# PREDICTING VIRAL YOUTUBE VIDEOS

## UNDERSTANDING POTENTIAL DRIVERS FOR SUCCESS

MIT
MANAGEMENT
SLOAN SCHOOL

Dilan SriDaran · May Oo Khine · Seth Chatterton · Nick Antoniou

8 December 2023

# 1 Executive Summary

## 1.1 Scope and Objectives

YouTube is one of the world's largest digital platforms, and is a primary hub for creators, advertisers, and viewers. Understanding factors that contribute to views is crucial for creators aiming to optimize their strategies and for marketers aiming to allocate resources effectively. This project aims to:

- **Predict viral video views:** We aim to predict the views a video will accumulate at discrete time intervals from its date of upload. This can be used by marketers to better understand the potential audience of a video and better inform their use of marketing budget.
- **Understand the predictors of views:** Whilst not an exact science, the project aims to better understand factors that contribute to higher views. This can be used by content creators to understand and optimize their controllable decisions to maximize their potential viewership.

Whilst a global platform, we focus on views in the United States, YouTube's second largest user base.

This report is a high-level report that focuses on key insights. Technical detail is available on request.

## 1.2 Key Findings

We built a model with a high degree of accuracy, with an out-of-sample $R^2$ of 0.77 when predicting the number of views of a viral video (Section 3). Notable features that drive higher views are summarized below, with more detail provided in Section 4.

Table 1: Summary of key predictors of high views, overall and by categories (approximated from SHAP plots).

| | Overall | Sports | Music | Gaming | News | Education |
|---|---|---|---|---|---|---|
| Upload Time | | | | | | |
| Evening/nighttime | ✓✓ | | ✓✓ | ✓ | | |
| Thumbnail | | | | | | |
| Moderate brightness | ✓✓ | | ✓✓ | | | |
| Moderate contrast | ✓ | | | | | |
| Low contrast | | ✓ | | ✓ | | ✓ |
| Title | | | | | | |
| Uppercase | | | | ✓ | | |
| Contains quote | ✓✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Long or short | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Contains word "official" | ✓ | ✓ | ✓ | | ✓ | ✓ |
| Contains word "trailer" | | | | ✓ | | |
| Contains word "shorts" | | | | ✓ | | ✓✓ |
| Limited stop words | ✓ | | | | | |
| Previous Video History | | | | | | |
| High likes | ✓✓ | | ✓✓ | ✓✓ | | |
| High views | ✓✓ | ✓✓ | ✓✓ | ✓✓ | ✓✓ | ✓✓ |
| High comments | | | ✓✓ | ✓ | ✓✓ | |
| High volume of uploads | | ✓✓ | | | | |

Key: ✓✓ Very strong predictor ✓ Moderately strong predictor

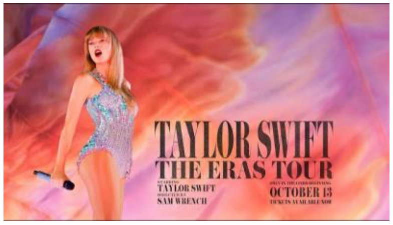
# 2 Data

The project leverages a Kaggle dataset that includes several months of daily trending YouTube videos across the United States, with up to 200 listed videos per day. The dataset includes videos up to and including 22 October 2023, the date of data download. Noting that since audience preferences are dynamic and ever-evolving, we do not consider any data prior to 2022, as this is unlikely to be reflective of current behavior. The data considered therefore includes 130,398 observations.

The raw data includes the video title, channel title, category, publish time, views, descriptions, and thumbnail links. We performed additional feature engineering to this data, guided by a desktop review of factors that typically influence the popularity of videos. This included adding features related to:

- **Prior videos:** Unsurprisingly, research suggests that the most important features to predict views are views, comments, likes, and dislikes (herein collectively referred to as "engagement metrics") of creators' prior videos. We created features related to these engagement metrics for the creator's most recent video upload, and all their viral uploads over the preceding year.
- **Thumbnails:** Thumbnails provide a first glimpse to viewers about what a video is about and help them decide if they want to watch it. We perform image analysis to identify brightness, saturation, contrast, and color palette, and detect the presence of faces and text in the image.

Table 2: Illustrative example of thumbnail analysis.



| | |
|---|---:|
| Brightness: | 135.83 |
| Saturation: | 122.60 |
| Contrast: | 44.46 |
| Colors: | red, brown |
| Face: | Yes |
| Text: | Yes |

- **Titles:** Alongside thumbnails, titles are the other point of first contact with viewers. We include features such as length, sentiment, casing, stop words, punctuation, quotations, and key words.
- **Upload time:** Alongside their content, the other key mechanism that creators can control is upload time, as this affects when their videos will start to appear within their potential audiences' feed. We create variables to identify the time and day of publish, including whether the published day is a weekend or a US public holiday. Times are in Pacific Standard Time.

