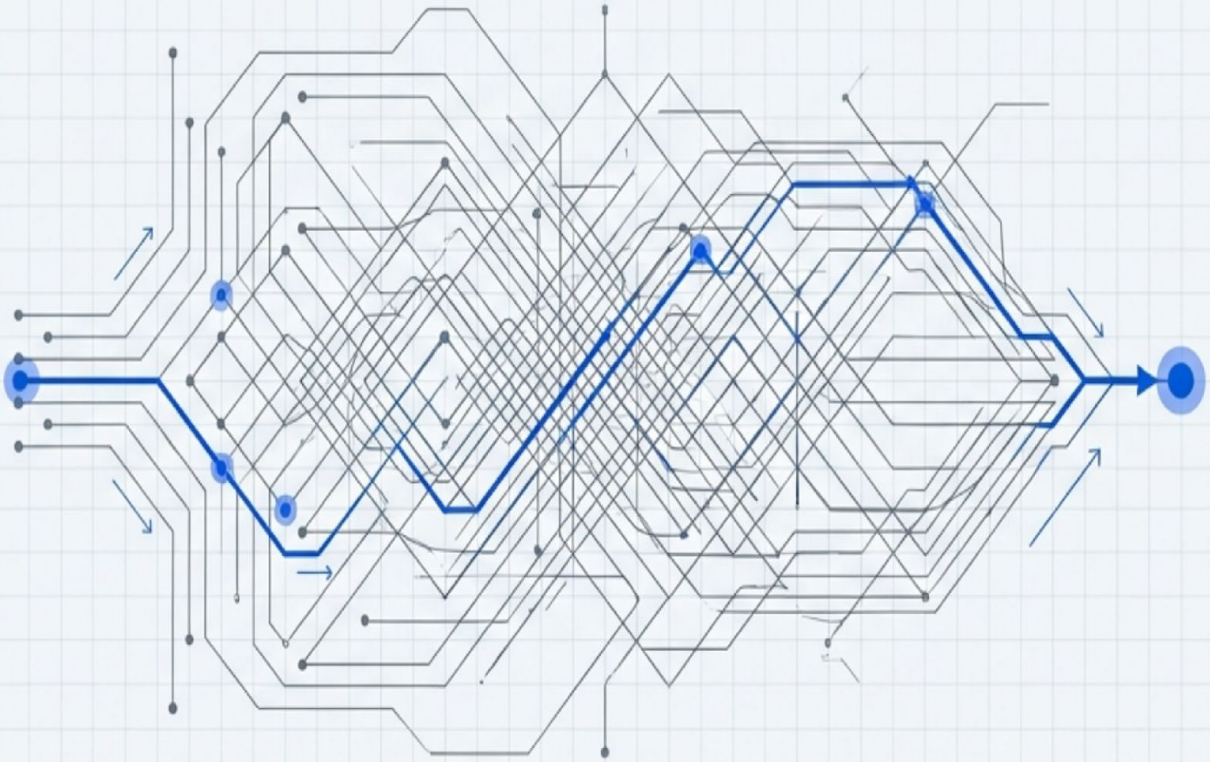


Cracking the Code of YouTube Success

An Analytical Investigation into the Drivers of Video Performance

The central question: What factors truly drive the performance of a YouTube video?

This analysis moves beyond conventional wisdom to build a predictive model model based on empirical evidence.

Goal

To understand the key drivers of YouTube video performance and build a robust model to predict view counts.

Evidence

A Kaggle dataset containing performance metrics for thousands of YouTube videos.

Method

Prepare the Evidence

Data cleaning and feature engineering.

Follow the Clues

Exploratory Data Analysis (EDA).

Build the Case

Machine learning with a Random Forest model.

Our investigation is built on a foundation of core video metrics.

The dataset provides two primary categories of evidence, which we use to derive further analytical ratios.

Engagement Metrics (Audience Actions)

- view_count (The Target Variable)
- like_count
- comment_count

Content Metrics (Video Attributes)

- duration_seconds
- video_age_days

Derived Ratios (Relative Engagement)

- engagement_rate
- likes_to_views_ratio
- comments_to_views_ratio

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

The first step in any investigation is to ensure the evidence is clean and reliable.

