YouTube Video Performance Analytics

Exploratory Data Analysis & View Prediction

Abhishek Verma

Table of contents

Project Overview

- **Goal:** Understand what drives YouTube video performance and build a model to predict view counts.
- **Dataset:** YouTube Video Analytics (Kaggle).
- **Approach:**
- Data cleaning & feature engineering
- Exploratory Data Analysis (EDA)
- Machine learning using Random Forest
- **Deliverables:** Insights, performance model, visualizations.

Dataset Structure

Each row represents a **single YouTube video** with:

- **Engagement metrics:**
  `view_count`, `like_count`, `comment_count`
- **Content metrics:**
  `duration_seconds`, `video_age_days`
- **Derived ratios:**
  `engagement_rate`, `likes_to_views_ratio`, `comments_to_views_ratio`
- **Target variable:**
  `view_count` (transformed using log1p)

These features allow us to analyze both *raw popularity* and *relative audience engagement*.

Preprocessing Steps

- No missing values in key metrics.
- Converted:
  - `published_at` → proper datetime
  - ISO-8601 duration text → numeric `duration_seconds`
- Dropped `favorite_count` (always 0).
- Engineered:
  - `video_age_days`
  - engagement ratios (likes/views, comments/views)

Why it matters:
Clean, numeric features allow accurate EDA and machine-learning modeling.

Distribution of Key YouTube Metrics

Interpretation

- All metrics are **heavily right-skewed**.
  → A small percentage of videos achieve very high views/likes/comments.
- Most videos fall within a moderate range with a **long tail of viral outliers**.
- Duration varies widely, showing multiple content styles:
  - short-form (<1 min)
  - mid-length (5–20 min)
  - long-form (1–3 hours)

Why this matters:
Skewed data makes linear regression unreliable — motivates the use of a **non-linear model** and **log-transform** for the target variable.

Correlation Between Metrics

Interpretation

- **View Count ↔ Like Count** → Strong positive correlation. Larger viewership tends to generate more likes.
- **Comments** correlate positively but more weakly.
- **Engagement rate** correlates differently from raw counts. High engagement does not always mean high views.
- Duration has weak correlation with views.

Implication:
View prediction requires a model that can capture **interaction effects**, not just linear trends.

Views vs Likes

Interpretation

- Clear upward trend: more likes → more views.
- Dense cluster of “average” videos + a few extreme viral hits.
- Strong evidence that **likes are a leading indicator** of video performance.

Views vs Comments & Duration

Interpretation

- Comment counts rise with views but scatter is high.
- Duration shows **no clear linear pattern** with views:
  - Very long videos do not necessarily perform well.
  - Viewers may prefer relevance over length.

Conclusion:
Comments and duration help, but are not dominant predictors.

Viewer Engagement Behavior

Interpretation

- Duration is plotted on a **log scale** for readability.
- Engagement rate is scattered across all durations.
- Very long videos often have **lower engagement**.
- No universal “best duration” exists.

Insight:
Audience response depends more on content quality than video length.

Modeling Strategy

Objective: Predict view_count using video features.

Steps (Part 1)

1. **Train/Test Split** (80/20)
2. **Log Transform Target**
   `y = log1p(view_count)` to reduce skew.
3. **Selected Features (1/2):**

- like_count, comment_count
- engagement_rate

Modeling Strategy (cont.)

Steps (Part 2)

3. **Selected Features (2/2):**

- likes_to_views_ratio
- comments_to_views_ratio
- duration_seconds
- video_age_days

4. **Model Used: Random Forest Regressor**

- Handles non-linearity
- Robust to outliers
- Captures interactions between features

Model Performance

1. R^2 Score ≈ 0.79

- Model explains **~79% of the variation** in view counts.
- Strong performance for a dataset with extreme outliers.

2. RMSE ≈ 17 Million Views

- Typical prediction error is ~17M views in the original scale.
- This sounds large, but consider the data range:
  - Some videos exceed **200M–300M views**.
  - Viral outliers are hard for any model to predict precisely.

Interpretation

- For *normal videos*, predictions are quite accurate.
- For *viral hits*, uncertainty is naturally higher.

Feature Importance (Random Forest)

Key Insights:

- **`like_count` dominates** — strongest predictor of views.
- Engagement metrics (`engagement_rate`, ratios) matter next.
- Time-based and duration-based features contribute modestly.
- Overall: **Audience actions beat content length in predicting success.**

Prediction Distribution: Actual vs Predicted

Interpretation

- **Red dashed line** = perfect prediction (actual = predicted).
- Most points lie close to this line → good model fit.
- Large deviations occur only for:
- very viral videos
- extreme high-engagement or trending content

Conclusion:
The model is reliable for typical videos, less so for extreme outliers — which is expected.

Practical Insights for Creators

- **Likes** are the strongest signal of performance.
- Engagement (interactions relative to views) matters more than raw length.
- Video duration alone does not predict success.
- Viral hits are unpredictable — driven by external amplification (e.g., trends).

Limitations

- Dataset does not include:
- Thumbnails
- Titles / CTR (click-through rate)
- Audience retention
- External traffic sources
- Model does not understand **video content** (no NLP, no image analysis).

Future Work

- Add **NLP-based features** from titles and descriptions.
- Include publish timings & category information.
- Apply advanced models:
  - Gradient Boosting / XGBoost
  - SHAP for explainability
- Deploy the model in a **Streamlit app** for interactive what-if analysis.

Summary

- Completed full workflow:
  Cleaning → EDA → Modeling → Insights
- Achieved **$R^2 \approx 0.79$** using Random Forest.
- Identified **likes and engagement** as top predictors.
- Established a reproducible, extensible framework using:
- Jupyter Notebook
- Quarto presentation
- Project folder structure suitable for GitHub

Speaker notes