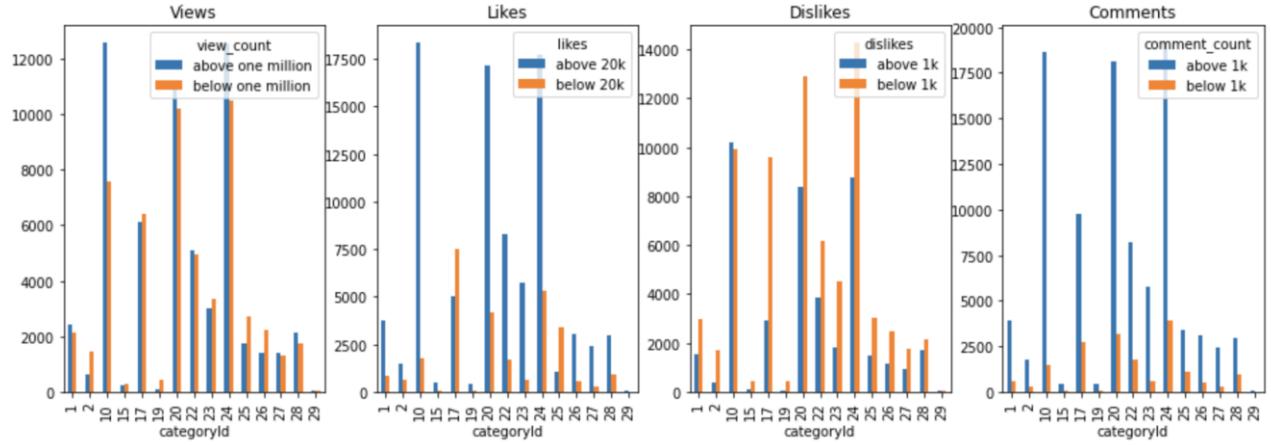

Figure 2. Histogram of 4 variables in each category

From figure 2, we classify each category into 2 parts. For the number of views, we separate the data into above one million and below one million, for likes, we separate into above 20,000 or not, for both dislikes and comments, we separate by 1,000. It is easy to find that with more views, there are more likes and comments and fewer dislikes. Also, the category of 10,20, and 24 are the most popular one figure 1, which are music, gaming, and entertainment.

ANOVA is a very useful tool to test if there are significant differences between groups. In this dataset, every trending video is marked with its category, we will first try One-Way ANOVA on categories with likes, dislikes, and view count and all F observed values are very large with p-values less than 0.05. Therefore, we can conclude that different groups marked with different categories have significant differences among them.

Table1. One Way ANOVA

index	sum of square	df	F
view_count~category	3.4607216564772982	14.0	265.8106390206175
Residual	88.08900809837485	94723.0	NaN
likes~category	87.63139676319422	14.0	1572.5610740231316
Residual	377.03322364290415	94723.0	NaN
dislikes~category	153.06010365163078	14.0	341.09693148239154
Residual	3036.0687325498493	94732.0	NaN

Checking assumptions for ANOVA:

The normality of data has been checked in the previous part.

Since we have a large dataset, we can apply CLT to this dataset and we treat the condition satisfied as n is large enough

ANCOVA is a strategy that combines linear regression and ANOVA. We have proved that there are significant differences between each category with different features including view counts, likes, and dislikes. Therefore, to quantify the differences between them, we use ANCOVA to adjust the mean for every feature with respect to their categories.

Since we have over 10 categories, we choose three categories that have similar view counts but different genres that may have different audiences. Moreover, since the view count, likes, and dislikes are large numbers without normal distribution, we log them to move on to our analysis.

