What Factors Affect the Number of Views on YouTube?

An Analysis using PCA and SVM

Xu Zichuan[1], Ito Aoi[2], Bamba Gen[3], and Akagi Ryusei[3]

[1]Big Data
[2]Network Architecture
[3]Nonlinear Math Science

Group 2

Abstract

The proliferation of digital content has made YouTube a primary platform for information dissemination and entertainment. Understanding the key drivers of video viewership is crucial for content creators aiming to maximize their impact. This study investigates the factors influencing YouTube video views by applying Principal Component Analysis (PCA) and Support Vector Machine (SVM) techniques to a dataset of 6,677 videos. We executed an extensive feature engineering pipeline, creating 37 analytical features related to thumbnail aesthetics (e.g., RGB intensity, visual complexity), title structure, channel authority, and publication timing. PCA was utilized to reduce dimensionality and identify latent factors, revealing that thumbnail color properties, engagement ratios, visual complexity, and title strategies are the most significant dimensions of variance. To model viewership, we developed two types of SVM models. An SVM regressor predicted the logarithm of view counts, achieving a coefficient of determination (R^2) of 0.642 on the test set. For classification, we critically compared two scenarios: a model using all features (including post-publication data like 'Like Count') achieved a misleadingly high accuracy of 97.6% due to data leakage. In contrast, a methodologically sound model using only pre-publication features yielded a more realistic predictive accuracy of 53.0%. This highlights the challenge of forecasting success and provides a robust framework for actionable, pre-publication strategies for content creators.

1 Introduction

YouTube has evolved into a global stage where content creators compete for audience attention. For creators, achieving high viewership is not only a measure of success but also a key to monetization and influence. Consequently, there is a significant interest in deciphering the formula for a successful YouTube video. We aim to analyze the factors influencing the number of views on YouTube using data-driven methodologies, as outlined in our project brief. By identifying these factors, content creators can develop effective strategies to boost their viewership.

This analysis considers a wide range of elements, including video metadata (title, tags), thumbnail properties[1], engagement metrics, and channel statistics. To unravel the complex relationships within this data, we employ a two-pronged machine learning approach. First, Principal Component Analysis (PCA) is used to reduce the dimensionality of the feature space and pinpoint the most significant underlying factors. Second, Support Vector Machine (SVM) models are developed to both predict the number of views (regression) and classify videos into performance tiers.

We detail our analytical pipeline, from data preparation and extensive feature engineering to model implementation and evaluation, concluding with a set of actionable recommendations for YouTube content creators based on pre-publication factors. Our code is open-sourced at [2].

2 Methodology

Our analytical process involved data loading and preprocessing, extensive feature engineering, and the application of machine learning models.

2.1 Data and Feature Engineering

The initial dataset comprised 6,677 YouTube videos. A minor issue of 45 missing values in the 'Tags' column was handled during preprocessing. Recognizing that raw data has limited predictive power, we engineered a comprehensive set of 37 features available for analysis:

- **Thumbnail Analysis**: Brightness, colorfulness, RGB channel values, RGB intensity and variance, and binary flags for the presence of people, text, and graphics were extracted. A `visual_complexity` score was synthesized from these features.
- **Title Analysis**: We quantified title length, word count, and the presence of numbers, all-caps words, questions, exclamations, and brackets. A `title_engagement_score` was created based on these elements.
- **Engagement Ratios**: Post-publication metrics like `like_to_view_ratio` and `comment_to_view_ratio` were calculated. These are used for descriptive analysis but are excluded from predictive modeling to prevent data leakage.
- **Temporal Features**: `publish_hour`, `publish_day_of_week`, and `publish_month` were extracted.
- **Channel Metrics**: We included channel-level data such as subscriber count and average video views to represent channel authority.
- **Categorical Encoding**: Variables like 'Channel Name' and 'Video Length Category' were numerically encoded.