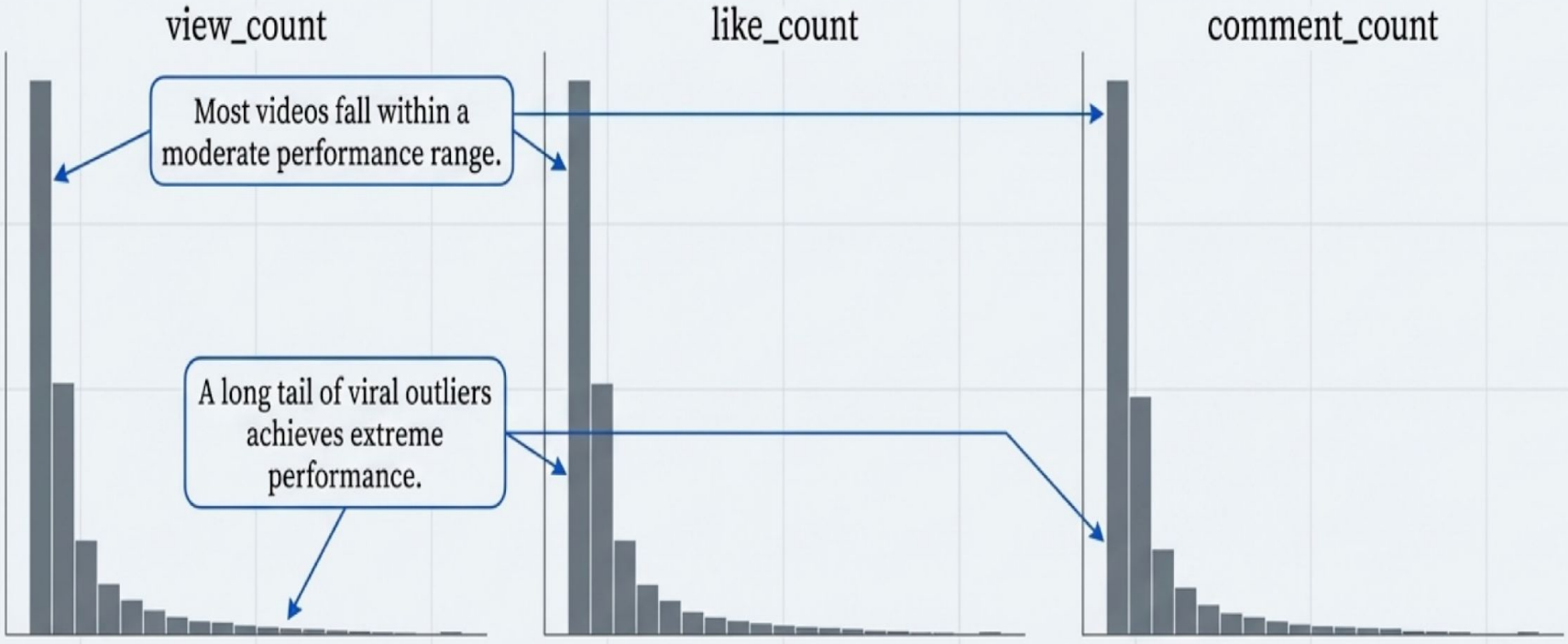
All key metrics were checked for missing values, and raw data was converted into a format suitable for robust analysis.

Key Actions	
Data Conversion	`published_at` → datetime; ISO-8601 duration → numeric `duration_seconds`.
Feature Engineering	`video_age_days` was calculated to account for time-based effects. Engagement ratios (e.g., likes/views) were created to measure audience interaction relative to reach.
Data Cleaning	`favorite_count` was dropped as it contained no variance (all values were 0).

Why this matters

Clean, numeric features are essential for accurate exploratory analysis and machine learning.

