Predicting YouTube Video Views and Analyzing Influential Factors Using Machine Learning

Jin Nakamura, Meng Siyuan, Haruto Ohto, and Ryotaro Hada
Graduate School of Information Science and Technology, The University of Osaka

***Abstract*—The number of views on a YouTube video is a key indicator of a channel's popularity and influence. The objective of this research is to predict video view counts using machine learning, and to identify the primary factors that determine these counts and elucidate their relative importance. To achieve this objective, we constructed a custom dataset for 6,062 videos, encompassing basic metadata and visual features extracted from thumbnail images. Using this dataset with the video view count as the target variable, we compared multiple predictive models and adopted LightGBM, which demonstrated the best performance. The constructed model achieved a predictive accuracy with a coefficient of determination of $R^2 = 0.4528$ in 5-fold cross-validation. Furthermore, an analysis of feature importance revealed that the number of channel subscribers is overwhelmingly the most important variable for predicting view counts. Subsequently, video duration, and visual features such as the thumbnail's colorfulness and brightness, were also shown to have a significant impact on view counts.**

***Index Terms*—YouTube, Views, LightGBM**

I. Introduction

YOUTUBE [1] has evolved from a simple video-sharing platform into the world's largest digital media ecosystem, with over 2.7 billion active users consuming billions of hours of content daily [2]. This unprecedented scale has transformed YouTube from an entertainment platform into a critical business infrastructure where content creators, brands, and organizations compete for audience attention and engagement. The number of video views has emerged as the primary metric of success, directly correlating with advertising revenue, brand influence, and marketing effectiveness [3] [4].

For content creators, understanding the factors that drive viewership represents a fundamental challenge with significant economic implications. A single viral video can transform an unknown creator into a global influencer, while established channels struggle to maintain consistent viewership despite substantial resources. This unpredictability highlights the complex interplay of factors influencing viewer behavior, ranging from obvious elements like content quality and thumbnail design to subtle factors such as posting timing and algorithmic recommendations.

Previous research has identified several potential factors affecting YouTube video performance, such as the visual attributes of thumbnails, metadata optimization, and channel characteristics [5] [6] [7]. For example, Jang et al. [5] demonstrated the significance of thumbnail visual features in predicting brand channel views, establishing a foundation for understanding the role of visual appeal in viewer engagement. However, existing studies have often focused on single factors or specific content categories, leaving a gap in comprehensive, multi-factor analysis across diverse video types.

The challenge of predicting YouTube views is further complicated by the platform's sophisticated recommendation algorithm, which considers numerous undisclosed factors to determine content visibility [8]. The influence of this algorithm, combined with the viral nature of social media sharing and the subjective quality of content, creates a prediction problem characterized by high dimensionality and significant noise. Furthermore, existing public datasets, such as those available on Kaggle [9], also present challenges, as they may lack important features like ”subscriber count” or ”thumbnail brightness,” or the data may be approximately seven years old and thus not reflective of current trends.

In response to these challenges, this study proposes a comprehensive machine learning approach to predict YouTube video views by integrating multiple data sources and analytical techniques. We combine traditional metadata analysis with computer vision techniques to extract meaningful features from thumbnail images, channel statistics, and temporal patterns. Our methodology compares several approaches, including models combined with Principal Component Analysis (PCA) [10] to identify the most influential factors and ultimately utilizes a gradient boosting model [11] to achieve robust prediction performance. The primary contributions of this research are threefold. First, to develop a comprehensive feature engineering pipeline that captures both explicit metadata and implicit visual characteristics. Second, to provide quantitative evidence on the relative importance of various factors affecting viewership. And third, to create and provide a unique new dataset that overcomes the shortcomings of existing datasets.

This research addresses a critical knowledge gap by providing content creators, marketers, and platform researchers with data-driven insights into the mechanisms of YouTube success. By identifying actionable factors that significantly influence viewership, our findings can inform strategic decisions in content creation, thumbnail design, and publishing optimization. Furthermore, our methodology establishes a framework for the systematic analysis of social media content performance that can be adapted to other platforms and media types.

II. Proposed Method

In this study, we propose a methodology for predicting YouTube video view counts, which consists of two main processes: A: Dataset and Feature Engineering, and B: Model Construction.

A. Dataset and Feature Engineering

The dataset used in this research is composed of basic metadata collected from YouTube, derived secondary features, and visual features extracted from thumbnail image analysis.