2.2 Exploratory Data Analysis (EDA)

EDA revealed that the 'View Count' distribution is heavily right-skewed (Figure 2), necessitating a log transformation for the regression task to stabilize variance. The correlation matrix (Figure 1) showed strong correlations between like and comment counts (0.71) and between different thumbnail color channels (e.g., G and B at 0.91), but weak direct correlations with 'View Count', suggesting the need for more complex modeling.

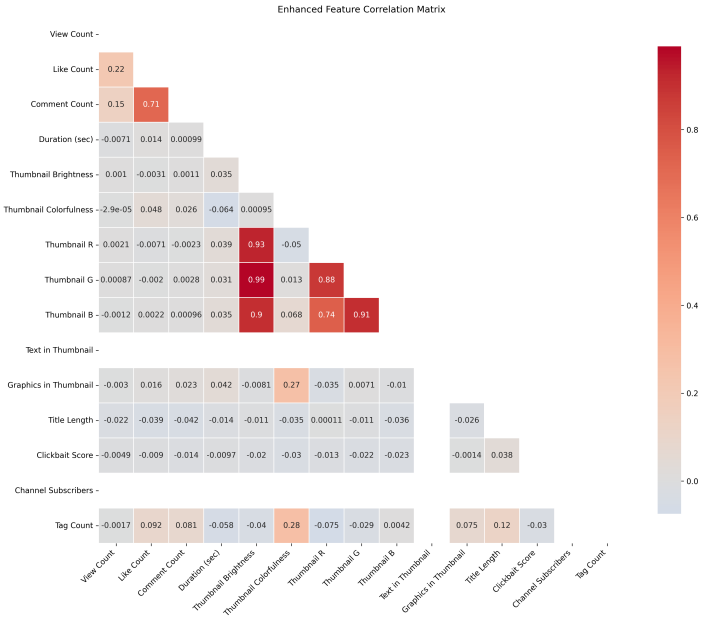


Figure 1: Enhanced Feature Correlation Matrix.

2.3 Principal Component Analysis (PCA)

PCA was applied to the 37 engineered features to identify the principal axes of variation. The goal was to transform correlated features into a smaller set of uncorrelated variables (principal components) that capture most of the information.

2.4 Support Vector Machine (SVM)

SVM was chosen for its effectiveness in high-dimensional spaces. We conducted three distinct analyses:

- **SVM Regression:** An SVR model was trained to predict the log-transformed 'View Count' using all 37 features. A grid search identified optimal hyperparameters as a radial basis function (RBF) kernel with $C = 10$ and $\gamma =' scale'$.
- **SVM Classification (with Data Leakage):** For comparison, an SVM classifier was trained to categorize videos into 'Low', 'Medium', and 'High' view tiers using all 37 features, including post-publication engagement metrics. The optimal parameters were a linear kernel with $C = 100$.
- **SVM Classification (Pre-Publication):** To build a realistic predictive model, a second classifier was trained using only the 32 features available *before* a video is published. Optimal parameters were an RBF kernel with $C = 1$ and $\gamma = 0.1$.

3 Results and Analysis

The application of our methodology yielded significant insights into the factors driving YouTube views.

3.1 Principal Component Analysis Results

The PCA results show that the first few components capture distinct, interpretable aspects of the data's variance. As shown in Figure 3, approximately 19 components are needed to explain 95% of the total variance, indicating the multidimensional nature of video success. The feature loadings (Figure 4) reveal the underlying structure:

- **PC1 (16.1% Variance):** Heavily loaded with thumbnail color features like `rgb_intensity` (0.45) and `ThumbnailBrightness` (0.45). This represents a **'Color Psychology'** dimension.
- **PC2 (9.6% Variance):** Dominated by `engagement_score` (0.54) and the underlying engagement ratios. It can be interpreted as the **'Engagement Efficiency'** component.
- **PC3 (8.3% Variance):** Shows strong loadings for `ThumbnailColorfulness` (0.38) and `visual_complexity` (0.37). This component captures **'Visual Complexity'**.
- **PC4 & PC5 (15.6% Combined Variance):** Strongly associated with title features like `HasNumbers`, `TitleLength`, and `TitleWordCount`, representing a **'Title Strategy'** dimension.

