# Predicting YouTube Video Performance: A Comparative Modeling Analysis

Final

Data Bootcamp

2025

# Predicting YouTube Video Performance

## Predictive Task

- Multi-class classification
- Predict YouTube view category: 0–10K, 10K–50K, 50K–100K, 100K–500K, 500K+

## Why This Is Hard

- Highly skewed view counts
- Nonlinear relationships
- Strong creator-level effects

# Baseline Model Comparison

**Models Evaluated**

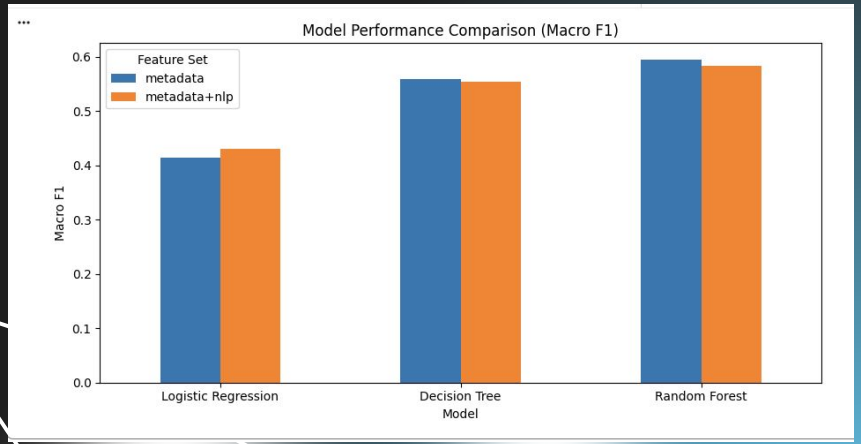
- Logistic Regression
- Decision Tree
- Random Forest

**Key Pattern (show bar chart of Macro F1)**

- Linear models underperform consistently
- Tree-based models dominate across feature sets

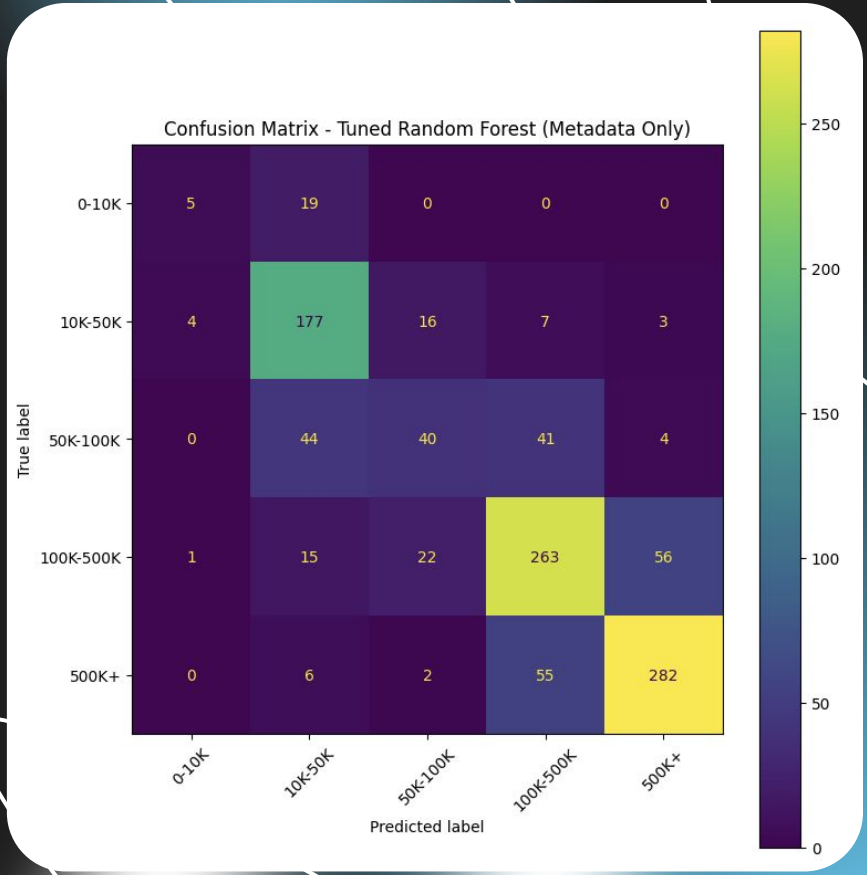
The poor performance of Logistic Regression suggests that video popularity is driven by nonlinear interactions between engagement metrics, timing, and channel identity. Tree-based models are better suited to capture these interaction effects, which explains their superior performance across both feature sets.