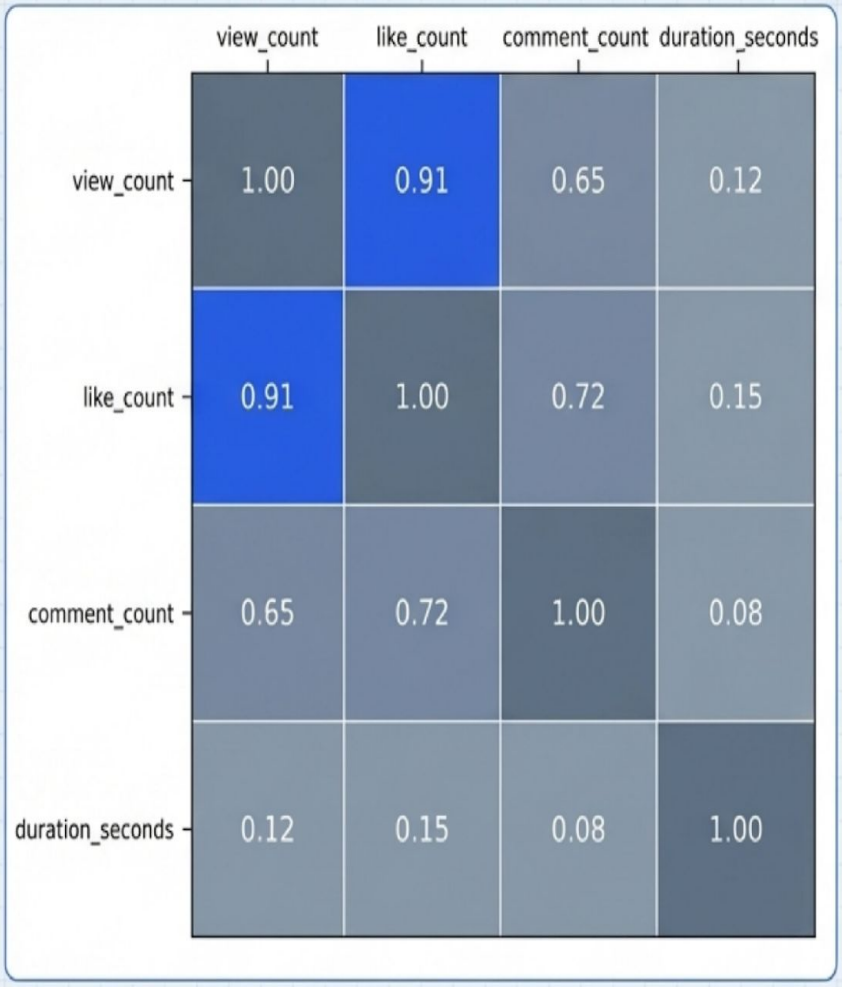
The landscape of YouTube performance is dominated by a few viral outliers.



All key metrics are heavily right-skewed, meaning a small percentage of videos earn a disproportionately high number of views, likes, and comments.

Implication: This skewness makes linear regression unreliable and motivates our use of a non-linear model (Random Forest) and a log-transformation for the target variable.