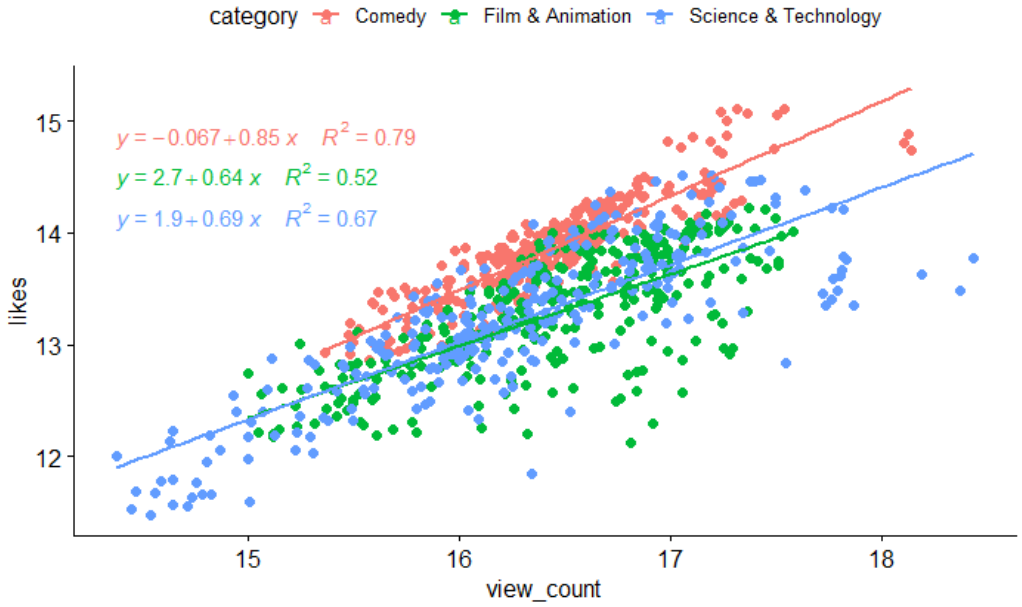

Figure 3. ANCOVA on likes vs. view count

From the results in figure 3, we can see that they basically have similar slopes with different intercepts. The category Film&Animation and Science$Technology have a very close slope compared with each other. Therefore, ANCOVA will be a good model for predicting view count by likes in different categories with different means. We can also see that there are positive correlations between likes and view count. From this result, we observe that likes will increase as view count increases, and trending videos generally have very large view counts because more audiences favor the content. Comedy with the largest slope shows that more viewers will click likes after watching than the trending videos from the other two categories.

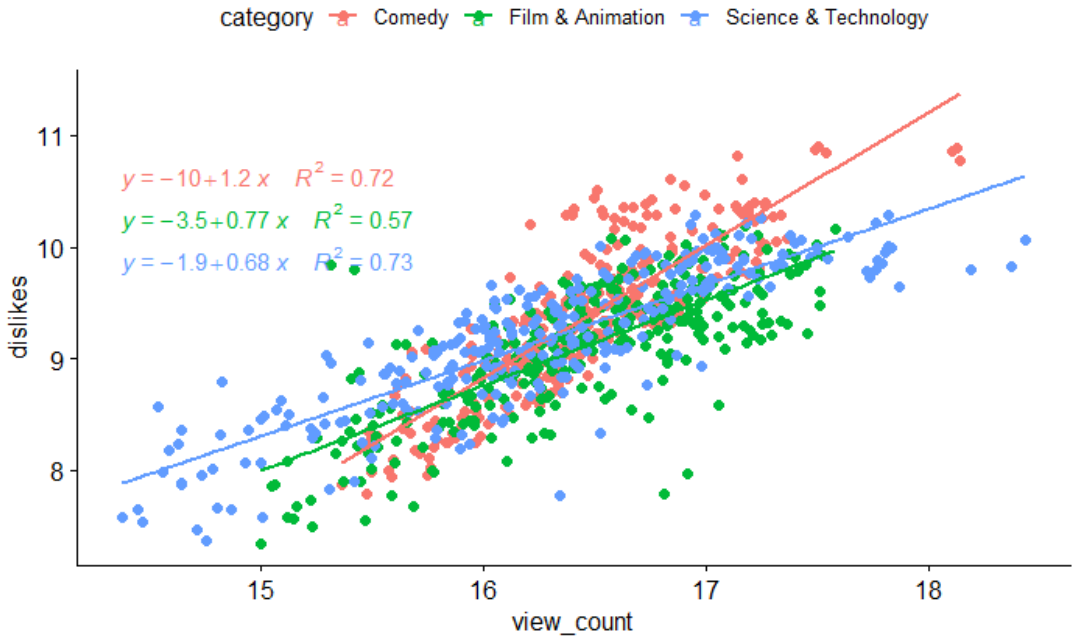

Figure 4. Dislikes vs View Count

We can see that Film&Animation and Science&Technology have a close slope shown in figure 4, similar to the previous ANCOVA on Likes vs. View count. But we observe that Comedy has a much larger slope of 1.2, the largest slope among the three categories. Therefore, we conclude that a larger view count will lead to an obvious increase in both likes and dislikes. The Comedy category has more significant effects on feedback than the other two categories.

After a Youtube video is published, it generally trends for several days and fades out gradually. We know that if a Youtube video performs well on the trending page, this video will be revealed by more people and last longer on the trending list. In this research, the author first extracts the maximum values for the other predictor variables and then creates a new variable total_trending_days to evaluate the number of trending days for each video. We are still using the 70%-30% split to create the training and testing data. We want to find the best regressor for this model to predict the number of views. The predictors for machine learning are the categoryId, likes, dislikes, comment_count, comments_disabled and ratings_disabled. The three regressors are the decision tree, Lasso, and Ridge. For the three machine learning methods, the research uses both the default parameter. To evaluate each regressor, the accuracy is a good scale for the performance. Table 2 below is the matric for the 3 regressors with their corresponding accuracy. It is easy to find that the decision tree regressor is the best with the highest accuracy.

