Prediction Model for Youtube Views

David Lanorias, James Pitt, Le Thien Nhan Nguyen

INFO 4330 - Data Warehousing and Data Mining (S50)

Ismail El Sayad, PhD

November 30, 2023

Table of Contents

Abstract

Various data modeling techniques are employed in order to identify the relationship between Youtube video views and other metrics such as comments, likes, and dislikes. If the relationship is proven to affect views in a significant way, it can provide a means of focus for content creators looking to improve their metrics.

Introduction

Youtube content creators, whether established or aspiring, always struggle with ensuring their video content can be successful. Though success with video production is difficult to quantify, there are attributes that we can observe to be correlated to what can be considered successful. These attributes are rooted in user feedback and provide a means of telling both the user and content creator whether or not the video has substantial reach and is of sufficient quality.

Topic

We are exploring YouTube Trends (2017-2018) in Canada, focusing on various trending videos and their metrics, and the relationship between these metrics.

Motivation

Social media consumption has dominated the world and it is a key area for advertising and revenue. Understanding YouTube's trending dynamics becomes a crucial task, and provides our motivation for this analysis. These insights can be leveraged for businesses and content creators to optimize their digital presence and strategies.

Problem Statement

If you are not online as a business or influencer, you do not exist. The effort of digital advertising, creating a social media brand and building a following is challenging. The key problem is identifying what drives viewer engagement and how to maximize ad revenue from getting views.

Proposed Solution

Our analysis aims to mine YouTube's trending video data to uncover valuable insights and trends. We are focusing on engagement metrics such as likes, comments, and views to help guide content strategy of those interested in having an online presence. By identifying what factors affect videos the most, creators can have a much clearer idea of what kind of metrics are a priority focus and can adjust their content accordingly.

Objective

To gain actionable insights from YouTube data. We aim to identify the most influential metrics and popular categories.

Implementation

Dataset

- We are going to be using a publicly available dataset containing Youtube statistics for the top trending videos of 2017 - 2018 in the United States.
- We divide the data into 2 half and each set is either use for training the model or testing the result
- We scale the Views variable for better readability and interpretability when dealing with such large numbers.

Modeling

We utilized three different data models for the purpose of comparison between them, Linear Regression, Multivariate Regression, and Polynomial with Interaction Terms. We believe each model will yield different results based on the type of data and their relationship.
For Linear Regression, we will be using the Likes count to predict Views count on future videos. For the Multivariate Regression Model, we added multiple additional variables such as Like, Dislike and comment_count to see if these would affect our prediction. With the Polynomial model and Interaction Terms we will use the Polynomial features to examine the interactions between the Likes and the Comment_count.

Discussion

Results and Model Comparisons

Univariate Model:

- RSS: 6384.342
- R^2: 0.655
- MSE: 0.312
    - Good R-squared value but higher MSE compared to others.

Univariate Model/No Outliers:

- RSS: 4204.271 (↓34.15% from Univariate Model)
- RS: 0.636 (↓2.90% from Univariate Model)
- MSE: 0.252 (↓19.23% from Univariate Model)
    - Improvement in both RSS and MSE compared to the first model, but slightly lower R-squared.

Interpretation

The R^2 value of 0.655 suggests that 65% of the variance in the views can be explained by the likes. This is not entirely accurate however as our MSE value remains quite high at 0.312, meaning that there are some elements to the view count that are beyond the data provided. The presence of outliers, particularly in data like social media views, will skew the data. The Model with outliers removed provided significant improvements in the RSS and MSE

Multivariate Model:

- RSS: 5192.256 (↑23.50% from Univariate Model/No Outliers)
- R-squared: 0.719 (↑13.11% from Univariate Model/No Outliers)
- MSE: 0.254 (↑0.79% from Univariate Model/No Outliers)
    - Best R-squared value and comparable MSE to the no-outliers model, this indicates it explains variance in the data well while maintaining accuracy.

Interpretation

The R-squared value of 0.719 suggests that the model is significantly more adept at predicting video views. This indicates that the three input variables—Likes, Dislikes, and Comment_count—contribute suitable variance for views prediction. Despite this model having a slightly higher Mean Squared Error (MSE) compared to the Univariate/No Outliers model, the difference is marginal enough to be considered roughly equal. The Model also an improvement in RSS value compare to Univariate/No Outliers

Polynomial, Interaction Term Model:

- RSS: 5417.035 (↑4.33% from Multivariate Model)
- R-Squared: 0.707 (↓1.67% from Multivariate Model)

- MSE: 0.265 (↑4.33% from Multivariate Model)
    - Similar R-squared to the multivariate model but higher MSE.