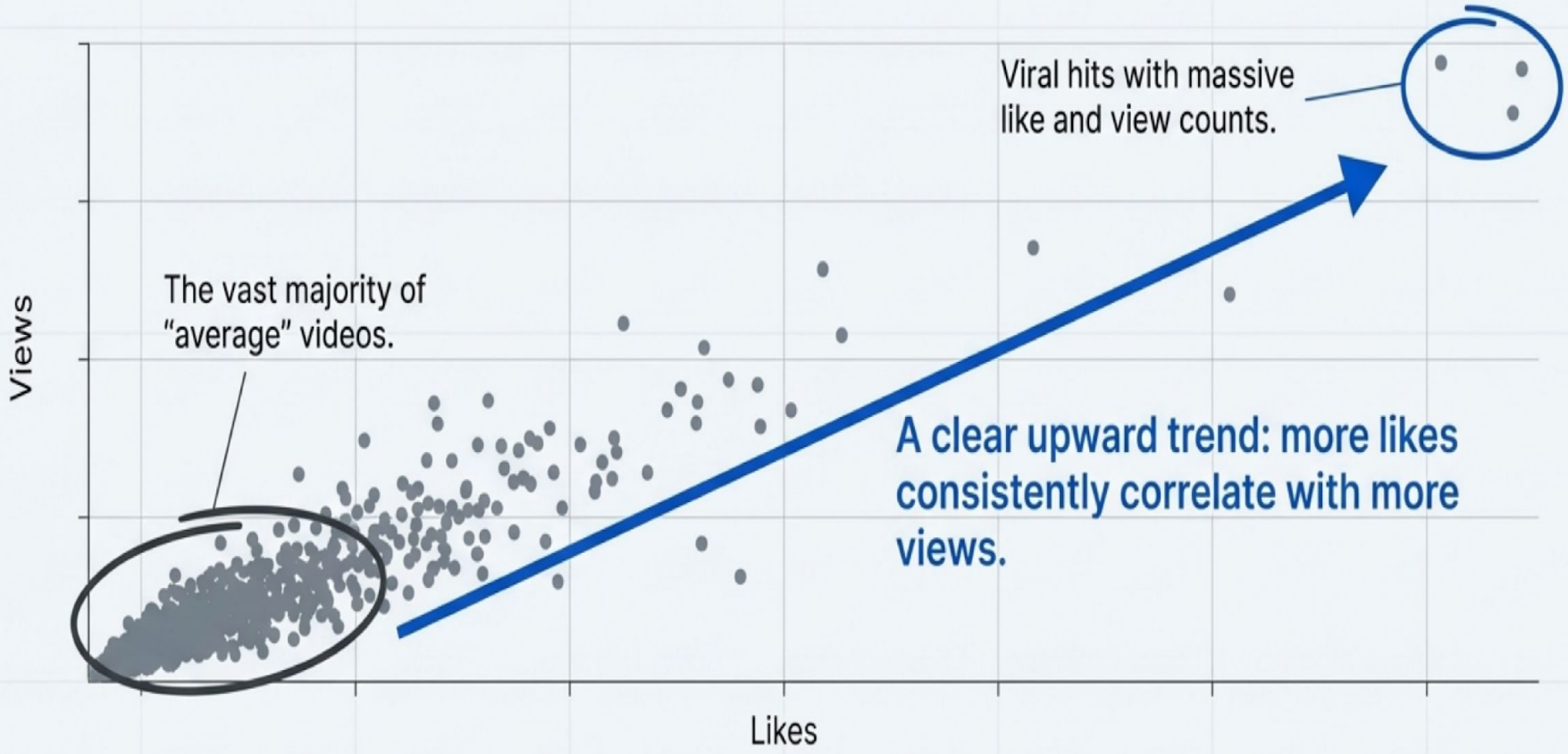
The first clue: Likes and views are inextricably linked.



- Strong Positive Correlation: `view_count` ↔ `like_count`. As one goes up, the other tends to follow suit very closely.
- Weaker Positive Correlation: `view_count` ↔ `comment_count`. A positive link exists, but it's less consistent than with likes.
- Weak Correlation: `view_count` ↔ `duration_seconds`. Video length shows little direct linear relationship with total views.

This initial finding suggests that audience approval (likes) is a much stronger signal of a video's reach than its basic attributes like length.

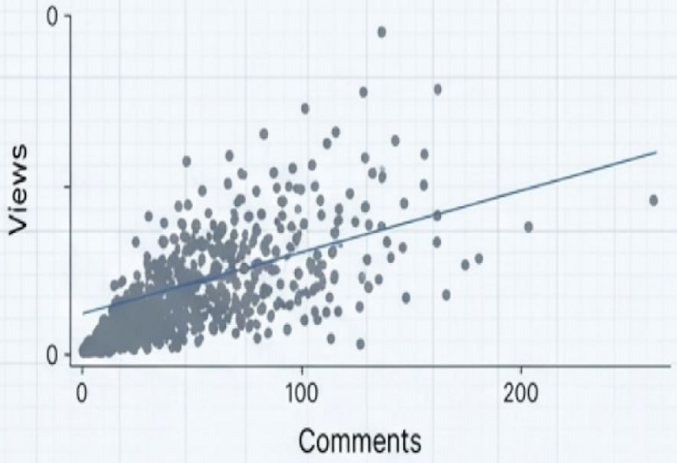
A closer look confirms likes are a leading indicator of video performance.



The visual evidence is clear. While not perfectly linear, the relationship between likes and and views is the most powerful one we've uncovered so far.