Table 2. Three Machine Learning Method with Accuracy

Method	Accuracy using default parameters
Ridge	79.51%
Decision Tree	87.77%
Lasso	79.60%

4. Conclusion

Through this project, we acquire the results of analyzing Youtube trending video data from August 2020 to March 2022 and thus propose some conclusions and assumptions in our report:

We successfully use the multi-linear regression model to imply that variables likes, dislikes, comment_count, and published_hour have a significant influence on the views of a trending video. YouTubers can improve their video quality based on these aspects of their video feedback. We

successfully utilize the 3 machine learning methods to compare the best model accuracy in predicting the number of views for a Youtube video and the best model is the decision tree method. The top 3 popular categories of trending videos are Music, Entertainment, and Science&Technology. Users could refer to this list to follow the trending content, and YouTube could fund to simulate more high-quality videos in other categories. Understanding these results will not only help YouTube develop better features and algorithms to improve user engagement and earn profits but also benefit YouTubers to improve their video quality and content in order to compete for opportunities of having more trending videos.

This project still can have lots of improvements. Like people can use R first to do the forward and backward selection to find the best model to predict the quality and then use Python to do the above. Also, people can calculate not only the accuracy score to compare the model but the MLE and MAP.

Since we use likes and dislikes to predict the number of views, for some videos, it occurs some people are not objective in evaluating the video, and also there may exist some fans to click a video many times to give a better number of views for their celebrities. Also, we only use 10 variables in this paper, the factor of influence are more than that mentions in this paper. Moreover, since we use the data from 2020 to 2022, covid-19 may influence the youtube video data.

References

[1] jgolani2. “EDA and ML Insights on Youtube Trending Dataset.” Kaggle, Kaggle, 3 Feb. 2022, https://www.kaggle.com/code/jgolani2/eda-and-ml-insights-on-youtube-trending-dataset/notebook.

[2] Kim, Young Hoon, Dan J. Kim, and Kathy Wachter. "A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention." Decision support systems 56 (2013)

[3] McCay-Peet, Lori, and Anabel Quan-Haase. "A model of social media engagement: User profiles, gratifications, and experiences." Why engagement matters. Springer, Cham, 2016.

[4] Sharma, Rishav. “YouTube Trending Video Dataset (Updated Daily).” Kaggle, 3 May 2022, https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset?resource=download.

[5] “YouTube Statistics 2022.” Official GMI Blog. https://www.globalmediainsight.com/blog/youtube-users-statistics/.

[6] H. M. M. Caldera, S. Perera, G. S. N. Meedin and I. Perera, "Classification of Trending Videos in YouTube," 2021 From Innovation To Impact (FITI), 2021, pp. 1-6, doi: 10.1109/FITI54902.2021.9833039.

[7] Castillo-Sánchez, J. J., Mejuto, J. C., Garrido, J., and García-Falcón, S. Influence of wine-making protocol and fining agents on the evolution of the anthocyanin content, colour and general organoleptic quality of Vinhão wines. Food Chemistry, 97(1), 130-136 (2006)s.

[8] X. Cheng, C. Dale and J. Liu, "Statistics and Social Network of YouTube Videos," 2008 16th Interntional Workshop on Quality of Service, 2008, pp. 229-238, doi: 10.1109/IWQOS.2008.32.

[9] G. M. H. C. Gajanayake and T. C. Sandanayake, "Trending Pattern Identification of YouTube Gaming Channels Using Sentiment Analysis," 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), 2020, pp. 149-154, doi: 10.1109/ICTer51097.2020.9325476.

[10] Choe, M.G., Park, J.H., Seo, D.W. (2019). How Long Will Your Videos Remain Popular? Empirical Study of the Impact of Video Features on YouTube Trending Using Deep Learning Methodologies. In: Xu, J., Zhu, B., Liu, X., Shaw, M., Zhang, H., Fan, M. (eds) The Ecosystem of e-Business: Technologies, Stakeholders, and Connections. WEB 2018. Lecture Notes in Business Information Processing, vol 357. Springer, Cham. https://doi.org/10.1007/978-3-030-22784-5_19

[11] Amudha, S., Niveditha, V. R., Kumar, P. R., Revathi, M., & Radha, S. (2020). Youtube Trending Video Metadata Analysis Using Machine Learning.

[12] Reed, C., Elvers, T., & Srinivasan, P. (2011, August). What's trending? mining topical trends in UGC systems with YouTube as a case study. In Proceedings of the Eleventh International Workshop on Multimedia Data Mining (pp. 1-9).

[13] S, Surbhi, et al. “Difference between ANOVA and Ancova (with Comparison Chart).” Key Differences, 11 Jan. 2017, https://keydifferences.com/difference-between-anova-and-ancova.html.

[14] Team, DataCamp. “Lasso and Ridge Regression Tutorial.” DataCamp, DataCamp, 25 Mar. 2022, https://www.datacamp.com/tutorial/tutorial-lasso-ridge-regression#introduction-to-lasso-regression.