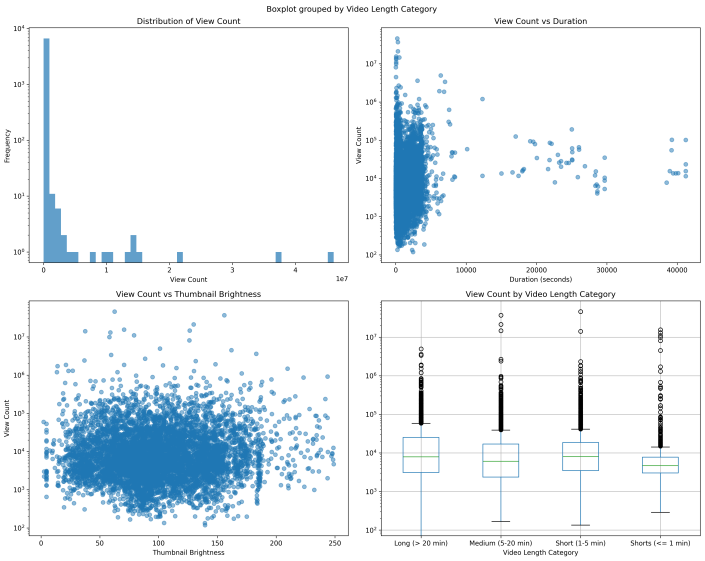


Figure 2: EDA Overview: View Count Distribution, Scatter Plots, and Boxplots by Video Length Category.

3.2 SVM Regression Results

The SVM regression model achieved a training R^2 of 0.795 and a testing R^2 of 0.642. An R^2 value of 0.642 indicates our model can explain approximately 64.2% of the variance in the log-transformed view counts of the test data. For a complex social media prediction task, this is a strong descriptive result. Figure 5 shows the scatter plot of predicted versus actual log views, confirming the model's power as the points cluster around the diagonal line.

3.3 SVM Classification Results

The two classification models produced starkly different results, providing a crucial insight into predictive modeling.

Analysis with All Features: The classifier using post-publication data achieved a near-perfect testing accuracy of 97.6% (Figure 6a). This high performance is attributable to data leakage; features like 'Like Count' are consequences, not predictors, of view count.