# 3 Model

We developed a model to predict the number of views a viral video would obtain. To ensure no data leakage, we have delineated a temporal boundary for training (1 January 2022 to 31 December 2022), validation (1 January 2023 to 31 May 2023), and testing (1 June 2023 to 22 October 2023).

We explored a number of regression models, ranging from simple linear, lasso, and ridge regressions and CART, to more complex random forests and XGBoost. These models were compared to two baseline approaches. The first predicts views using average views across the training data, whilst the second predicts views based on views attained by that creator's prior upload. The evaluation criterion was out-of-sample $R^2$, however, out-of-sample mean absolute error (MAE) is also reported.

Table 3: Out-of-sample performance metrics for models tested.

| Model | $R^2$ | MAE[1] |
|---|---|---|
| Baseline 1 | -0.01 | 3.31 |
| Baseline 2 | 0.61 | 2.19 |
| Ridge | 0.23 | 3.43 |
| Lasso | 0.23 | 3.47 |
| CART | 0.64 | 2.18 |
| Random Forest | 0.76 | 1.73 |
| XGBoost | 0.77 | 1.75 |

Given the high dimensionality of the data and the complexity of the problem, it is unsurprising that the linear models perform quite poorly. The non-linear models (CART, Random Forest, and XGBoost) perform much stronger. Whilst the CART model is most interpretable, we believe the additional complexity of the XGBoost model is justified, given its $R^2$ is 20.3% higher than the CART model.

## 4 Managerial Insights

The key value of the model is not necessarily the precision per se, but rather the understanding it provides about what factors are most influential to gains in viewership. The discussion below provides an overview of features that drive viewership, with reference to videos overall and also four distinct categories: “Music,” “Sports,” “Education,” and “News.” Supporting technical details are provided in Appendix A and B, and there are over 25 other categories that can be explored further if required.

Note, whilst engagement on prior videos are extremely influential factors, these are not within the direct control of creators. Creators should therefore focus on insights on the thumbnail, title, and upload time.

**Prior videos:**

- **Higher previous engagement:** Unsurprisingly, videos are more likely to generate higher views if that creator’s previous videos generated more views, as these are likely to boost the profile and subscriber count of the creator. Interestingly, for “News” videos, high previous comments are a strong predictor, likely because engaging creators typically post thought-provoking and potentially polarizing content to spark debate. By contrast, views for “Music” and “Gaming” are more strongly driven by previous likes, as audiences typically watch these categories for pleasure or enjoyment (manifested through likes) rather than discussions.
- **Higher volume of uploads:** Frequency or volume of uploads is typically not a strong driver of views, except for “Sports.” This is intuitive as sports videos are typically characterized by a small number of channels holding exclusive rights to broadcast specific sports content, and a recent trend towards very high-volume uploads of small clips. For example, NBC Sports uploaded over 50 English Premier League videos in a 24-hour window on 6 December 2023.

**Thumbnails:**

- **High brightness:** Videos with moderate-to-high brightness attract higher views, most notably for “Music” videos, as dynamic use of color and composition can help to catch interest. However, analysis suggests that too much brightness can overwhelm and put off audiences.

[1] MAE is reported in millions.

- **Mixed contrast effects**: Generally, thumbnails with moderate contrast perform better. However, for “Sports,” “Gaming,” and “Education,” it appears that lower contrast attracts more views. This is an interesting result, and may warrant further investigation.

**Titles:**

- **Long or short:** For categories such as “Education” and “News” that are watched for intellectual purposes, longer titles perform better as they are better able to provide informative content to persuade viewers to engage with their content. By contrast, for categories that are viewed for entertainment, such as “Music” and “Gaming,” short and punchier titles better attract views.
- **Direct quotes:** Direct quotes appear to attract views across all categories, which may reflect either “click-bait” or the use of quotes from popular or influential figures to attract attention.
- **Upper case:** The use of uppercase words appears to be a strong mechanism to attract the attention of audiences for “Gaming” videos. This is unsurprising as this is a commonly used technique to convey a sense of excitement, which typically resonates with gaming audiences. By contrast, the use of all caps may not be considered appropriate or professional for “Education” or “News” videos.
- **Minimal stop words:** YouTube previews have a finite number of characters, and therefore the use of stop words wastes valuable space to convey topics to audiences.
- **Keywords:** Specific keywords tend to attract views. For example, words such as “official” typically garner interest in “Sports,” “News,” and “Education,” whilst words such as “trailer” or “shorts” typically do well in “Gaming” videos.