| | feature_set | model | accuracy | f1_macro |
|---|---|---|---|---|
| 3 | metadata | Random Forest | 0.723164 | 0.595428 |
| 2 | metadata | Decision Tree | 0.648776 | 0.558802 |
| 1 | metadata | Logistic Regression | 0.603578 | 0.414260 |
| 0 | metadata | Dummy (most frequent) | 0.336158 | 0.100634 |
| 7 | metadata+nlp | Random Forest | 0.724105 | 0.583325 |
| 6 | metadata+nlp | Decision Tree | 0.647834 | 0.553580 |
| 5 | metadata+nlp | Logistic Regression | 0.606403 | 0.431019 |
| 4 | metadata+nlp | Dummy (most frequent) | 0.336158 | 0.100634 |



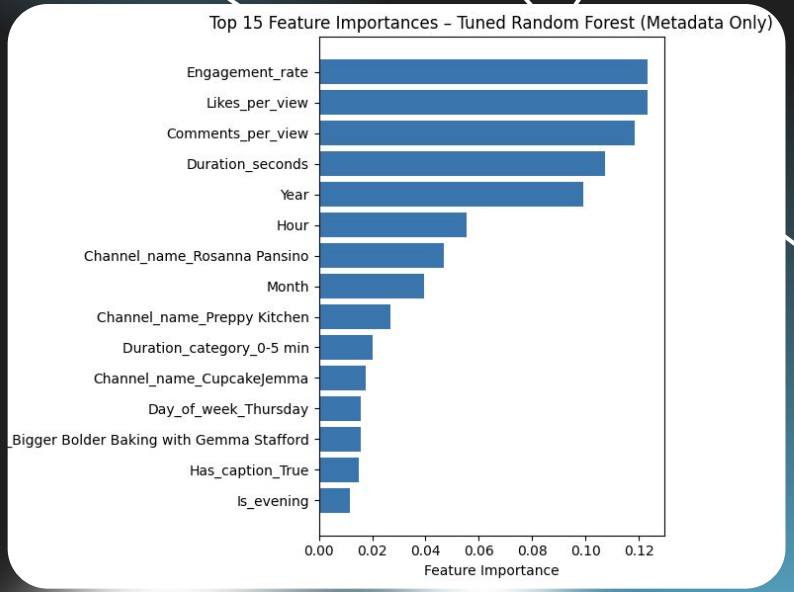
# Confusion Matrix Analysis — Random Forest

- **Excellent at the Top:** The model is highly accurate at identifying "viral" content (100K–500K and 500K+ buckets). If a video is a hit, the model catches it.
- **Struggles in the Middle:** The 50K–100K bucket is the weakest link. The model often confuses these "mid-tier" videos with their immediate neighbors (10K–50K or 100K–500K).
- Errors are "Safe": The mistakes are almost exclusively "off-by-one." The model virtually never confuses a viral giant (500K+) with a flop (0–10K). It captures the hierarchy, even when it misses the exact boundary.



# Feature Importance & Drivers

- **Interaction Efficiency is the Primary Signal.**
  - The top 3 features are all "per view" or rate-based metrics. Total volume doesn't predict success as well as the density of engagement.
- **Duration is Critical**
  - Duration_seconds is the #4 driver (and the #1 non-interaction feature). This suggests the model distinguishes between short-form viral hits and long-form "searchable" content.
- **Channel Identity is Secondary**
  - While creators like Rosanna Pansino appear in the top 10, they are below universal metrics. This implies the "rules of popularity" like high engagement apply generally, regardless of who is posting.
- **Year Matters**
  - Year is a top-5 feature, likely capturing the channel's growth trend over time (e.g., a video posted in 2023 naturally gets more views than one in 2018 due to subscriber growth).



# Testing a Higher-Capacity Model: Gradient Boosting

**Key Findings**

- Gradient Boosting achieves comparable Macro F1 to the tuned Random Forest.
- Despite greater model expressiveness, performance gains are marginal.
- This suggests diminishing returns from additional model complexity given the current feature set.

