The Anatomy of Successful YouTube Content: A Metadata-Based Study of Factors Contributing to Channel and Video Success

Manan Patel
University of Tennessee, Knoxville
mpatel65@vols.utk.edu

Jake Shoffner
University of Tennessee, Knoxville
jshoffn3@vols.utk.edu

***Abstract*— We addressed the challenge faced by online creators who rely solely on their YouTube income, which heavily depends on the popularity of their videos to generate a satisfactory income. To increase their chances of success, creators must regularly analyze trends and data to identify the most profitable content to create. Our work focused on utilizing predictive modeling techniques such as regression and classification to identify key variables such as video category IDs, view counts, and channel subscribers, and identified the most influential factors on our models.**

Keywords— YouTube, Metadata, Machine Learning, Predictive Models, Regression, Classification

I. INTRODUCTION

We identified the particular variables that influence the popularity of videos using machine learning models, which can then help the content creators focus on certain variables more than others while making their content. We utilized a data-driven approach to answer questions such as what would be the amount of views that a video gathers based on watch time; what variables influence the amount of subscribers that a channel has.

Furthermore, we implemented predictive modeling techniques such as regression and classification to answer these questions. First, we used regression models, linear and multiple regression, to figure out which variable is most effective in predicting video view count and subscriber count. We then used decision trees and random forest models for predicting video category IDs and classification ranges of video view counts. We then presented results obtained from these models in this paper.

II. METHODOLOGIES

1. Data Collection

We started our project by finding a dataset that can give very detailed insights about YouTube channels and videos. Using Kaggle, the "YouTube Videos and Channels Metadata" dataset was a prime target for our research [1]. This dataset contains over 500,000 unique video entries on the YouTube platform, containing 26 total features that helped us build an understanding of how videos and channels become successful on the platform.

2. Preprocessing

We preprocessed our data using the Pandas library for any entries with NA values, or in our case, negative one values. We did this type of data cleaning for all features of a subset for a specific machine learning model, to only remove rows that have these undesirable values. This is to ensure that as much data as possible is preserved, and we only clean the rows that absolutely need to be removed for our models to produce valid results.

3. Feature Selection

We performed feature selection in three ways. One, by using Pandas to do simple statistical analysis of features. Second and a bit better, is to use all available features in a decision tree or random forest model and then look at the visualization of the most important features. We learned the most impactful features proportional to each other, making it easy to see which features create the best model. Lastly, by using a heatmap for our regression models, we found what features correlated the most with what other features. This is the simplest and best approach for feature selection in general and what we would recommend moving forward.

4. Predictive Modeling

We discovered the possibility of predicting: video view counts, video category IDs, and channel subscriber counts from the information provided in our dataset. Not only this, but we want to know how accurate our predictions are and what features are the best contributors for making these predictions.

We harnessed the power of linear regression and multiple linear regression from the scikit-learn python library for predicting video view counts and channel subscriber counts. These are great models for predicting quantitative variables such as these.

We used decision tree and random forest models from the scikit-learn python library for predicting video category IDs and classification ranges of video view counts. To clarify, we used classification instead of regression, meaning that the video view model, while accurate, is not precise in determining the video views since the views are in a wide range of numbers.

5. Evaluation

In evaluating our machine learning models, it is important to know what the best way is to measure their success.

With our linear regression and multiple linear regression models, we evaluated the R squared (coefficient of determination) values of each model. As for the most important features, we used a heatmap and tested models with different features against one another to see which model has the best R squared value of the group.

With our decision tree and random forest models, we compared our predictions, that used test data predictors as input, against our target test data. To understand the most important features in our model, we used the built-in *feature_importances_* data from scikit-learn.

III. Results

The predictive models, their accuracies, and their most important features are outlined below.

1. Linear and Multiple Linear Regression Models

a. Predicting Video View Counts

We discovered through a heatmap that views/elapsedtime was one of the better predictors for a regression model in predicting view counts. Through this model, we achieved an R squared value of 91.47%, which makes this model an incredible fit for view predictions.

b. Predicting Channel Subscriber Counts

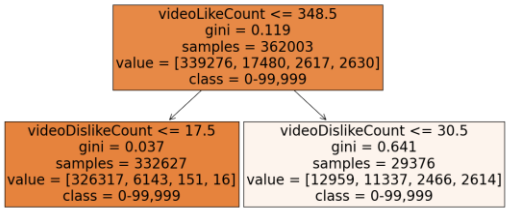
Through multiple linear regression, we created a valid predictive model for channel subscriber counts. With an accuracy of 78.86%, this model holds up decently for making predictions. We concluded that the best predictive variables were channelViewCount, subscriberCount, videoViewCount, and videoLikeCount.

2. Decision Tree and Random Forest Models

a. Predicting Video Category IDs

We did two variations of decision trees with low accuracy scores. This led us to use a random forest model to get the best accuracy we could, which ended up resulting in a disappointing 34.03% accuracy. We concluded that this was a bad predictive model and would use other methods for predicting video category IDs in the future.

b. Predicting Range-based Video View Counts



We did three different depth-variations of decision tree models. The higher the depth, the more accurate the model. Depth-4 attained the lowest accuracy of 95.54%, while max depth attained the highest accuracy of 98.39%. The most important features for this model were videoLikeCount, videoDislikeCount, views/subscribers, subscriberCount, and channelViewCount.

IV. Lessons Learned and Future Work

When building a predictive model, choosing the right type of model is essential. While using a decision tree classification model can increase accuracy, it may come at the cost of specificity compared to a regression model decision tree. Our decision tree model identified five crucial predictor variables that had a significant impact on our predictions, including videoLikeCount, videoDislikeCount, views/subs, subscriberCount, and channelViewCount. In contrast, our linear regression model found that the most reliable predictor variable was views/elapsedtime. However, in our multiple linear regression model, we determined that the most important predictor variables were channelViewCount, subscriberCount, videoViewCount, and videoLikeCount.

Future work in this study includes analyzing how the relationships between the predictor variables and the outcome variable change over time and exploring whether updating the models with new data can improve their accuracy.

V. References

[1] The Devastator. 2023, "YouTube Videos and Channels Metadata" Source. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/revealing-insights-from-youtube-video-and-channe.