**Upload Time:**

- **Evenings:** Videos typically attract higher views when published during evenings (Pacific Time). This is likely because these are prime viewership times for individuals. The exception to this is “Sports” and “News,” where there are no clear time patterns, suggesting a focus on immediate post-event dissemination, prioritizing timeliness over specific time-of-day considerations.

## 5 Limitations and Next Steps

- **Biased data:** The model dataset considers only viral videos. This biased dataset may overlook non-viral videos that could provide valuable insights which should be considered in future work to create a more balanced and comprehensive dataset.
- **Expand scope:** The analysis is limited to a specific region and timeframe, potentially missing out on global trends and the latest shifts in YouTube's audience engagement. The model should be expanded in scope and potentially consider viewer demographic data to understand how different audience segments interact with and influence the popularity of YouTube content.
- **Key features:** The model doesn't include features such as subscriber count, media presence, engagement levels, and collaboration impacts. Future work should explore these features, as we believe them to be crucial in understanding a video's potential reach and virality.
- **Preference evolution:** The model fails to account for the rapidly changing content trends and viewer preferences on YouTube, and YouTube’s algorithm updates, which can significantly impact video popularity. This can be addressed by continuously updating the model to reflect the latest content trends, viewer preferences, and algorithm changes on YouTube.
- **Techniques:** Future work may utilize techniques like natural language processing and video content analysis to delve deeper into the qualities that make content engaging shareable.

## Appendix A: Feature Importance

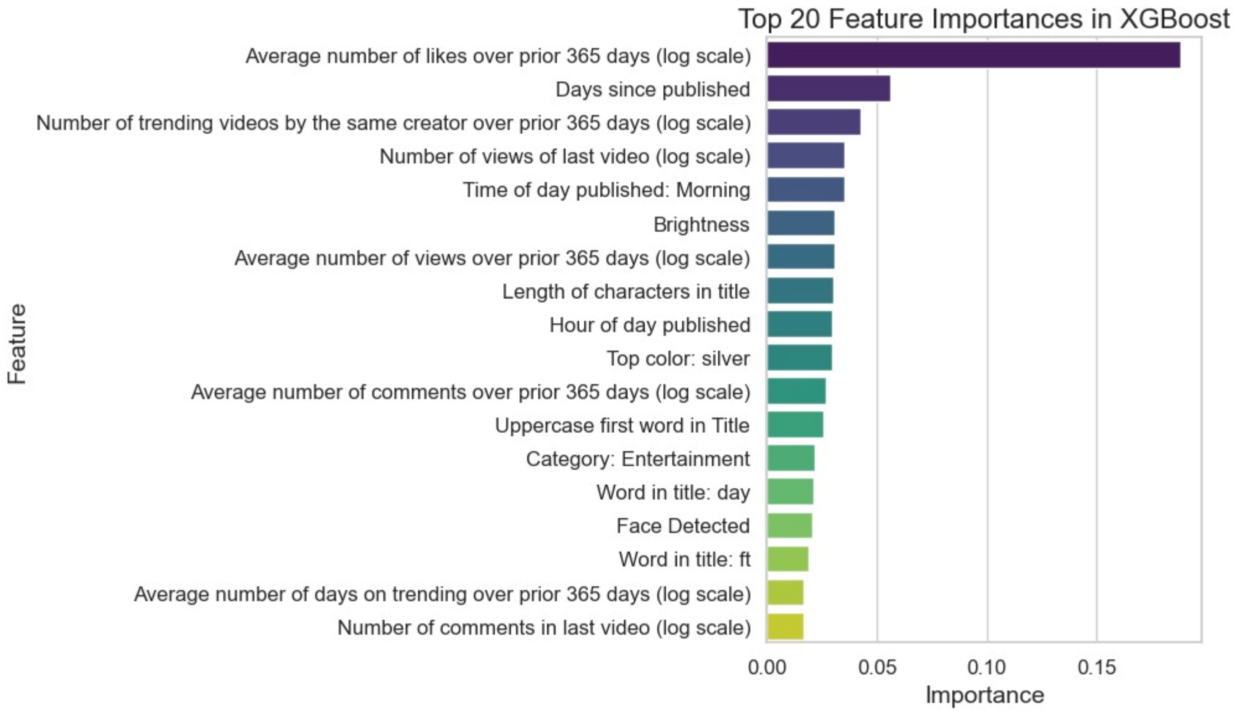


Figure 1: Top 20 features for XGBoost model.