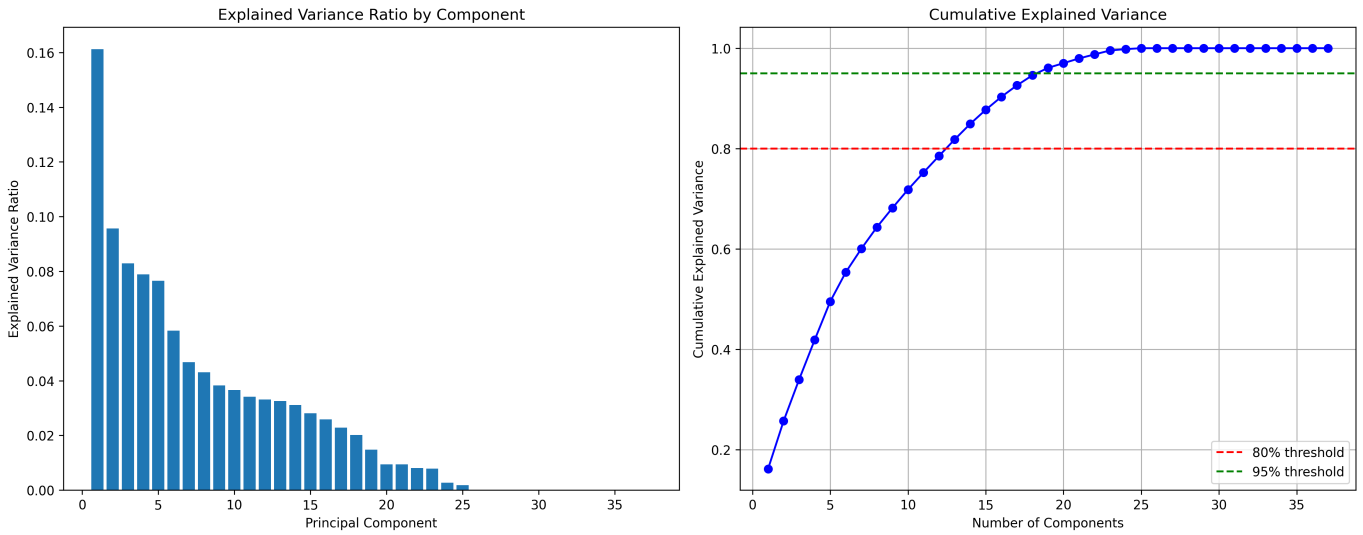


Figure 3: Explained Variance by Principal Component. The right plot shows that 13 