1) Primary Metadata and Channel Information: Primary metadata and channel information are the raw data collected directly through methods such as the YouTube API [12]. The main data points acquired are as follows:

- `views`: The cumulative number of views for a video. This serves as the target variable for prediction in our study.
- `subscribers`: The number of subscribers to the channel that posted the video.
- Other Information: Basic data used for video identification and generating temporal features, including `video_id`, `title`, `category_id`, and `published_at`.

2) Derived Feature Engineering: From the collected primary data, we engineered features suitable for model input.

- Content Features:
  - `video_duration`: The length of the video in seconds. As its distribution is right-skewed, we applied a log-transformation ($\log(x+1)$) before model input.
  - `tags_count`: The total number of tags assigned to a video. Since a significant portion of videos (58.7%) had no tags, this zero value was also utilized as meaningful information.
  - `description_length`: The character count of the video description. This feature also has a high percentage of zero values (35.3%).
- Temporal Features: Based on `published_at`, we engineered the following features to capture temporal patterns influencing viewership:
  - `days_since_publish`: The number of days elapsed from the publication date to the time of data collection. This feature is a powerful predictor, as a longer time span naturally provides more opportunities for views to accumulate.
  - `hour_published`: The hour of the day (0-23) when the video was published. This allows the model to learn the relationship between posting time and audience activity, as publishing during peak hours can affect initial viewership.
- Channel Features: The `subscribers` feature has a right-skewed distribution. To capture its impact on different scales, we used the raw `subscribers` count and also generated its log-transformed version, `log_subscribers`, as an additional feature.

3) Visual Features from Thumbnails: To quantify the impact of thumbnails on viewer click behavior, we used the OpenCV image processing library [13] to extract the following visual features from each thumbnail image.

- Color Features:
  - `brightness`: We quantified the overall brightness of an image by converting it from the BGR color space to HSV and calculating the mean value of the V (Value) channel across all pixels.
  - `colorfulness`: An index representing the diversity and vividness of colors in an image. It was calculated based on the formula $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}$, which utilizes the opponent color space of red-green and yellow-blue [14].
- Structural Features:
  - `object_complexity`: The total number of objects detected within the image using the YOLOv3 object detection model [15] was used as a feature.
  - `element_complexity`: A score representing the overall intricacy of visual components in the thumbnail. This was calculated by analyzing the number and density of keypoints and contours, serving as a proxy for the quantity of distinct graphical elements in the image.
- Facial Detection Features:
  - `has_face`: Using the Haar Cascade classifier [16] implemented in OpenCV, we generated a binary feature that is 1 if at least one face is detected and 0 otherwise.

4) Excluded Features: Features such as `likes` and `comment_count` were not used, as they represent post-viewing outcomes and are thus unavailable at the time of prediction. Using them would lead to data leakage. Furthermore, the full text of `title` and `description` was not utilized as a feature, since Natural Language Processing (NLP) was not implemented within the scope of this research.

B. Model Construction

1) Model Selection: In this study, we compared the performance of several machine learning models, including LightGBM [17], XGBoost [18], Random Forest [19], and Linear Regression. As a result, LightGBM, a gradient boosting model that balances high predictive performance with computational efficiency, was selected as our final model, as it demonstrated the best performance in a comparative analysis (see Section III-A for details).

2) Model Configuration: For the model configuration, we first addressed the distribution of the target variable, `views`. Given its large scale and right-skewed nature, we applied a log-transformation ($\log(x+1)$) to create `log_views`, which served as the final target variable to stabilize the training process. The main hyperparameters for the adopted LightGBM model were optimized through cross-validation, resulting in the following settings:

- `num_leaves`: 31
- `max_depth`: 6
- `learning_rate`: 0.05
- `n_estimators`: 200

3) Evaluation Framework: We established a robust framework to evaluate the model. The primary metric for measuring predictive accuracy was the Coefficient of Determination (R^2), which indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

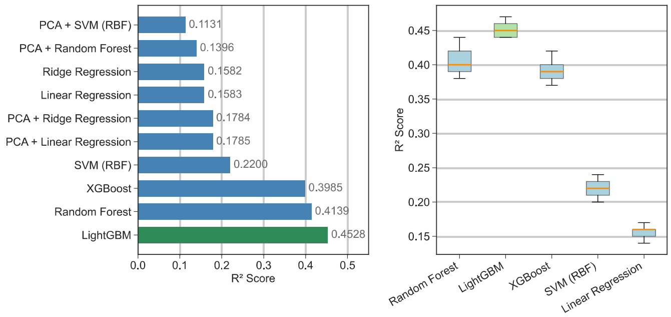
To objectively assess generalization performance, we split the entire dataset into a training set (80%) and a test set (20%). Subsequently, we conducted 5-fold Cross-Validation on the training data to evaluate the mean performance and its variance, thereby confirming the model's robustness.