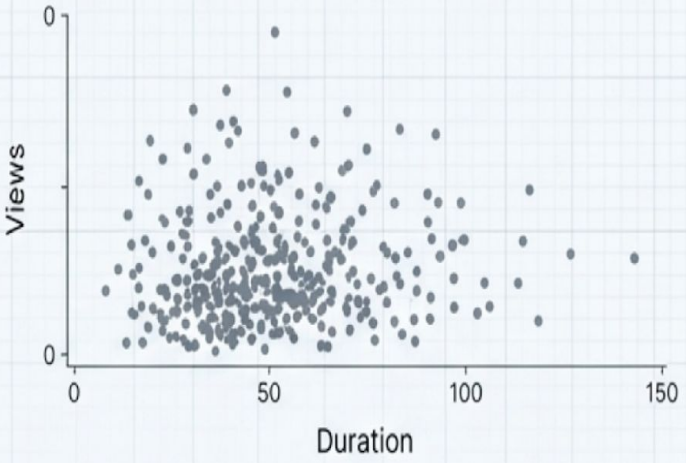
Other potential leads, like comments and duration, show a much weaker connection to views.

Views vs. Comments



Comment counts do rise with views, but the relationship is noisy and far less predictable than likes.

Views vs. Duration



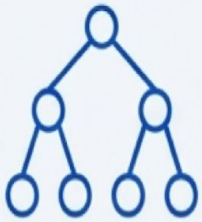
There is no clear linear pattern between duration and views. Very long videos do not necessarily perform better, suggesting viewers value relevance over length.

Comments and duration are part of the story, but they are not the dominant predictors we are searching for.

To connect all the clues, we built a predictive model to quantify what matters most.

We moved from exploration to prediction, training a Random Forest Regressor to learn the complex, non-linear patterns in the data.

Modeling Strategy



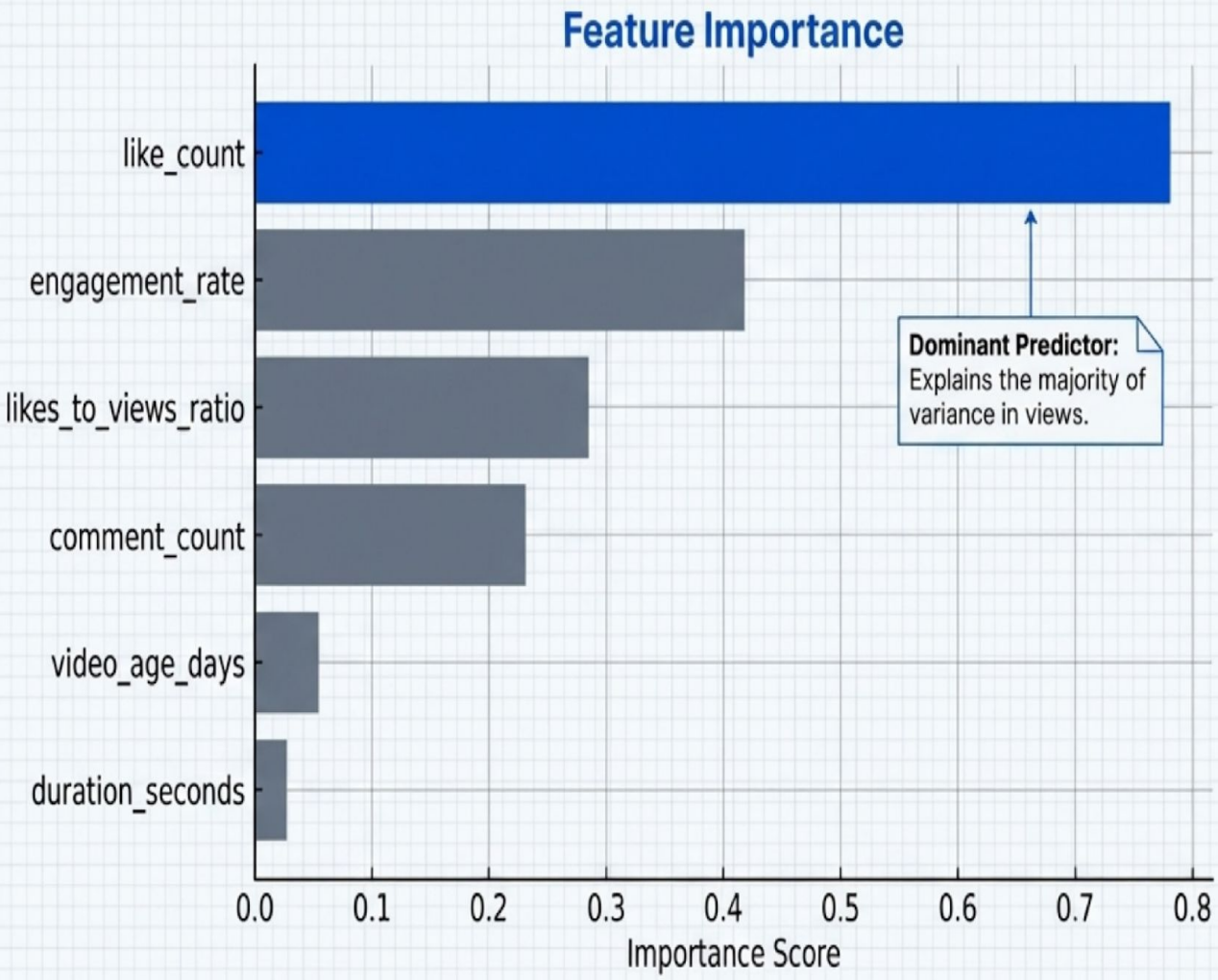
Model Choice: Random Forest Regressor.

