Youtube Trending Prediction

Navyam Garg
navyam22317@iiitd.ac.in

Saksham Kapoor
saksham22431@iiitd.ac.in

Tizil Sharma
tizil22543@iiitd.ac.in

Abstract

This project focuses on the fast growth of digital media and how viral videos on YouTube can make a big impact. Predicting which videos will trend is useful for improving how people engage with content and how creators can earn money. Using machine learning, this project looks at how data can help predict trends in a constantly changing online space.

1. Introduction

The rise of social media platforms has transformed content creation and consumption, with YouTube standing out as one of the most influential video-sharing platforms. Predicting whether a video will trend on YouTube has become an area of interest for content creators, marketers, and researchers alike. The ability to accurately forecast video trends can enhance strategies for content optimization, audience targeting, and marketing campaigns. This project aims to leverage machine learning techniques to analyze video data and predict the likelihood of a video becoming a trending topic on YouTube.

2. Literature Survey

2.1. Springer (2023): Machine Learning for YouTube Popularity Prediction

This study employs advanced machine learning techniques, particularly XGBoost, to predict the popularity of YouTube videos. The model integrates metadata features like video quality (HD/SD) and duration with traditional attributes such as views and likes. It uses feature fusion and selection to optimize computational efficiency and enhance prediction accuracy. The tuned XGBoost model achieves 88% accuracy, outperforming traditional decision trees. The study also emphasizes Min-Max normalization for handling data inconsistencies, demonstrating the importance of preprocessing in achieving robust predictive performance. **Paper Link**

2.2. MDPI (2023): Optimized Predictive Modeling for Video Virality

The research focuses on refining machine learning methodologies for predicting YouTube video popularity by leveraging metadata. It implements a tuned XGBoost algorithm that incorporates fused features derived through normalization and correlation-based selection. The study compares baseline models with tuned versions, highlighting significant improvements in precision and recall. The XGBoost model achieves high computational efficiency by reducing dimensionality while maintaining an accuracy of 88%, making it a benchmark for metadata-driven predictive modeling. **Paper Link**

2.3. Cedric Richier et al. (2015): YouTube Popularity Dynamics

This work introduces a classification-based approach to model and predict YouTube video popularity. By first categorizing videos into specific classes, the model captures the dynamic nature of popularity evolution. The predictive process incorporates metadata and employs advanced classification algorithms to reduce error rates compared to baseline methods. This early integration of categorization into the modeling process highlights the importance of structured data preprocessing in machine learning for popularity prediction. **Paper Link**

2.4. Quyu Kong et al. (2018): Hawkes Intensity for Popularity Forecasting

The paper presents HIPie, a system combining the Hawkes Intensity Process with machine learning to predict and analyze video popularity. The model uses metadata and historical popularity trends to forecast virality. HIPie integrates predictive models with an interactive visualization platform, enabling users to diagnose and compare the success of videos and channels. By leveraging real-time data processing and machine learning, it provides actionable insights into the mechanics of viral content. **Paper Link**

3. Dataset and preprocessing

Dataset Link Kaggle

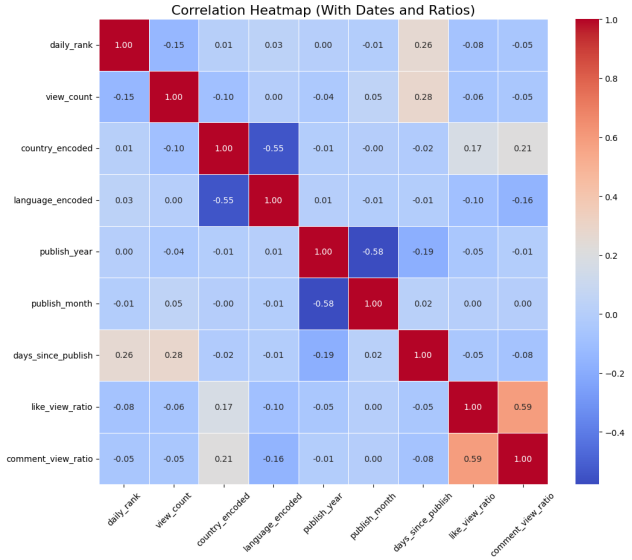

Figure 1

Overview

The dataset consists of a total of 2,025,984 entries, each representing a unique video tracked on the platform. It contains 18 columns that capture various metrics and attributes associated with these videos, allowing for in-depth analysis and insights.