Interpretation

The model performed well despite a slightly higher error rate (MSE: 0.265), it offers good insights into the synergistic effects of likes, comments, and dislikes on video popularity. The interaction terms show how combinations of engagement metrics can non-linearly influence the view counts. The single metric approach of the Univariate model failed to capture this nuance.

Categorizing Videos and Trends

By utilizing the existing category_ids from the dataset, we are able to create a model to better understand the popularity and trends of YouTube. When we graph the data and compare by category and views, we find that the two categories, music and movies, vastly outperform any other category in terms of viewership.

```
categoryMapping = {
    "1": "Film & Animation",
    "2": "Autos & Vehicles",
    "10": "Music",
    "15": "Pets & Animals",
    "17": "Sports",
    "18": "Short Movies",
    "19": "Travel & Events",
    "20": "Gaming",
    "21": "Video Blogging",
    "22": "People & Blogs",
    "23": "Comedy",
    "24": "Entertainment",
    "25": "News & Politics",
    "26": "Howto & Style",
    "27": "Education",
    "28": "Science & Technology",
```

```
  "30": "Movies",
  "31": "Anime/Animation",
  "32": "Action/Adventure",
  "33": "Classics",
  "34": "Comedy",
  "35": "Documentary",
  "36": "Drama",
  "37": "Family",
  "38": "Foreign",
  "39": "Horror",
  "40": "Sci-Fi/Fantasy",
  "41": "Thriller",
  "42": "Shorts",
  "43": "Shows",
  "44": "Trailers"
}
```

Fig 1. Category mapping model

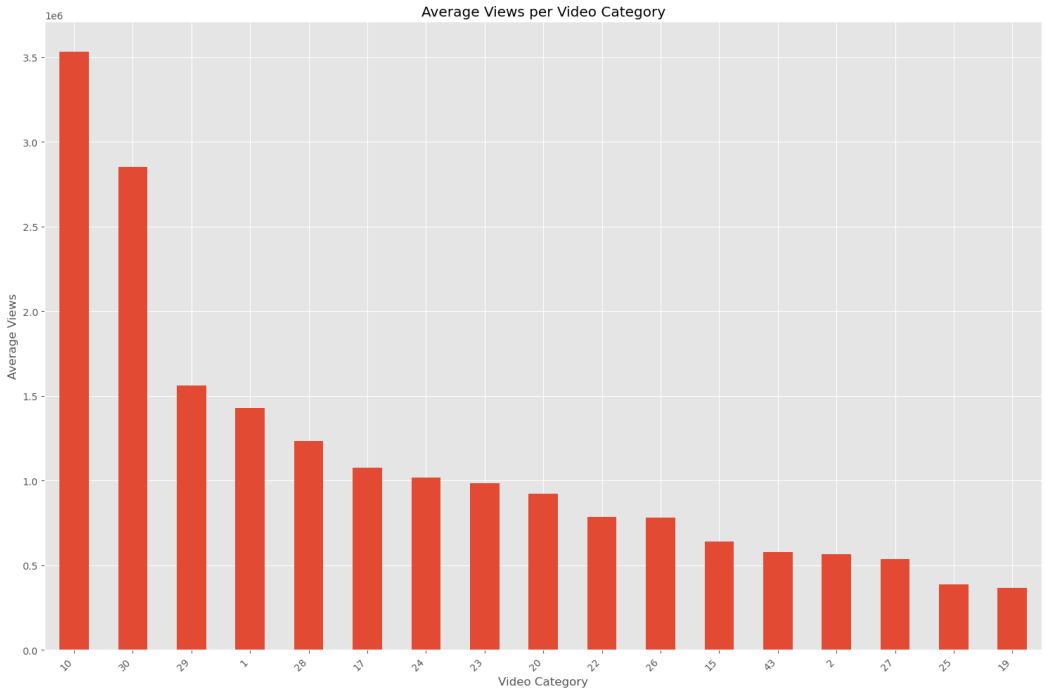

Fig 2. Graph that compares average views to video category

Recommendation

Comparing the R-squared and MSE of the multivariate model with the other models, it is very clear that the multivariate model is better for predictions of view counts. After conducting

these models, the recommendation remains the same: for success on YouTube, focus on content in categories like Music and Movies and optimize engagement metrics such as ‘likes’ above all else, while also paying attention to comments and minimizing dislikes. This approach is crucial for content creators and businesses on Youtube seeking to get better reach and build a brand.

Conclusion

Winner? - Multivariate Regression Model

The Multivariate Regression Model, using likes, dislikes, and comments as predictors, outshines other models in predicting YouTube video views and success.(2nd MSE and 1st R^2) This model effectively captures the complex relationship between various engagement metrics and video popularity.

References

J, M. (2019, June 2). *Trending Youtube Video Statistics.* Retrieved from Kaggle: ` https://www.kaggle.com/datasets/datasnaek/youtube-new/data