- Why? It excels at handling non-linearity, is robust to outliers (like our viral hits), and automatically captures interactions between features.

Target Variable: `log1p(view_count)` to normalise the extreme skew.

Features Used: All key metrics were included: `like_count`, `comment_count`, `duration_seconds`, `video_age_days`, and all derived engagement ratios.

The verdict is in: Audience engagement, led by likes, is the the single most powerful predictor of success.



Key Insight

The model's findings are unequivocal. Audience actions, particularly likes, are far more predictive of a video's ultimate view count than inherent content attributes like its duration.

The top three features are all derived from audience engagement, highlighting its critical role.

Our model's predictions align closely with reality, explaining ~79% of the variation in views.

Performance Metrics

R^2 Score

≈ 0.79

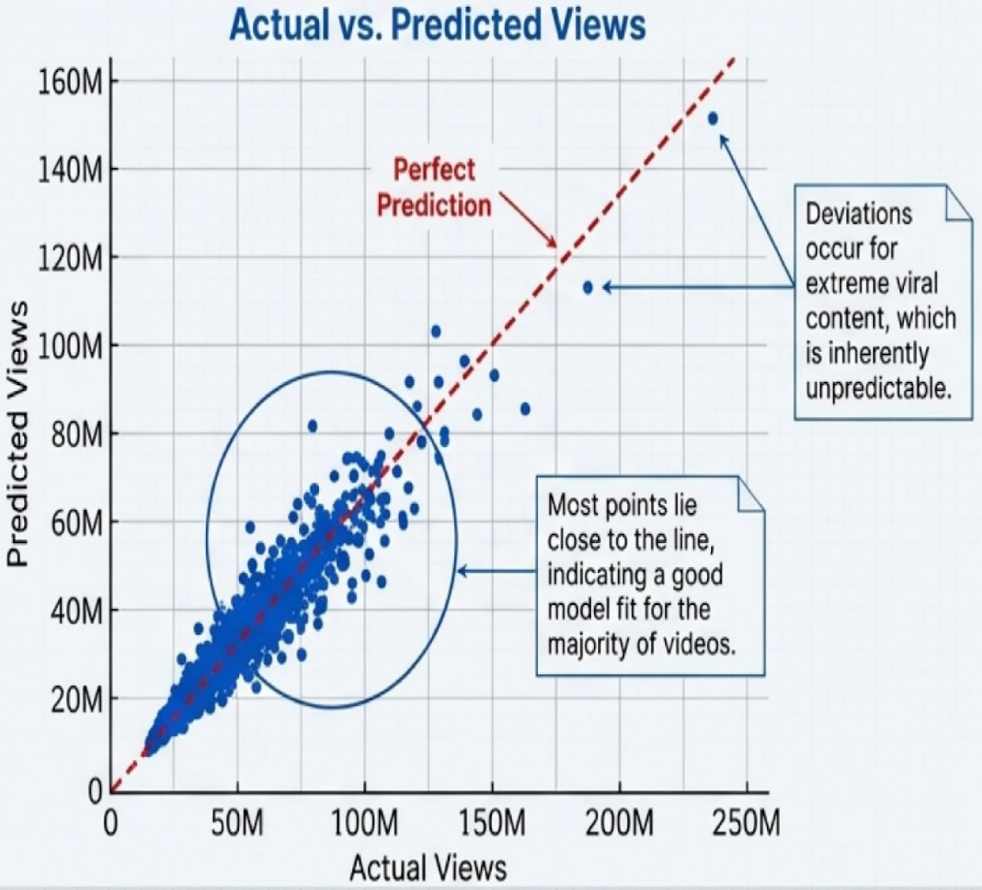
Interpretation: The model successfully explains approximately 79% of the variance in YouTube view counts.