**Interpretation**

- Engagement and timing features already capture most of the predictive signal.
- Boosting does not substantially improve class separation for lower-frequency view categories.
- Model performance appears constrained by feature quality rather than model capacity.

| | feature_set | model | accuracy | f1_macro |
|---|---|---|---|---|
| 3 | metadata | Random Forest | 0.723164 | 0.595428 |
| 2 | metadata | Decision Tree | 0.648776 | 0.558802 |
| 1 | metadata | Logistic Regression | 0.603578 | 0.414260 |
| 0 | metadata | Dummy (most frequent) | 0.336158 | 0.100634 |
| 8 | metadata+nlp | Gradient Boosting (tuned) | 0.725047 | 0.617007 |
| 7 | metadata+nlp | Random Forest | 0.724105 | 0.583325 |
| 6 | metadata+nlp | Decision Tree | 0.647834 | 0.553580 |
| 5 | metadata+nlp | Logistic Regression | 0.606403 | 0.431019 |
| 4 | metadata+nlp | Dummy (most frequent) | 0.336158 | 0.100634 |

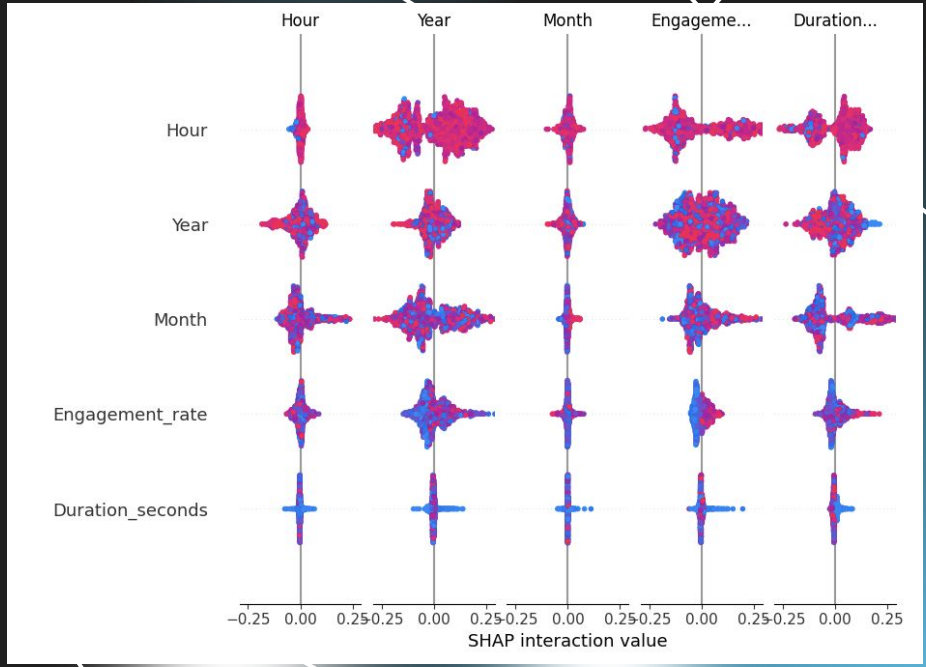
# Model Interpretability (SHAP)

SHAP values quantify how each feature contributes to a prediction relative to the model's baseline.

- **Engagement_rate is the dominant driver** of predicted view tier across all interactions.
- **Engagement_rate × Duration_seconds shows strong nonlinear effects:** engagement signals are interpreted differently for short vs. long videos.
- **Timing features (Hour, Month, Year) matter primarily in combination with engagement** rather than as standalone predictors.

Why this explains model performance

- These interaction patterns validate why tree-based models outperform Logistic Regression:
- the task depends on nonlinear and conditional relationships, which linear models cannot capture.



# Conclusion

**Learnings**

- Tree-based ensemble models (Random Forest, Gradient Boosting) consistently outperform linear models, confirming that YouTube video performance is driven by nonlinear feature interactions rather than additive effects.
- Engagement-based metadata (likes per view, comments per view, engagement rate) is the strongest predictor of performance, outweighing textual or sentiment-based signals.
- Adding NLP features provides limited marginal improvement, suggesting that sentiment scores may be noisy or insufficiently expressive for this task.

**Best Model**

Tuned Random Forest (metadata-only) was selected as the final model:

- Strong macro F1 performance
- Robust across popularity tiers
- More interpretable and stable than more complex boosting approaches

**Key Modeling Insight**

The model learns relative popularity well (misclassifications occur mostly between adjacent buckets), indicating that performance tiers exist on a continuum rather than as sharply separable classes.