YouTube Popular Videos

Written by: Codie James of Brainstation

Date: Dec 11, 2022

Objective

The objective of this report is to answer the question of which features have the most importance in relation to the success of a YouTube video.

Summary

The data used in this report was collected directly from YouTube and later uploaded to Kaggle. Kaggle is a data repository that can be accessed by anybody. The data itself has just over 40,000 different data points from different videos containing statistical information for each. The information I am working with is a snapshot of a certain time and is not continually updated.

Importance

I believe that by answering this question I will be able to advise which numbers have statistical significance and should be targeted. This can also be used to see the most growth, lower costs and gain the most exposure possible. Previously there has been a lot of theory crafting on what exactly makes a video popular and I hope to answer some of this question.

Analysis Methods

During my analysis I first had to clean up some of the data and remove features that were not important to the question. Some of these features include:

- Video ID
- Thumbnail URL
- Publish time

These are largely random upon creation or can not be used to predict a video's success by views. The next step I took was to leave only the columns of data that contained numerical values to be used as prediction metrics. With this data I did not have to worry so much with missing values as the data was very robust in nature.

Modelling Insights

Exploring the data I had to work with I had a few columns of interest:

- Views
- Likes
- Dislikes
- # of comments

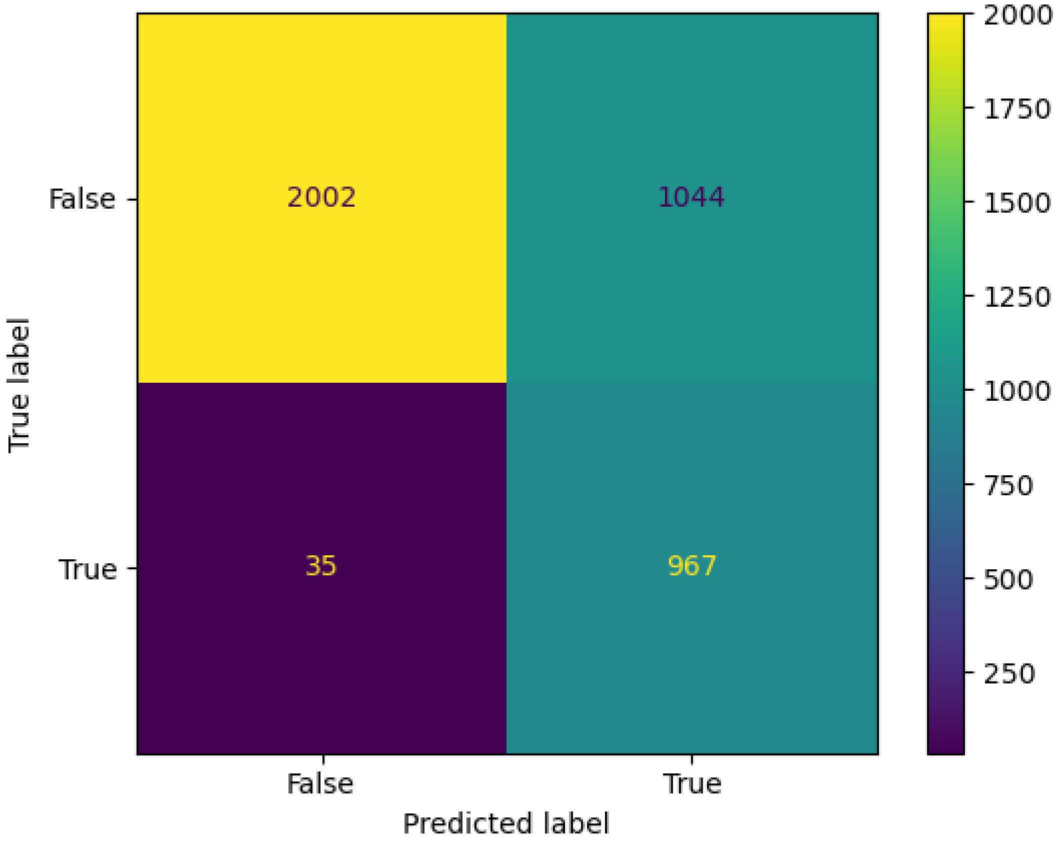
I decided to create a ratio of like to dislike where a positive number means more likes and a negative number means more dislikes. This is important in discovering if likes or dislikes have more of a factor on video views.