RMSE

≈ 17 Million Views

Interpretation: While this number seems large, it is relative to a dataset where top videos exceed 200M views. The model is highly accurate for typical videos, with higher uncertainty for unpredictable viral hits.

Visual Confirmation



The Investigator's Briefing: Practical Insights for Content Creators

The analysis provides four key strategic takeaways for maximising video performance.

Focus on Likes

Likes are the single strongest signal of future performance. Content that actively encourages this form of engagement is likely to perform better.

Prioritise Engagement Over Length

Ratios like likes-to-views matter more than raw video duration. A shorter, highly engaging video is superior to a long, meandering one.

Duration is Not a Silver Bullet

There is no universal 'best duration.' Success is found across all video lengths. Focus on content quality that respects the viewer's time.

Acknowledge Unpredictability

Viral hits are, by nature, outliers driven by external factors (e.g., social trends, media coverage) that are difficult to model and cannot be planned for.

The case is closed, but the investigation continues.

The Unsolved Mysteries

This analysis did not include several critical data points that could provide deeper insights:

- **Content & Packaging:** Thumbnails, titles, descriptions (CTR).
- **Audience Behaviour:** Audience retention data.
- **Traffic Sources:** External vs. internal traffic drivers.

The Next Investigation

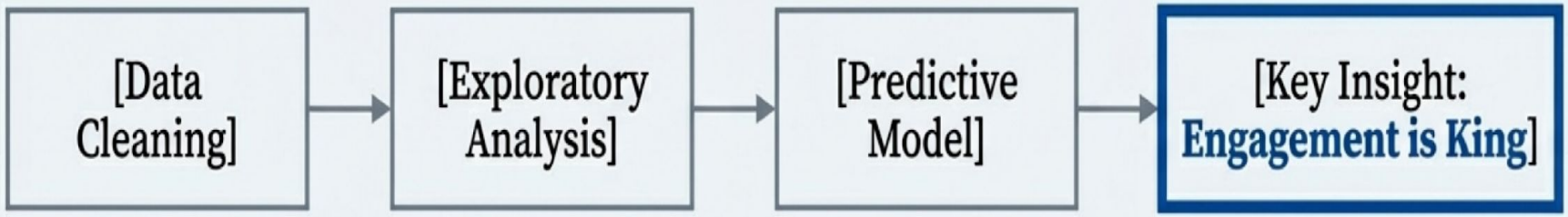
Future work can build on this foundation to create an even more powerful model:

- **Advanced Features:** Use NLP to extract features from titles and descriptions.
- **Smarter Models:** Implement Gradient Boosting (XGBoost) and use SHAP for enhanced explainability.
- **Interactive Tools:** Deploy the model in a Streamlit app for 'what-if' analysis.

From Raw Data to Actionable Insight: A Reproducible Framework

Summary:

- We conducted a full-stack data science workflow from data cleaning and EDA to predictive modelling.
- Our Random Forest model achieved a strong **R^2 of ≈ 0.79**, establishing a reliable predictive benchmark.
- The investigation concluded that **likes and engagement metrics are the definitive predictors** of YouTube video views, far outweighing content length.



Project built with a reproducible, extensible framework using Jupyter Notebook and Quarto.