## Appendix B: SHAP Plots

### Overall

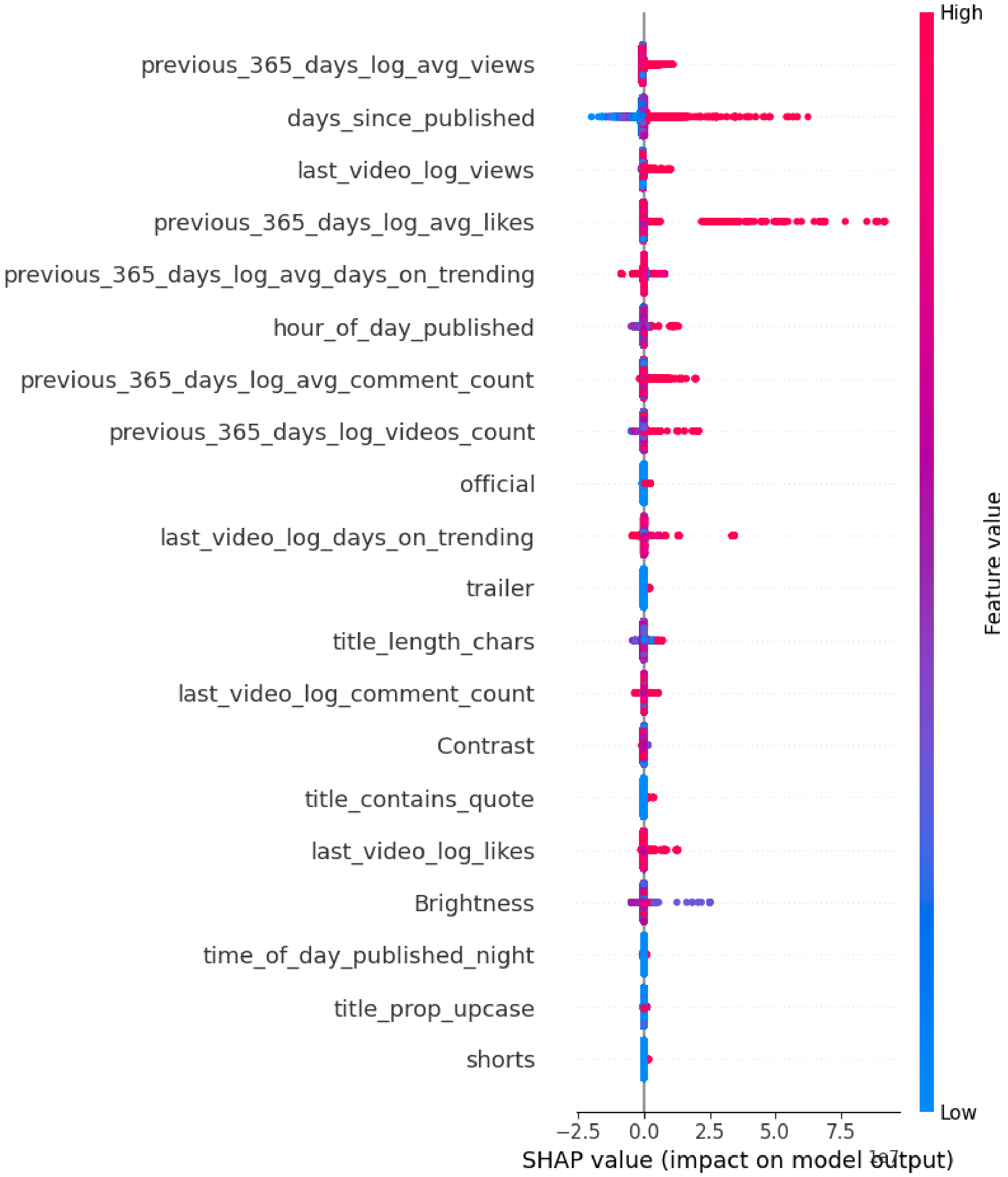


Figure 2: SHAP value plot for all videos.

