Predicting YouTube Video Views from Video and Thumbnail Features

Lean Ting Jin
Matrikelnummer 6956985
ting-jin.lean@student.uni-tuebingen.de

Finn Springorum
Matrikelnummer 6124977
finn.springorum@student.uni-tuebingen.de

Christian Traxler
Matrikelnummer 6969273
christian.traxler@student.uni-tuebingen.de

Anna Chechenina
Matrikelnummer 6987499
dummymail1@uni-tuebingen.de

Abstract

We are are planning to use the YouTube Data API v3 [1] to see which factors of a YouTube video best predict how many videos it will receive.

1 Introduction

YouTube is the second most visited website in the world. In 2024, creators uploaded x hours of content to YouTube, and were paid a total of $\$y$ million in revenue (source: []). For YouTube creators, there is a massive financial incentive to maximize a video's view count.

The view count of a video depends on the YouTube algorithm. However, this algorithm is a black box. Despite this, it is widely believed that the number of views a video receives is strongly correlated with 3 main factors: (1) the click-through rate of the video's thumbnail, (2) the watch duration of the video and (3) the number of subscribers of the channel (source: []).

Increasing the watch duration of a video requires a significant investment in time and money. Meanwhile, the number of subscribers a channel has is also beyond the control of a creator. In contrast, it is easier for a creator to tweak the thumbnail of a video to try to improve the click-through rate. While multiple online resources [], and prominent YouTubers [] claim that more eye-catching thumbnails get more clicks, there does not seem to be any quantitative results on this topic in literature.

Admittedly, it is not clear how one can quantify the eye-catchiness of a thumbnail. However, it is reasonable to assume certain features such as image color, saturation and the presence of human faces are important. Therefore, the aim of this study is to investigate the relationship between the eye-catchiness of a video's thumbnail, and the number of views it receives. As a proxy for eye-catchiness, we use 6 features which are easily derived from the thumbnail: hue, saturation, lightness, contrast, sharpness and the number of faces.

2 Methods

Data Collection. We used the YouTube Data API [1] to collect data on videos from a variety of categories. There are 15 main categories of videos on YouTube. Given our interest in thumbnails, we decided to focus on video categories intended to appear in the feed of a user and elicit a response,

Project Report for *Data Literacy* 2024/25

which are: Comedy, Education, Entertainment, Gaming, How-to & Style, News & Politics, People and Blogs, and Sports.

We excluded videos explicitly sought after through searches (e.g., Film & Animation), videos which are unlikely to include human faces on their thumbnail (e.g., Pets & Animals), or videos from a very niche field (e.g., Nonprofits & Activism).

Due to the limited number of videos that can be retrieved for a given query, the date range from January 1st, 2015 to the time of collection was divided into disjoint sets, and 500 videos were requested for each period. This strategy enabled the acquisition of thousands of videos for a single query. To prevent duplicates while respecting the daily API limit, our data collection strategy entailed the request of up to 2,500 videos from a designated category, each accompanied by a singular generic keyword (e.g., most popular video games for the Gaming category), until the collection comprised precisely 10,000 unique videos from said category, yielding a dataset with 8 categories, for a total of 80,000 unique videos.

Video and Thumbnail Features. For each video V_i, we obtained the following information of interest: thumbnail, T_i, view count, N_i and subscriber count, S_i. From each thumbnail T_i, we extracted the following 6 features: hue (h_i), saturation (s_i), lightness (l_i), contrast (c_i), sharpness (s_i) and number of faces (f_i).

The hue is the angle of the mean color of the thumbnail in the HSV color space, with a range of 0 to 360 degrees. The saturation, lightness and contrast follow their standard definitions as implemented in the OpenCV Python library (normalized to [0,1]). The sharpness is the log variance of the Laplacian of the grayscale image. The number of faces is the number of distinct faces in the thumbnail as predicted by RetinaFace model implemented in the DeepFace library [2, 3].

Analysis. We are interested the the relationship between the view count N_i and the 6 features derived from the thumbnail T_i. However, before performing any analysis, we note that intuitively, a video V_i published by a channel with more subscribers S_i is likely to have a higher view count N_i compared to one with fewer subscribers, regardless of the thumbnail features. To remove this influence, we first fitted a linear regression model of (log) view count against (log) subscriber count. As the data appears log-normal, the base-10 logarithm is used to ensure that the residuals are homoscedastic and normally distributed. Only the residuals of this model, called the residual log view count, $R_i := \log(N_i) - \hat{\beta}_0^{(\gamma)} - \hat{\beta}_1^{(\gamma)} \log(S_i)$ were considered for the remainder of the analysis. $\gamma \in \{1, 2, \cdots, 8\}$ corresponds to one of the 8 categories that video V_i belongs to.

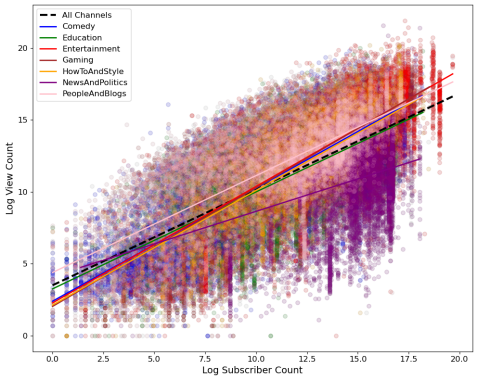

Figure 1: Regression of log view count against log subscriber count. The relationship is similar for all 8 video categories. For the combined dataset, the relationship is $\widehat{\log(N_i)} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(S_i)$, 95% CI for $\beta_0 = [3.445, 3.571]$, 95% CI for $\beta_1 = [0.663, 0.673]$, $R^2 = 0.467$.

Next, the correlation between the 6 thumbnail features was investigated. There did not appear to be strong multicollinearity between features. A linear regression model was fitted to each continuous feature (saturation, lightness, contrast, sharpness).

Meanwhile, the number of faces f_i was treated as a discrete feature, and 95% confidence intervals of the residual log view count R_i were computed for each face count $k \in \{0, 1, 2, 3\}$. The videos with thumbnails containing 3 or more faces were grouped into the same bin and analyzed together.

We also treated the hue h_i as a discrete feature. Since hue ranges from 0 to 360 degrees, we grouped the videos into 6 bins of equal width. For example, bin with a hue of 0-60 degrees would contain videos with reddish thumbnails. We then computed a 95% confidence interval for R_i for each bin.

Finally, a multiple linear regression model was fitted to the residual log view count R_i against hue, saturation, lightness, contrast, and number of faces.

3 Results

4 Discussion/Limitations

Discussion

Limitations As a project based on the YouTube API, there are couple of inherent limitations of the data collection. One issue with the data collection is the categories: each video is either user-defined or automatically assigned only one category. This means that there exists some error in the categories such that errors are likely and that some categories that we wanted to exclude (like music videos) are likely to be included and may skew results.
Also,

5 Statement of Contributions

Here is an example:

XX performed the correlation analysis, organized the data and code for the processing of dataset1 and subdataset2, and created the scatter plot. YY created the random forest regression model, performed the data cleaning for the xyz analysis / xyz database, and created the bar charts to display the regression results. ZZ researched and collected the raw data, restructured the pipeline for the data analysis, and proof-read the draft for the final report. AA performed the data cleaning for dataset1, and performed the Ridge and Lasso regularization. All members of the group contributed to writing the report.

References

[1] Google. YouTube Data API v3. https://developers.google.com/youtube/v3/docs.

[2] Sefik Ilkin Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilisim Teknolojileri Dergisi*, 17(2):95–107, 2024.

[3] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.