Analysis of Trending YouTube Videos: Finding Patterns in Viral Content

Vincent Park

9/28/2022

1. Abstract

As the digital world continues to grow, content creators frequently have trouble building a community and producing videos that will interest their audience. Especially as these people look toward the internet for both recreational and monetary reasons, finding out techniques to build a community is important in today's age. This paper analyzes the issues of video performance, revealing the patterns of what makes a video successful and viral. By training different models and testing different datasets, we were able to find the correlation between the potential chances of popularity and the video's content. Using the most accurate model, the Random Forest model, content creators can see whether or not they are likely to do well based on patterns found in trending videos.

2. Introduction

As content creation becomes more prominent, channels seek out more methods to grow their audience and create videos that will attract users. YouTube, with its hundreds of millions of daily users, shows how competitive content creation may be and how people may feel a lack of recognition for their hard work. To help smaller content creators get started, this paper analyzes how a video may perform, which would be a good indicator of how to grow and attract an audience. We will discuss the likelihood of a video to trend based on previous patterns in the trending page and its performance in relation to other trending videos. To find this likelihood, we will be solving a supervised learning regression problem since the dataset used is a collection of trending videos with numerical values. With our dataset, we train and test several different models. We are then able to find patterns within the data and predict how likely a video would trend based on the results from the models.

3. Background

For prior research, I consulted "Recurrent Neural Networks for Online Video Popularity Prediction," by Tomasz Trzcinski which demonstrates a method of proving a video's predicted popularity. In this paper, it analyzes a video frame by frame and compares trending popular content with the video's content, predicting how the video will perform based on the content before publication. This approach demonstrates a common method to analyze videos through patterns and tags, showing that videos with similar content typically have similar performance.

This academic paper uncovers how different types of content are favored or watched more than others, revealing how users' interests change. Since we want to capture the user's attention as quickly as possible, the paper also invites content creators to keep their videos interesting throughout the entire video to garner more watch time, resulting in more exposure.

4. Dataset

In the project, we will be using a dataset containing 40949 trending YouTube videos from 2017 to 2018.

Since we will be using numerical data, we need to manually transform features in our dataset into numbers; for example, I associated each channel with an identification number in order to train its data later on. In addition, I removed unnecessary textual features such as "description" to clean up the data further. Although the text description of a video could be a helpful predictor, we wanted to stick to a regression based approach for this project. After preprocessing the data and getting rid of non-numerical values, there were a total of 8 features that described each video and its performance through like, dislike, and comment counts. We added one more feature that measured a video's popularity score by using an equation that factors in both views and likes. A small sample of our dataset is shown in Figure 1.

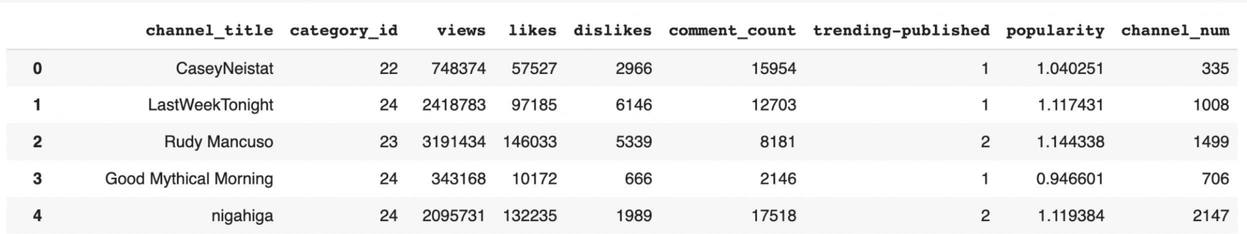
Views and likes are two very different ways to tell if a video is popular, so we decided to design a popularity equation based on both, as shown here:

$$0.7\left(\frac{log(v+1)}{\log\left(v_m\right)}\right)+0.3\left(\frac{log(l+1)}{\log\left(l_m\right)}\right)$$

where v is the number of views, v_m is the median number of views,

l is the number of likes, and l_m is the median number of likes

Figures 2 and 3 demonstrate that there is a good spread of popularity scores across the different video categories and all the videos as a whole. After the data is finally preprocessed, we separated the dataset into two sets: one set with comments and one set without comments factored in. This is because we noticed a strong log-linear relationship between comment count and popularity score, as seen in Figure 4. Each set had its data split into 80% training and 20% testing.