## Sports

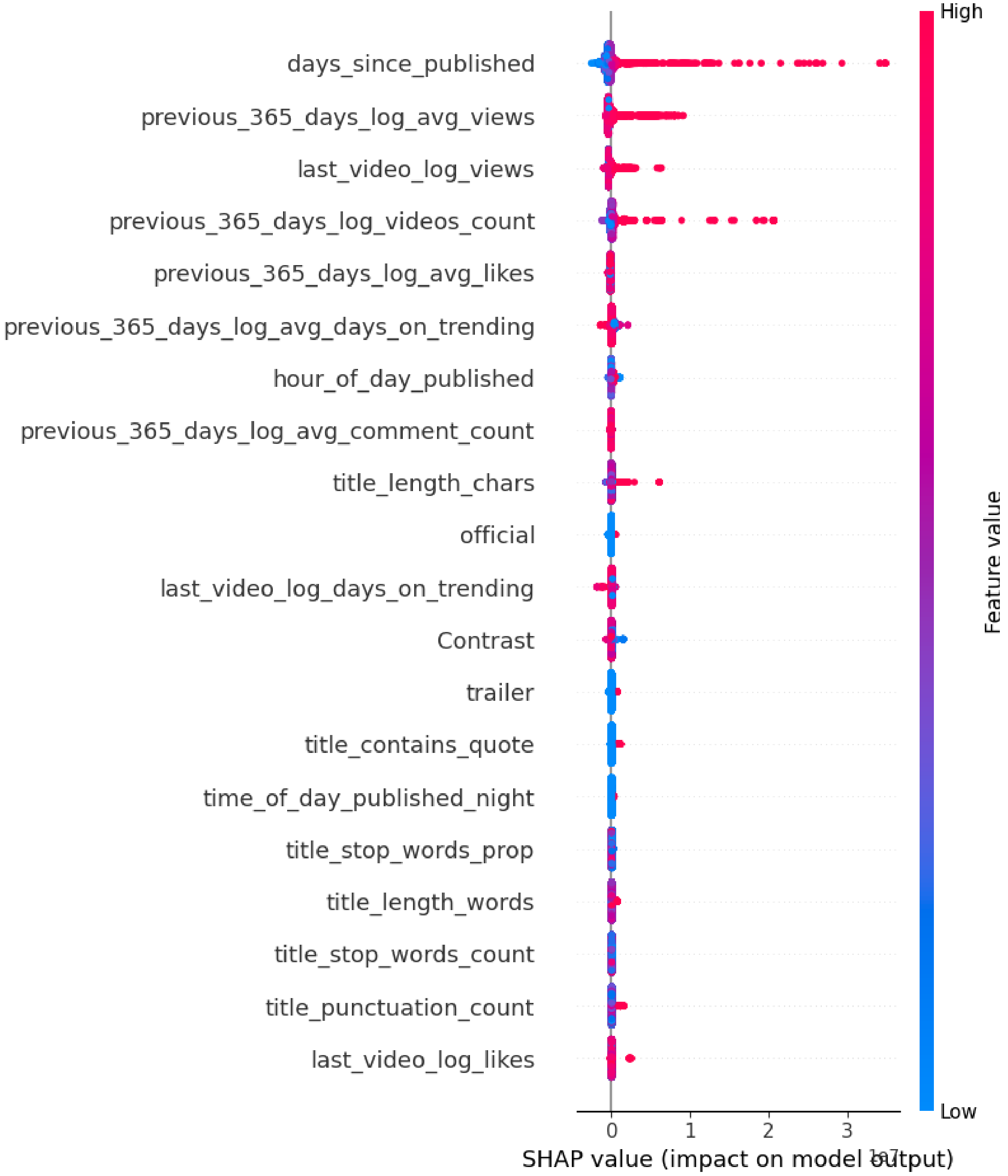


Figure 3: SHAP value plot for “Sports” videos.