Column Details

title, channel_name, daily_rank, daily_movement, weekly_movement, snapshot_date, country, view_count, like_count, comment_count, description, thumbnail_url, video_id, channel_id, video_tags, kind, publish_date, language

Preprocessing

The dataset was filtered to include data specific to India and the United States, representing a general population. Missing values were handled by applying a straightforward approach of dropping rows with `NaN` values to ensure data consistency and integrity for subsequent analysis.

EDA

We performed EDA analysis on the dataset, following are the plots:

4. Insights

- **Daily Ranks Distribution:** The distribution of daily ranks is approximately uniform, indicating that the ranking model performs consistently across the range of ranks.

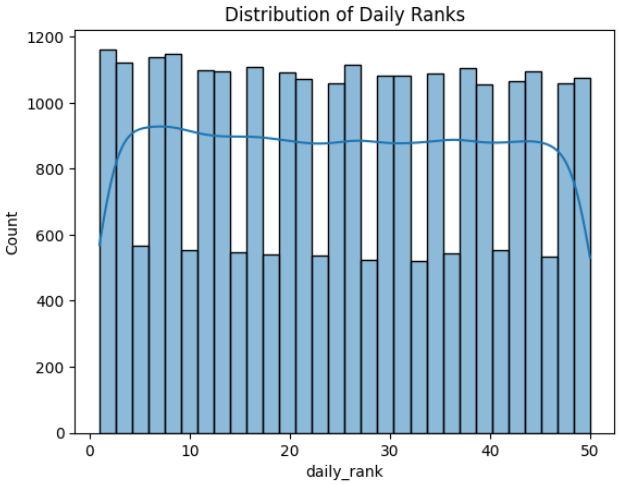

Figure 2

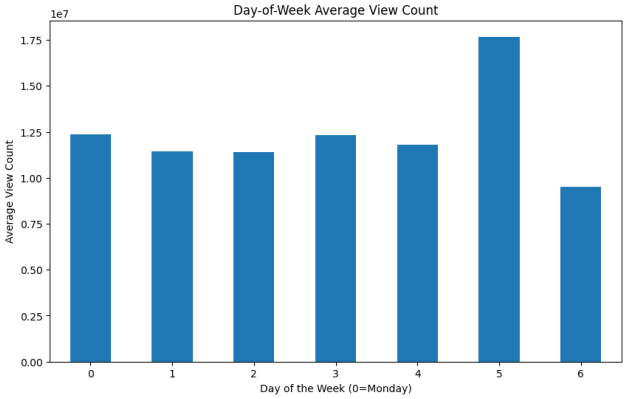

Figure 3

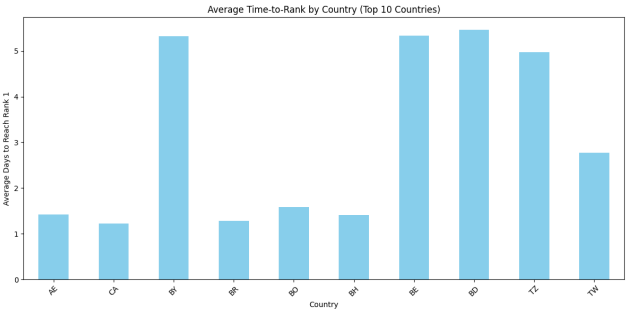

Figure 4

- **Day-of-Week Trends:** View counts peak on Fridays (day 5), suggesting higher user engagement leading into the weekend.
- **Time-to-Rank by Country:** Certain countries, such as Belarus (BY) and Belgium (BE), take significantly longer to reach Rank 1, highlighting possible regional

variations in content popularity.