III. Evaluation Results

This section reports the experimental results obtained using the methodology proposed in the previous section and discusses the insights derived from them.

A. Model Comparison

To select the final predictive model, we first conducted a comparative analysis of several machine learning algorithms, including LightGBM, Random Forest, XGBoost, and various linear models. A comprehensive summary of this comparison is presented in Figure 1.



(a) Model Performance Comparison

(b) Cross-Validation Score Distribution

Fig. 1. Comprehensive performance comparison across various models

As the results in Figure 1 indicate, the LightGBM model achieved the highest predictive performance with a Test R^2 score of 0.4528. The comparison also revealed that models combined with Principal Component Analysis (PCA) did not achieve high predictive performance. For example, when combined with PCA, the Random Forest model achieved an R^2 score of 0.1396, whereas the model without PCA reached 0.4139. This suggests that critical predictive information was not sufficiently captured after dimensionality reduction via PCA. While other non-PCA models like Random Forest ($R^2 = 0.4139$) also showed strong results, LightGBM demonstrated the most robust generalization performance. Therefore, based on its superior predictive accuracy, we selected LightGBM for the final analysis.

B. Predictive Performance

Applying the LightGBM model to the final complete dataset resulted in a mean Coefficient of Determination (R^2) of 0.4528 ± 0.0158 from 5-fold cross-validation. This indicates that our model can explain approximately 45% of the variance in video view counts.

Table I details the model's performance on a specific 80/20 training and test split. The difference in R^2 scores between the training set (0.6385) and the test set (0.4528) is 0.1857, which suggests that overfitting is controlled within an acceptable range.

TABLE I
Model Performance on Training and Test Sets

Data	R^2 Score
Training Set	0.6385
Test Set	0.4528

C. Analysis of the Dataset

1) Dataset Comparison: The results in Table II highlight the critical impact of specific dataset characteristics on model performance. The most decisive factor is the presence of the `subscribers` feature. A direct comparison between Dataset 2 and Dataset 3 reveals this. Adding the `subscribers` feature catapulted the R^2 score from 0.2575 to 0.4528, a 75.8% increase, confirming its overwhelming importance.

Furthermore, when the critical `subscribers` feature is included, the benefit of a larger sample size becomes evident. Comparing Dataset 1 with Dataset 3 shows that increasing the sample size improved the R^2 score from 0.3239 to 0.4528. This demonstrates that while feature quality is paramount, a larger dataset also contributes to a more robust and accurate model.

TABLE II
Dataset Comparison and Performance

Dataset	**Sample Size**	**Presence of `subscribers`**	**R^2 Score**
1	767	Yes	0.3239
2	6,062	No	0.2575
3	6,062	Yes	0.4528

2) Feature Importance: An analysis of the feature importance in the final model is presented in Table III. As the table shows, `subscribers` is the most important feature by a large margin. This is followed by `days_since_publish` and `video_duration`, which also hold significant predictive power. Furthermore, key visual features from thumbnails, namely `brightness` and `colorfulness`, were confirmed to have a substantial influence on predicting view counts.

TABLE III
Feature Importance of the Final Model

Feature	**Importance**	**Percentage**
`subscribers`	1013	19.3%
`days_since_publish`	680	13.0%
`video_duration`	659	12.6%
`brightness`	602	11.5%
`description_length`	511	9.8%
`colorfulness`	475	9.1%
`hour_published`	394	7.5%
`tags_count`	309	5.9%
`object_complexity`	207	4.0%
`log_subscribers`	194	3.7%
`element_complexity`	111	2.1%
`has_face`	84	1.6%

D. Practical Utility and Error Analysis

To evaluate the model's practical utility, we analyzed the errors between predicted and actual values. The key metrics are as follows:

- Median Relative Error: 68.8%
- Proportion of videos with error within 90%: 67.8%
- Proportion of videos with extreme prediction error (>500%): 11.1%

The median relative error of 68.8% indicates the difficulty of accurately pinpointing the view count for individual videos. However, the model predicts with an error margin of less than 90% for approximately two-thirds of the videos, suggesting it is capable of capturing general trends. On the other hand, the extreme prediction errors observed in over 10% of the videos suggest the existence of factors that our model cannot capture, such as unexpected viral hits.