### Education

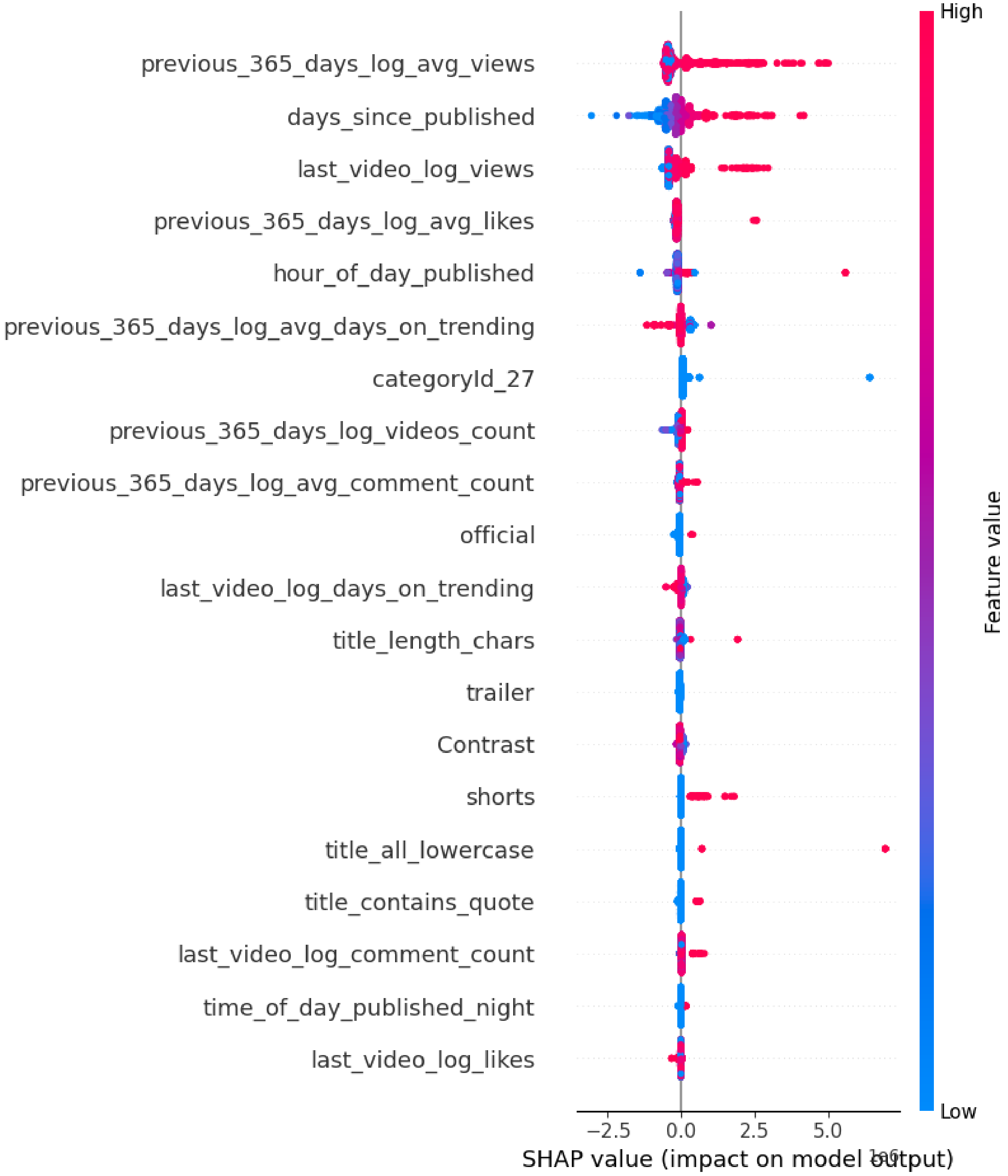


Figure 4: SHAP value plot for "Education" videos.