components are needed to cross the 80% threshold.

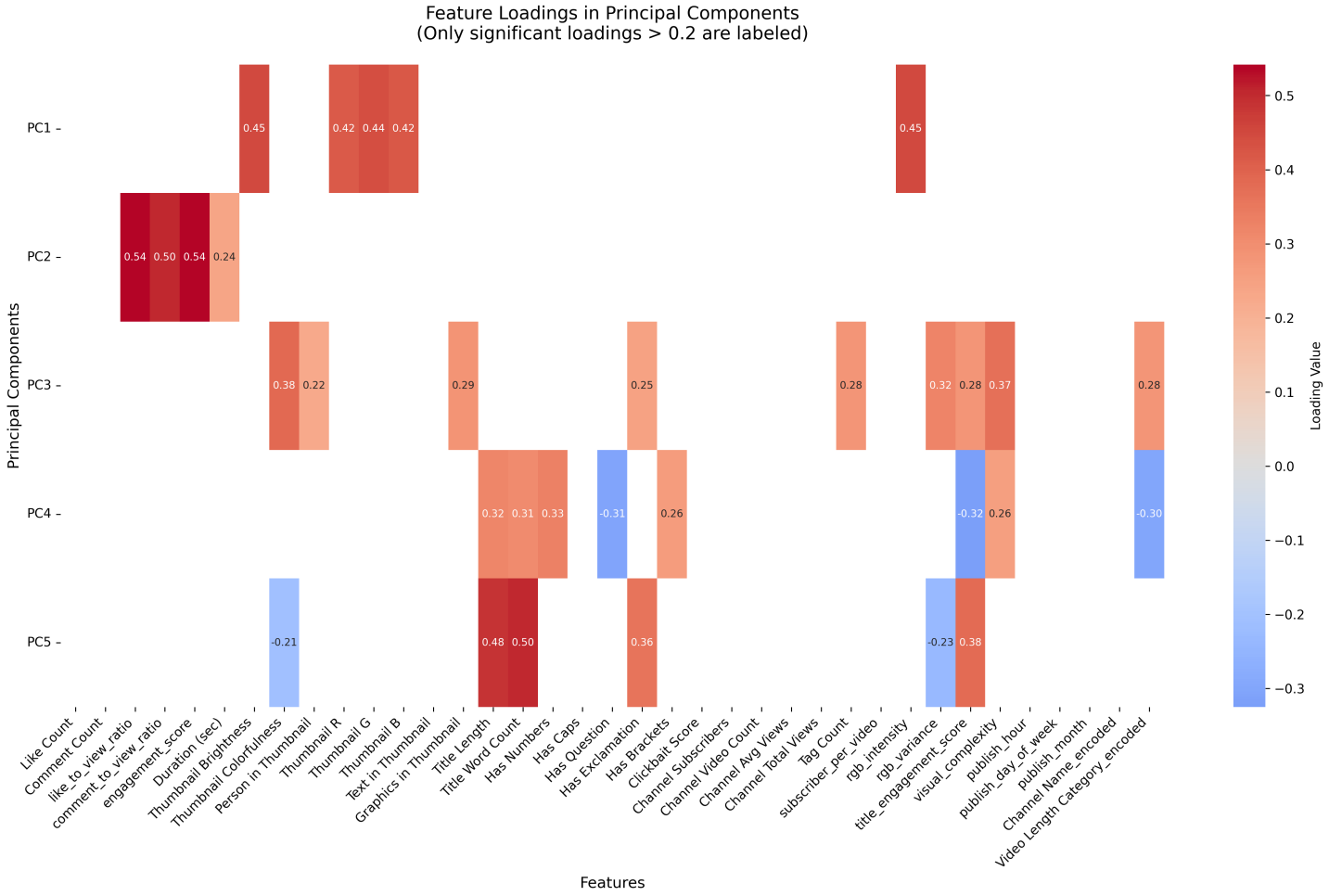
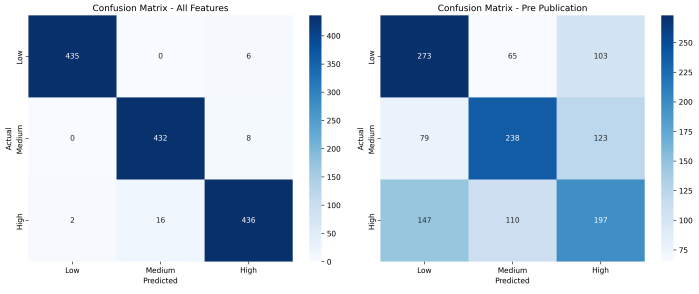