Model Choice and Results

For my results I decided to employ 3 different methods:

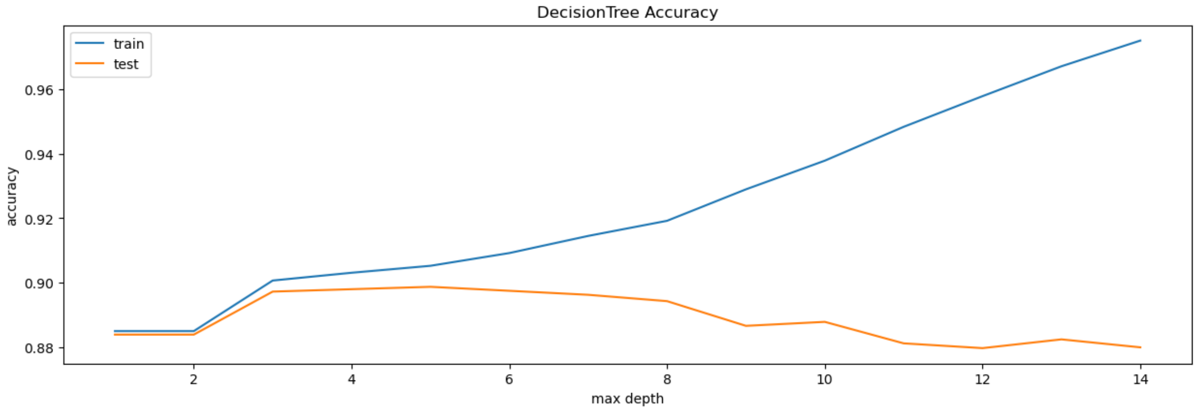
- Logistic Regression
- K-Nearest Neighbors
- Decision Tree Classifier

I decided to use multiple types of machine learning techniques in order to get a more robust and accurate model. I started with a logistic regression which predicts a yes or no outcome based on information you give it. I chose to target videos with at least 1 million views as my definition for success. Using this method resulted in a accuracy of appox.~ 91.2%.



Above our figure shows the predictions as True and False. Our correct predictions include the 2002 and 967 which were predicted correctly as both videos with over 1 million views and below 1 million views. Our values of 35 and 1044 are the predictions the model did not get correct.

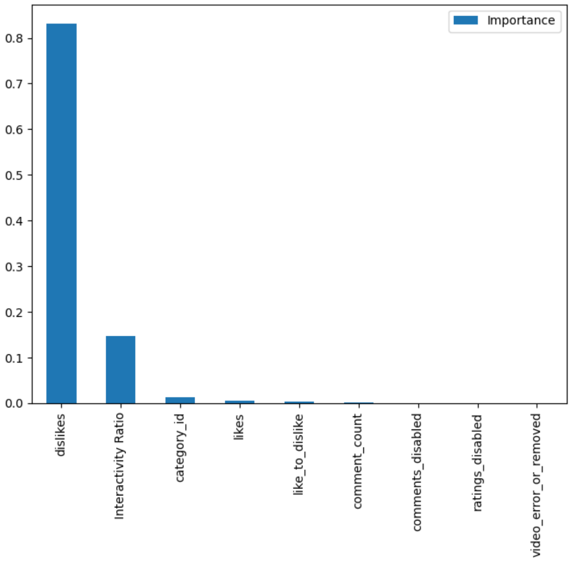
Here is an example of my testing and optimization of my models:



The above graph is showing how many features are used to determine a video's success. Based on the graph the best amount of features to use in this case would be 3. The reasons for this would be because the accuracy between both the training data and the testing data is closest together but still benefits from the best possible accuracy. Using a higher number of features results in the model being over-fit, this means it will do poor on data it has not seen before.

Findings and Conclusion

After going through this data I can conclude that there are a few features that are more important to look at to determine if a video will be successful. These can be used as a rule of thumb and should be targeted when creating a video on YouTube.



In order to produce a successful video it is important to have the lowest amount of dislikes possible. This will encourage the audience to share the video with family and friends. Having more interaction on each video will also net more viewership as YouTube will see this as good and recommend the video to more people.

Key points to target:

- Fewest Dislikes possible
- Encourage user interaction
- Target popular Categories
- Have more Likes than Dislikes