- **Correlations:** The heatmap reveals that daily rank is weakly correlated with view count (-0.15) but more influenced by the days since publishing (0.26), indicating the importance of recency.
- **Ranking Model Performance:** The actual vs. predicted daily rank scatter plot shows a strong diagonal trend, suggesting the model predicts ranks effectively with minor deviations.

5. Methodology

5.1. Loss Functions

Loss functions are critical in evaluating the performance of regression models. They quantify the difference between the predicted values and the actual values, guiding the optimization process during model training. We employed several loss functions to comprehensively assess our model's performance.

5.1.1 Accuracy with an Off Prediction Allowance by 5

Definition: This metric calculates the proportion of predictions that are within a specified allowance (in this case, 5 units) of the actual values. This metric provides a practical assessment of the model's performance by considering predictions that are reasonably close to the actual values as accurate. It is particularly useful in scenarios where exact predictions are less critical than being within an acceptable range.

$$\text{Accuracy} = \frac{\text{Number of predictions with } |y_i - \hat{y}_i| \leq 5}{n} \times 100\%$$

5.1.2 Weighted Sub Accuracy

Definition: Weighted Sub Accuracy is a customized metric designed to emphasize the importance of certain predictions over others based on their deviation from a reference value (e.g., 50 in this context). Weighted Sub Accuracy takes into account not only whether a prediction is off by a certain amount but also how significant that error is relative to a reference point (50 in this case). This metric is particularly useful when errors in certain ranges are more critical than others, allowing for a more nuanced evaluation of model performance. **Explanation**: **Calculation of Weighted Differences**:

$$\text{diff2}[i] = |(y_{\text{test}}[i] - \hat{y}_{\text{rf}}[i]) \times (y_{\text{test}}[i] - 50)|$$

$$\text{diff2}[i] = \frac{\text{diff2}[i]}{50}$$

$$\text{diff2}[i] = \text{int}(\text{round}(\text{diff2}[i]))$$

5.1.3 Spearman's Rank Correlation

Spearman's Rank Correlation is a non-parametric statistical measure that assesses the strength and direction of the monotonic relationship between two ranked variables. It evaluates how well the relationship between the variables can be described using a monotonic function, with values ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

5.1.4 Kendall's Tau

Kendall's Tau is a rank-based correlation coefficient that measures the strength and direction of association between two ranked variables. It evaluates the agreement between rankings by counting concordant and discordant pairs, with values ranging from -1 (perfect disagreement) to +1 (perfect agreement).

5.1.5 NDCG (Normalized Discounted Cumulative Gain)

NDCG is a metric for evaluating ranking quality, commonly used in information retrieval. It assesses the relevance of ranked items by assigning higher importance to relevant items appearing earlier in the ranking, normalized against the ideal ranking for comparative purposes.

5.2. Linear Regression

What it is: Linear Regression is a simple statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It predicts the output based on the weighted sum of input features.

Why we used it: Linear Regression provides a baseline model to understand the relationship between variables and assess the predictability of the dataset. It is particularly useful for continuous output prediction and for interpreting feature significance.

5.3. Random Forest

What it is: Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and combines their outputs for improved accuracy and robustness. It reduces overfitting and improves generalization by averaging predictions from various trees.

Why we used it: Random Forest is well-suited for datasets with non-linear relationships and interactions between features. Its ability to handle missing data, rank feature importance, and manage large datasets makes it a powerful tool for predicting video popularity.

5.4. XGBoost

What it is: XGBoost (Extreme Gradient Boosting) is a high-performance implementation of gradient-boosted decision trees. It sequentially trains models, where each subsequent model corrects errors from the previous one, optimizing for both speed and accuracy.