IV. Discussion

Based on the evaluation results from the previous section, we present the following discussion.

- The Decisive Importance of Subscriber Count: The finding that `subscribers` is the most critical feature suggests that a channel's brand power and established fanbase heavily influence view counts, often more than the appeal of an individual video.
- The Efficacy of Visual Thumbnail Features: The high importance of `colorfulness` and `brightness` suggests that visually appealing thumbnails are effective in encouraging viewer clicks. Conversely, the very low importance of the `has_face` feature suggests the mere presence of a face is not a significant factor in predicting view counts.
- The Influence of Video Duration and Posting Time: As the third most important feature, `video_duration` significantly affects a viewer's decision to watch. Furthermore, the importance of `hour_published` highlights the strategic value of posting during periods of peak audience activity.
- Model Limitations and Practical Implications: The model's R^2 score of 0.4528 and median relative error of 68.8% reflect the prediction difficulty, likely due to unobserved variables like recommendation algorithms and content trends. Despite these limitations, the identified factors hold practical value as strategic guidelines for content creators.

V. Conclusion

In this study, we developed a LightGBM model to predict YouTube view counts, achieving an R^2 score of 0.4528. The analysis confirmed that channel subscribers are the overwhelmingly dominant factor, followed by video duration and thumbnail visual features. This highlights the importance of an established fanbase over individual video characteristics.

While the model's predictive power is limited by unobserved factors like viral trends, our findings on the importance of channel scale and thumbnail optimization offer actionable guidelines for content creators. Future work should focus on introducing time-series models, adding NLP analysis, integrating external data, and applying causal inference.

Author Contributions

While each member had primary responsibilities as detailed below, all authors contributed equally to this work.

- **Jin Nakamura:** Project Lead and Data Analyst, responsible for overseeing the project, as well as for implementing the machine learning models and conducting the experiments.
- **Meng Siyuan:** Data Collector, responsible for gathering the YouTube video data used in the analysis.
- **Haruto Ohto:** Report Writer, responsible for drafting and compiling the final report based on the analysis results.
- **Ryotaro Hada:** Report Writer and Presenter, responsible for the final editing of the report and for preparing the presentation of the findings.

References

[1] Google LLC, "Youtube," accessed: 2025-07-23. [Online]. Available: https://www.youtube.com/

[2] Anadolu Ajansı, "Youtube counts 2.7b active users as it turns 19," accessed: 2025-07-23. [Online]. Available: https://www.aa.com.tr/en/world/youtube-counts-27b-active-users-as-it-turns-19/3139073

[3] S. Ni, "Exploring the preferences of us youtube users and factors related to youtube uploader's revenue," *Studies in Social Science & Humanities*, vol. 2, no. 1, p. 43–54, Jan. 2023.

[4] Google LLC, "YouTube Partner Program overview & eligibility," accessed: 2025-07-23. [Online]. Available: https://support.google.com/youtube/answer/72851

[5] 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, no. 6, pp. 8169–8177, 2024.

[6] W. Tafesse, "Youtube marketing: how marketers' video optimization practices influence video views," *Internet Research*, vol. 30, no. 6, pp. 1689–1707, 2020.

[7] M. Bärtl, "Youtube channels, uploads and views: A statistical analysis of the past 10 years," *Convergence: The International Journal of Research into New Media Technologies*, vol. 24, pp. 16–32, 02 2018.

[8] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, ser. RecSys '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 191–198.

[9] Kaggle, "Find Open Datasets and Machine Learning Projects," accessed: 2025-07-23. [Online]. Available: https://www.kaggle.com/datasets

[10] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[11] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, 11 2000.

[12] Google LLC, "Youtube data api," accessed: 2025-07-23. [Online]. Available: https://developers.google.com/youtube/v3

[13] OpenCV team, "OpenCV - Open Computer Vision Library," accessed: 2025-07-23. [Online]. Available: https://opencv.org/

[14] D. Hasler and S. Suesstrunk, "Measuring colourfulness in natural images," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5007, pp. 87–95, 06 2003.

[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 04 2018.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–I.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.

[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 5–32, Oct. 2001.