Figure 4: Feature Loadings in First Five Principal Components. Red indicates a strong positive loading, blue a strong negative loading.

Analysis with Pre-Publication Features: The classifier using only features available before publishing achieved a testing accuracy of 53.0% (Figure 6b). While modest, this accuracy is significantly better than random chance (33.3%) and represents a realistic measure of our ability to forecast a video's success tier. The confusion matrix shows the model is most confident in identifying 'Low' view count videos but struggles more with distinguishing 'Medium' and 'High' tiers.



(a) All Features (Leaked) **(b)** Pre-Publication (Realistic)

Figure 6: Confusion Matrices for SVM Classification. (a) shows inflated accuracy due to data leakage. (b) shows realistic predictive performance.

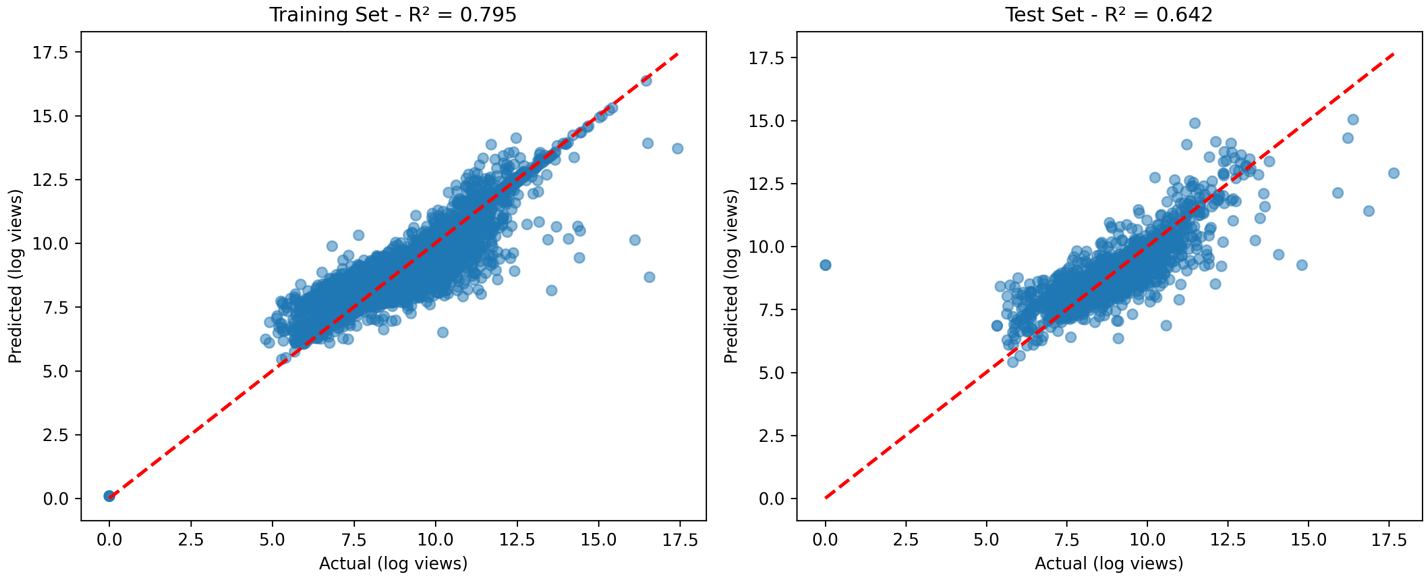


Figure 5: Predicted vs. Actual Log(Views) for SVM Regression on Training and Test Sets. The red dashed line represents a perfect prediction ($y = x$). The model shows good generalization from the training set to the test set.

4 Discussion and Recommendations

Our analysis confirms that YouTube success is multifactorial. The descriptive power of the models is high, but true prediction is challenging. The 45 percentage point drop in accuracy (from 97.6% to 53.0%) when excluding post-publication data starkly illustrates the difference between explaining success in hindsight and predicting it beforehand.

4.1 Limitations

It is important to acknowledge the limitations of this study.

1. **Generalizability**: The dataset appears sourced from specific channels (e.g., 'Google for Developers'), so findings may not apply to all genres.
2. **Causality**: This analysis reveals correlations, not causation.
3. **Omitted Variables**: The model does not include crucial data like video transcript sentiment, audio analysis, or external promotion efforts.

4.2 Recommendations for Content Creators

Based on the pre-publication feature analysis, we offer the following actionable recommendations:

1. **Advanced Thumbnail Design:** Go beyond simple brightness. Our PCA showed that the interplay of RGB colors (`rgb_intensity`), colorfulness, and visual complexity are key. Experiment with specific color palettes that stand out.
2. **Strategic Title Engineering:** The data shows a measurable impact from title structure. A/B test titles that include numbers, are framed as questions, and are of optimal length for your niche.
3. **Optimize for Visual Complexity:** Find a balance in your thumbnails. The model identified `visual_complexity` (a mix of text, graphics, and people) as a significant component. Avoid both overly simplistic and cluttered designs.
4. **Align Video Length with Format:** The `VideoLengthCategory` was consistently important. Analyze your own analytics to see if a certain duration (e.g., shorts, mid-length) performs best and align content production accordingly.
5. **Leverage Publishing Time:** While not the strongest factor, optimizing `publish_day_of_week` and `publish_hour` for when your audience is most active is a simple, effective strategy.
6. **Build Channel Authority:** Channel-level metrics like subscriber count were part of the predictive model. Focus on long-term growth strategies to build a loyal subscriber base, which boosts the performance of individual videos.

5 Conclusion

This study successfully identified key factors influencing YouTube viewership through a combination of PCA and SVM. Our extensive feature engineering allowed us to move beyond simple metrics and quantify abstract concepts like thumbnail design and title strategy. The PCA distilled these into interpretable dimensions related to color, engagement, and content presentation. The SVM models demonstrated strong descriptive power ($R^2 = 0.642$) and, crucially, provided a realistic baseline for predictive accuracy (53.0%) by carefully avoiding data leakage. The results provide strong, data-driven support for content creators to focus their efforts on optimizable, pre-publication elements like thumbnail color composition and title structure to enhance viewership.

References

[1] H. E. Jang, S. H. Kim, J. S. Jeon, and J. H. Oh, "Visual Attributes of Thumbnails in Predicting YouTube Brand Channel Views in the Marketing Digitalization Era," *IEEE Transactions on Computational Social Systems, vol. 11,* July 2023.

[2] `https://github.com/zichuanxu/youtube-analysis/`