Why we used it: XGBoost is known for its superior performance in structured data tasks, with high accuracy and efficiency. Its advanced features, like regularization, tree pruning, and handling of missing values, make it ideal for predicting complex patterns like video popularity trends.

5.5. GridSearch

What it is: GridSearch is a hyperparameter tuning technique that exhaustively tests all possible combinations of specified hyperparameter values to identify the best model configuration for performance optimization.

Why we used it: GridSearch ensures that the models (e.g., XGBoost, Random Forest) are fine-tuned for optimal accuracy and efficiency. By systematically exploring hyperparameter combinations, it helps in finding the configuration that minimizes errors and maximizes predictive capability.

6. Results and Analysis

The table compares the performance of XGBoost, Random Forest, and Linear Regression. XGBoost outperformed the others with the highest accuracy (36.18%) and weighted accuracy (72.05%), along with the strongest rank correlation (Spearman's 0.6289, Kendall's Tau 0.4614) and nDCG (0.9745).

Random Forest performed slightly below XGBoost with an accuracy of 34.19%, weighted accuracy of 71.49%, and similar rank correlation (Spearman's 0.6147, Kendall's Tau 0.4487). Its nDCG (0.9729) was also competitive.

Linear Regression lagged, with the lowest accuracy (24.32%) and rank correlation (Spearman's 0.4037, Kendall's Tau 0.2887), and an nDCG (0.9533) inferior to the other models.

6.1. Accuracy Summary

The following table summarizes the accuracy of each model based on various feature combinations:

The scatter plot compares the actual daily ranks against the predicted daily ranks for the XGBoost model. The diagonal trend indicates that the model captures the rank distribution well, though some deviations suggest room for improvement.

7. Conclusion

In conclusion, XGBoost emerged as the best-performing model due to its superior accuracy, ranking correlation, and relevance, making it the ideal choice for applications requiring both classification and ranking. Random Forest offers a viable alternative with slightly reduced performance, while Linear Regression is not recommended for tasks of this nature due to its comparatively weak performance.

Table 1. Model Performance Comparison (Transposed)

Metric	XGBoost	Random Forest	Linear Regression
Accuracy (%)	36.18	34.19	24.32
Weighted Acc.	72.05	71.49	66.98
Spearman's RC	0.6289	0.6147	0.4037
Kendall's Tau	0.4614	0.4487	0.2887
nDCG	0.9745	0.9729	0.9533

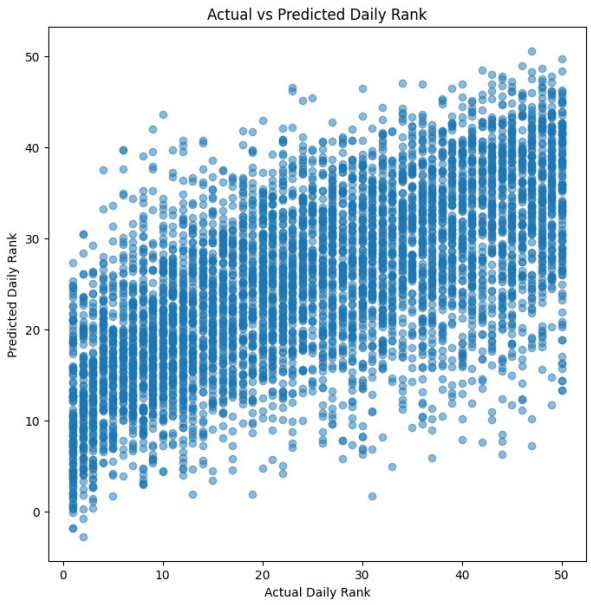

Figure 5. Actual vs Predicted Daily Rank for XGBoost Model

8. Contributions

- Saksham : Data Analysis, EDA, Model Training, Evaluation
- Tizil : Preprocessing, Model Development and Reporting
- Navyam : Literature Review, Loss search, GridSearch