## Gaming

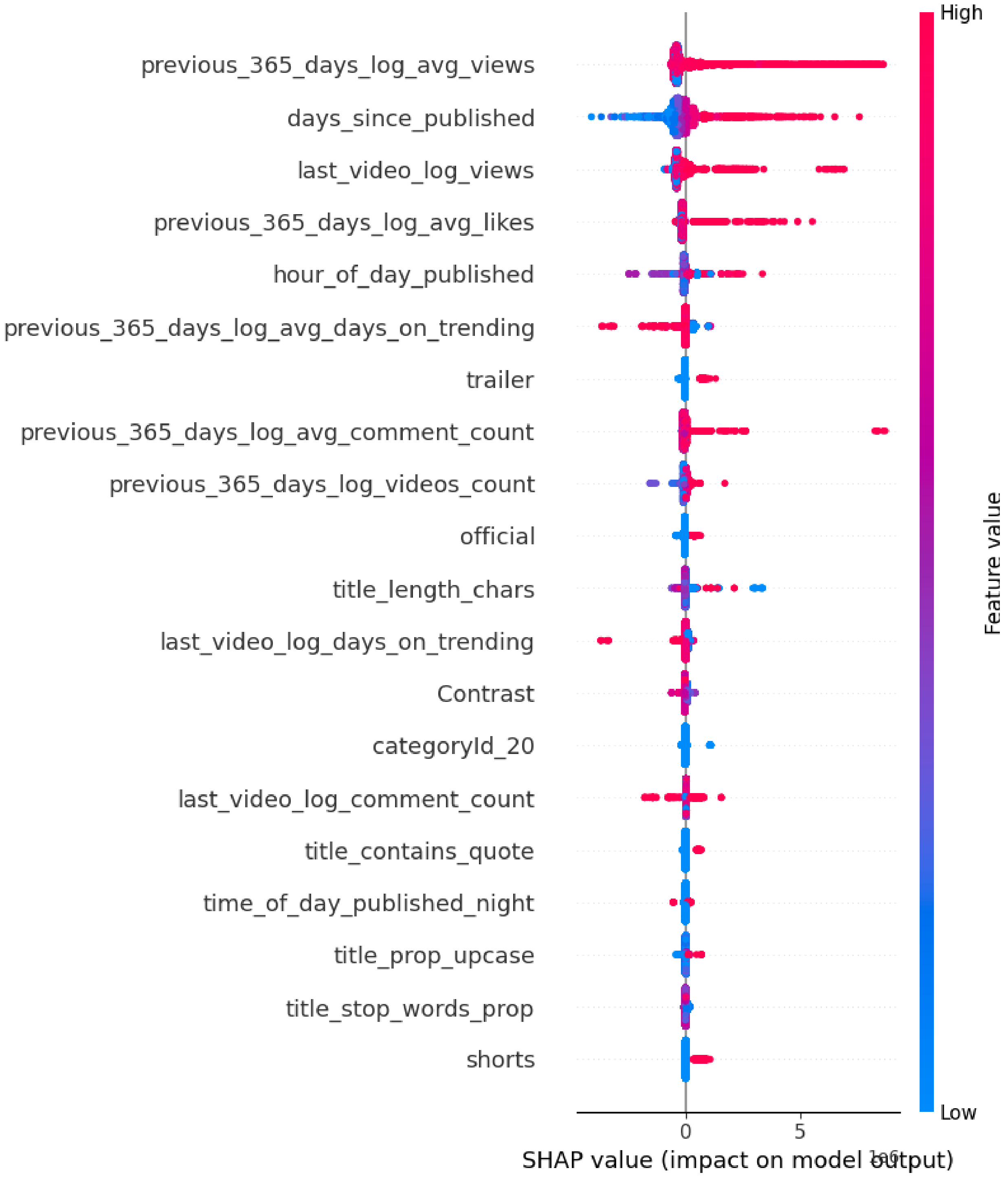


Figure 5: SHAP value plot for “Gaming” videos.