	channel_title	category_id	views	likes	dislikes	comment_count	trending-published	popularity	channel_num
0	CaseyNeistat	22	748374	57527	2966	15954	1	1.040251	335
1	LastWeekTonight	24	2418783	97185	6146	12703	1	1.117431	1008
2	Rudy Mancuso	23	3191434	146033	5339	8181	2	1.144338	1499
3	Good Mythical Morning	24	343168	10172	666	2146	1	0.946601	706
4	nigahiga	24	2095731	132235	1989	17518	2	1.119384	2147

Figure 1: Sample of the dataset

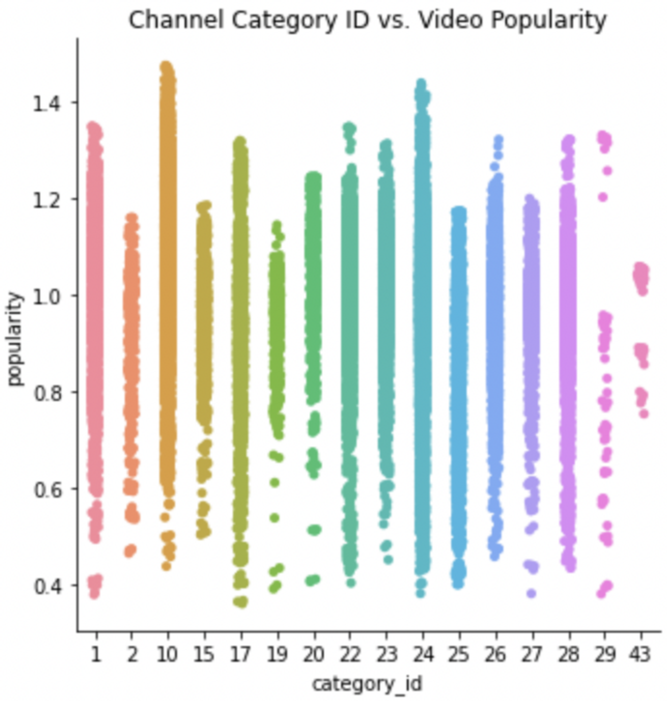


Figure 2: Categorical plot showing popularity score vs. video category

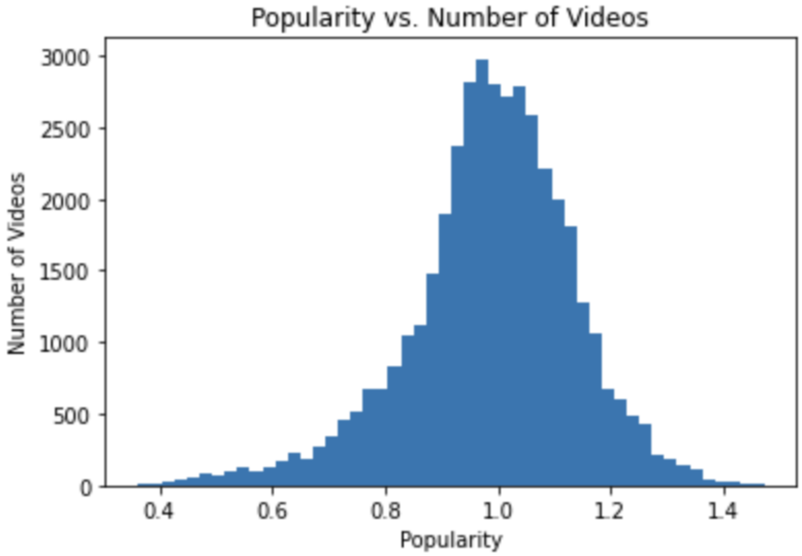


Figure 3: Histogram showing the distribution of popularity scores

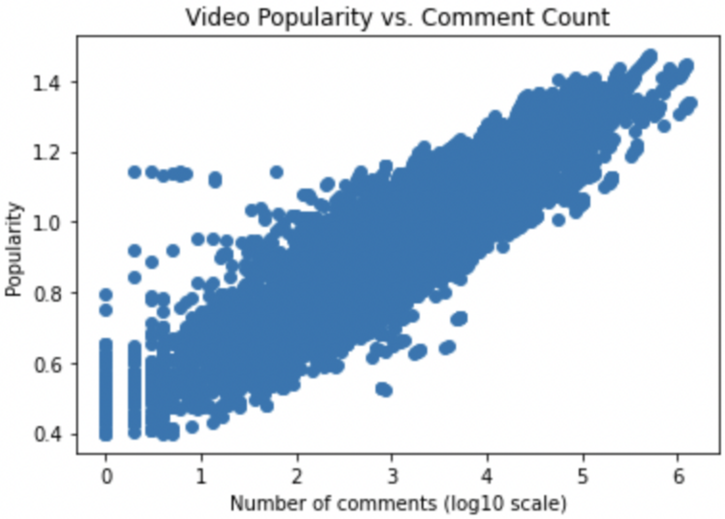
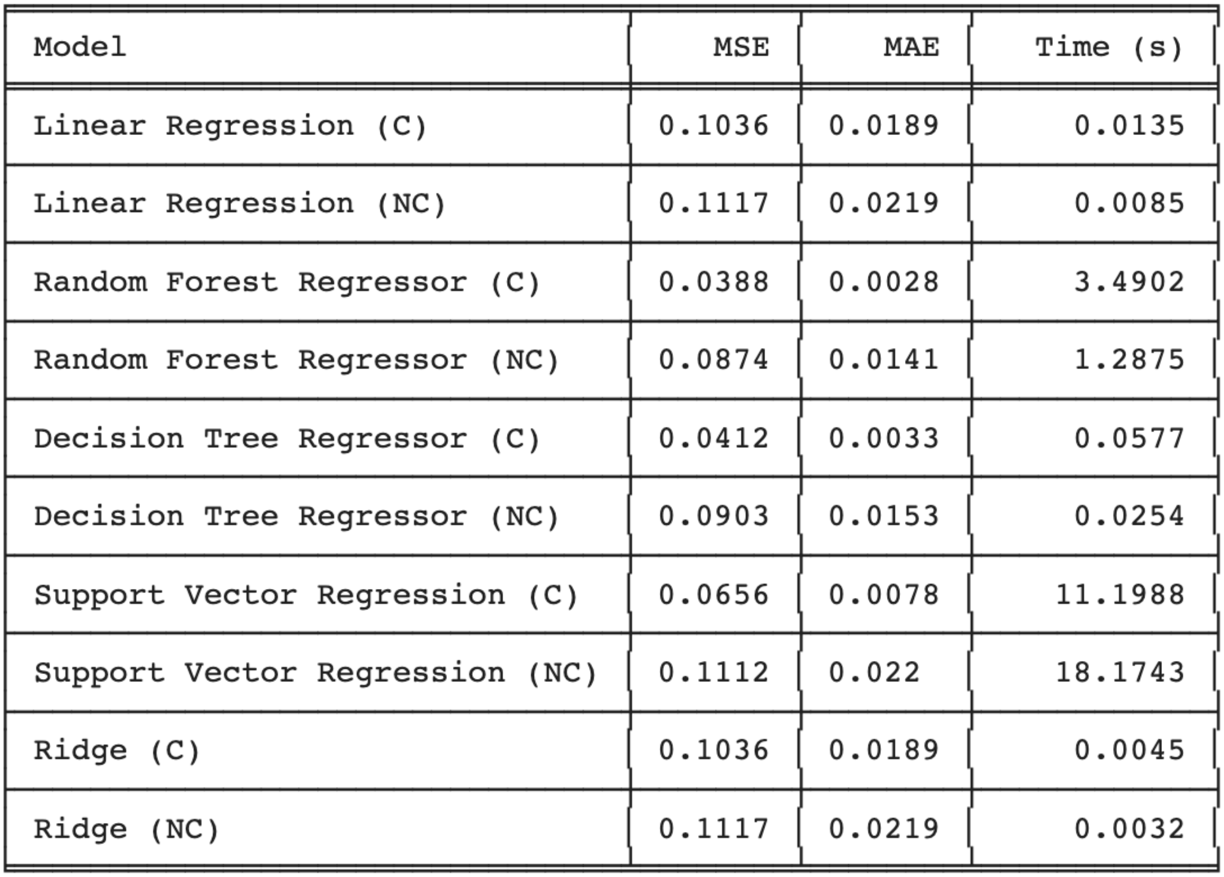


Figure 4: Scatterplot showing the relationship between popularity score and comment count

5. Methodology/Models

To solve our research problem, we used a total of 5 different machine learning algorithms to predict the performance of a video. Picked by their prevalence in machine learning, the five models that were used were Linear Regression, Random Forest Regressor, Decision Tree Regressor, Support Vector Regression, and Ridge Regression. With the popularity score, genre, and channel identification of each video, the models were evaluated using two metrics: mean absolute error and mean squared error, which would show how accurately the models could predict the popularity score of each video. Visually, the metrics would show how close each video in the test dataset is to the best line of fit of all the trained videos. In the table below, we show the results of each model, one with and without comments, and the amount of time it took to run. We included this time measurement because if this model were scaled up and trained on millions of YouTube videos, the time it takes to train a model would be a serious concern.



Model	MSE	MAE	Time (s)
Linear Regression (C)	0.1036	0.0189	0.0135
Linear Regression (NC)	0.1117	0.0219	0.0085
Random Forest Regressor (C)	0.0388	0.0028	3.4902
Random Forest Regressor (NC)	0.0874	0.0141	1.2875
Decision Tree Regressor (C)	0.0412	0.0033	0.0577
Decision Tree Regressor (NC)	0.0903	0.0153	0.0254
Support Vector Regression (C)	0.0656	0.0078	11.1988
Support Vector Regression (NC)	0.1112	0.022	18.1743
Ridge (C)	0.1036	0.0189	0.0045
Ridge (NC)	0.1117	0.0219	0.0032

6. Results and Discussion

The results for each model showed a very small error. For the results of the mean squared error, the Ridge Regression model and the Linear Regression model showed the worst error, having around 0.11 for the testing set with comments and 0.10 for the testing set without comments, and the Random Forest Regression model showed the best error, having around 0.03 for the testing set with comments and 0.08 for the testing set without comments.

Across all the models, the mean absolute error was usually better than the corresponding mean squared error. Once again, the Ridge Regression model and the Linear Regression model both had the worst errors, having around 0.021 for the set with comments and 0.018 for the set without comments, and the Random Forest Regression model had the best error values: 0.02 and 0.01.

Also, it's important to note that the Support Vector Regression model took the longest to calculate, needing almost 10 to 20 more seconds to train the entire model for the set with and without comments. Most of the other models took less than a tenth of a second to train, and the Random Forest Model took only a couple of seconds. Overall, the results of all the models demonstrate that there is definitely a relationship between the popularity of a video and its genre and channel.

7. Conclusion

The models all show a relationship between a video's genre, channel, and popularity, showing that type of content and previous history can help in making a successful video. This data not only shows concrete numerical evidence that certain types of videos do better than others, but it serves as a guide for smaller content creators on how to grow their channel using discovered patterns found within the trending tab. These results can also prove how certain factors in videos, such as likes, comments, or especially views, can greatly affect how viral a video may be; although the content is usually a true reflection of the quality of the video, its statistics are a better reflection of the video and the likelihood of virality. I think all my models performed well because certain types of videos that are related in some way tend to perform similarly; for example, cooking videos, especially by the same channel, will have around the same number of views and likes as the particular audience is usually attracted to the same types of videos. To develop my research further, I would analyze individual videos with frequently replayed points to specifically see what types of content users look for. I might also try to use a natural language processing approach to processing the text description data of each video to see if there are any interesting patterns between the description of a video and its trending potential. In addition, I think broadening the studied platform would show how users' demonstrated interests change over time, especially in popular platforms like Instagram and TikTok. This difference would also show how different types of content become viral all the time.

8. Acknowledgements

I would like to thank Amanda Wang for guiding me in the coding process, research paper construction, and advisory in my research topic. I would also like to thank Inspirit AI for the opportunity to participate in the Independent Project Mentorship program. This paper would not have been possible without their generous help.

9. References

[1] Trzcinski, Tomasz, et al. *Recurrent Neural Networks for Online Video Popularity Prediction.* Warsaw University of Technology, 2017.