## Music

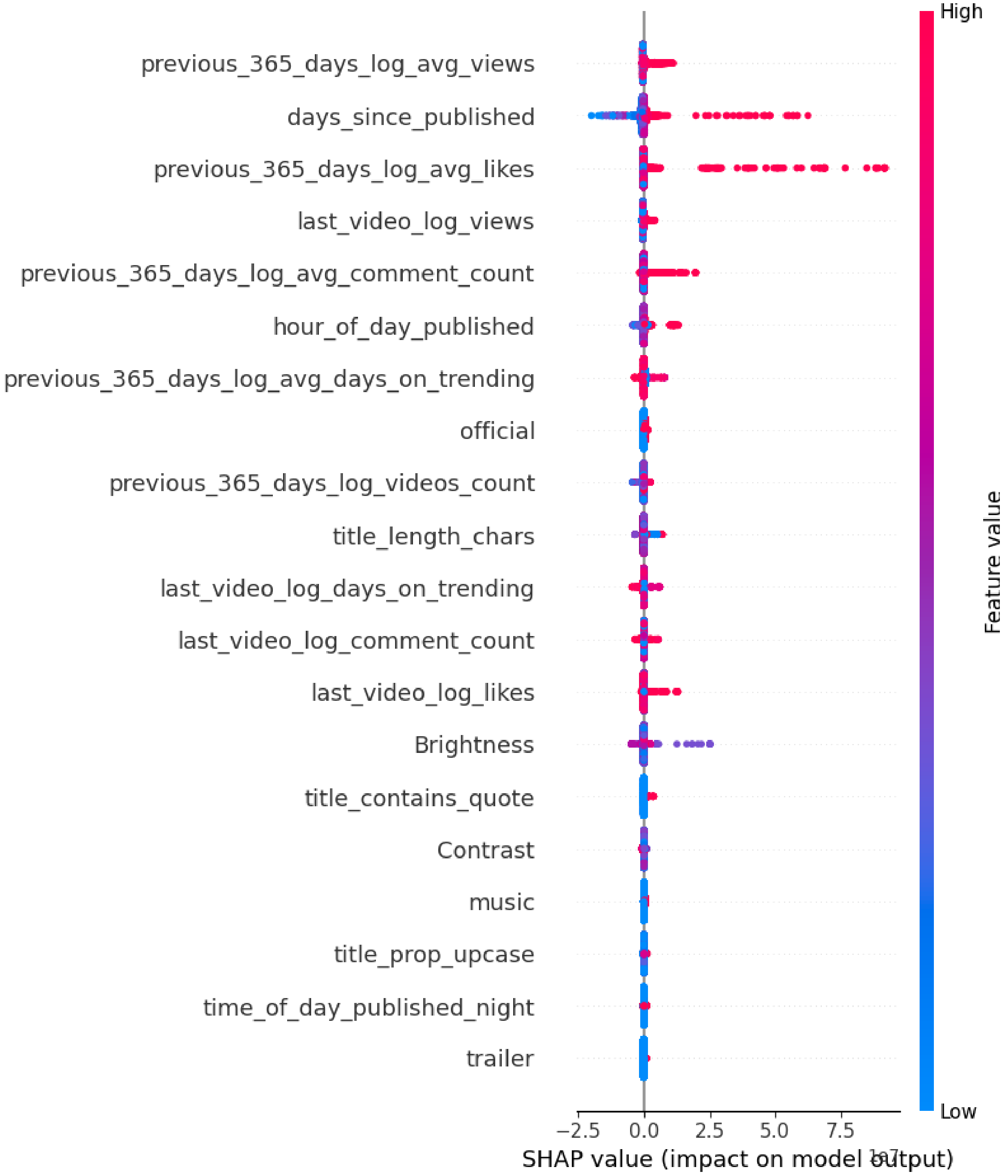


Figure 6: SHAP value plot for “Gaming” videos.

## News

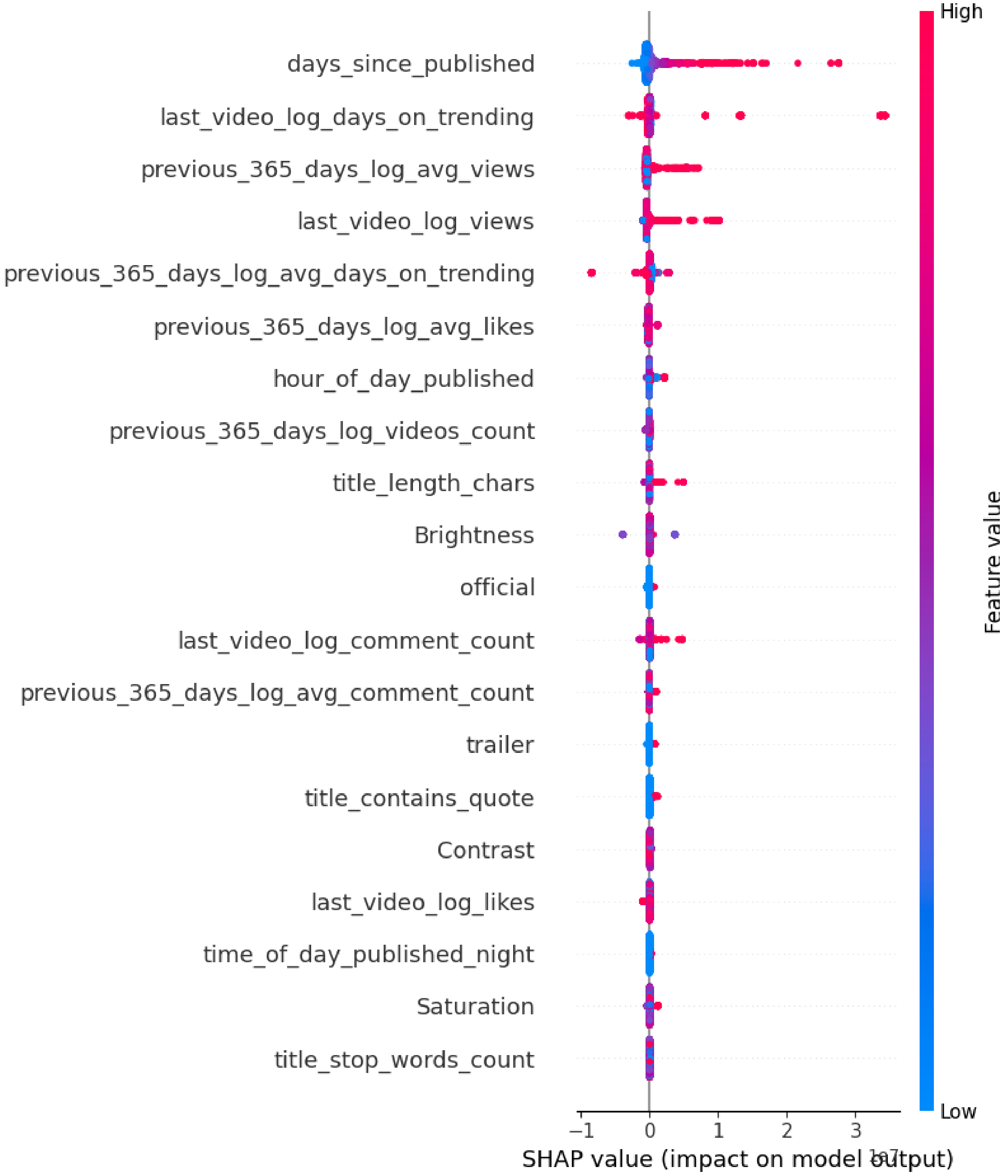


Figure 7: SHAP value plot for "News" videos.