BREAK THROUGH TECH

GOOGLE PROJECT

Predicting YouTube Virality Across 11 Countries

12/06/2025

Meet Our Team

Nancy Nakyung Kwak
Undergraduate Researcher
2nd Year Data Science
@ University of Texas at Austin

Woon Ket Wong
Challenge Advisor
Software Engineer
@ Google

Haziel Andrade Ayala
Break Through Tech
AI Studio Coach

Presentation Agenda

Motivation & Research Question	1 minute
Dataset & Definitions	4 minute
Exploratory Analysis: What Drives Virality?	4 minutes
Feature Engineering & Modeling	5 minutes
Results, Limitations & Next Steps	1 minute

How Rare Is ‘Viral’ Really?

- *“Only ~10% of videos in our dataset are ‘viral’ (top 1-% of views).”*

  But YouTube sees millions of uploads per day -> humans can’t review everything.
- **Question:**
  - **“Can we predict *early* which videos will go viral – and how quickly they’ll hit Trending – using just engagement and metadata?**

That’s what we set out to explore with 2.9M trending videos from 11 countries.

Why This Matters for YouTube & Creators

- YouTube's recommendation system heavily shapes *what* people watch.
- Creators want to know: *"What drives virality? What should I focus on?"*
- Google's mission: *"organize the world's information and make it universally accessible and useful"*

→ **our job is to turn raw trending data into actionable signals.**

Project Goal

- Build a model to **predict viral videos (top 10% by views)** using early engagement + metadata.
- Understand **what drives virality** and **how quickly** videos become trending across regions.

Data Understanding and Data Preparation

Dataset:

11 Countries, ~2.9M Trending Videos

```
US: shape after cleaning = (268704, 17)
GB: shape after cleaning = (268667, 17)
RU: shape after cleaning = (238539, 17)
BR: shape after cleaning = (268701, 17)
CA: shape after cleaning = (268633, 17)
DE: shape after cleaning = (268604, 17)
FR: shape after cleaning = (268646, 17)
IN: shape after cleaning = (251202, 17)
JP: shape after cleaning = (268629, 17)
KR: shape after cleaning = (265602, 17)
MX: shape after cleaning = (268528, 17)

Global dataset shape: (2904455, 17)
Out[2]:
```

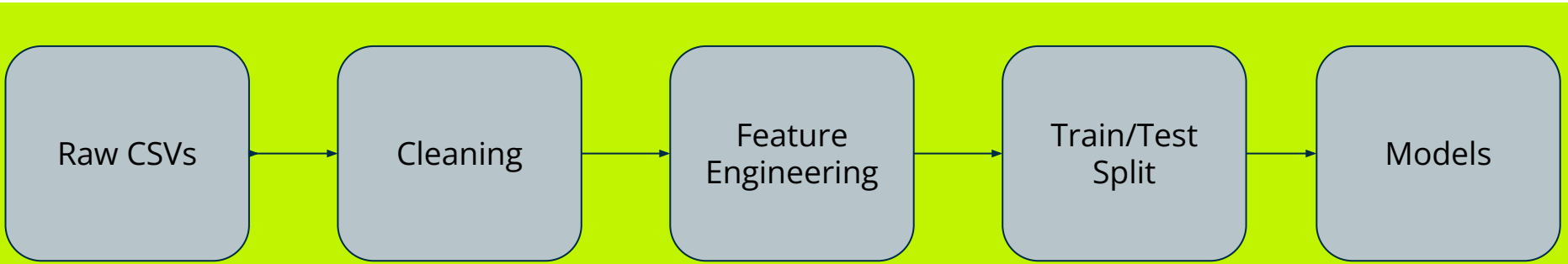
	video_id	title	publishedAt	channelId	channelTitle
0	3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis
1	M9Pmf9AB4Mo	Apex Legends | Stories from the Outlands – "Th...	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends
2	J78aPJ3VyNs	I left youtube for a month and THIS is what ha...	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g4vX-m6g	jacksepticeye

- Source:
  - Kaggle "YouTube Trending Video Dataset" (2020-2023)
  - https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset
- combined all country files

  → **single global dataset** with a "country" column
- Key rew fields:
  - timestamps, caegoryId, tags, view_count, likes, dislikes, comment_count, etc.

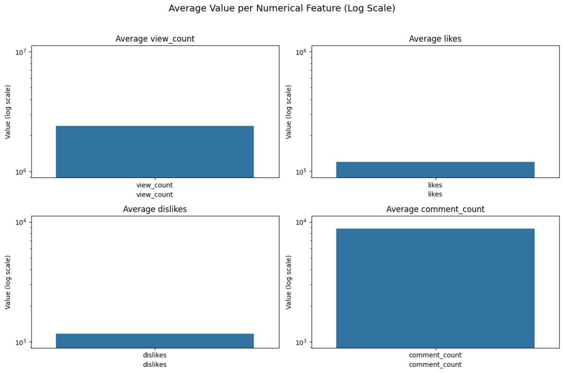
Cleaning & Feature Engineering (Global Pipeline)

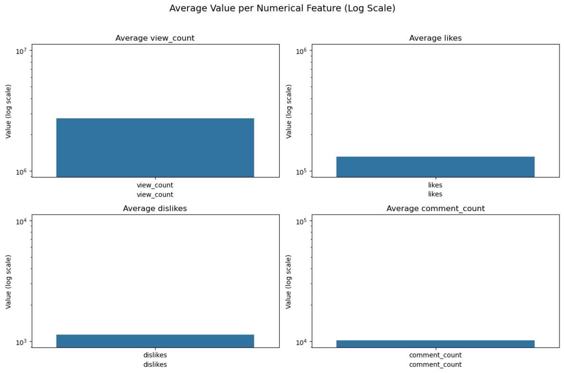
- Removed exact duplicates, filled missing descriptions with empty strings
- Converted timestamps to datetime, created:
  - “publish_hour”, “publish_dayofweek”
  - “time_to_trending_days”, “hours_to_trending”, “days_since_published”
- Lightweight text features:
  - “title_length” (characters) ; “tag_count” (split on |)
- **Target: viral = 1 if view_count in top 10% globally (threshold ≈ 3.54M views), else 0.**

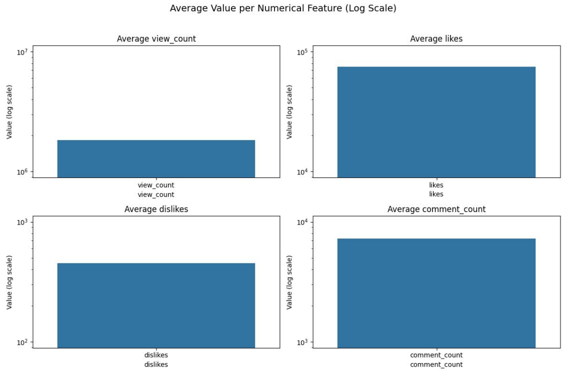


Engagement Distributions Are Extremely Skewed

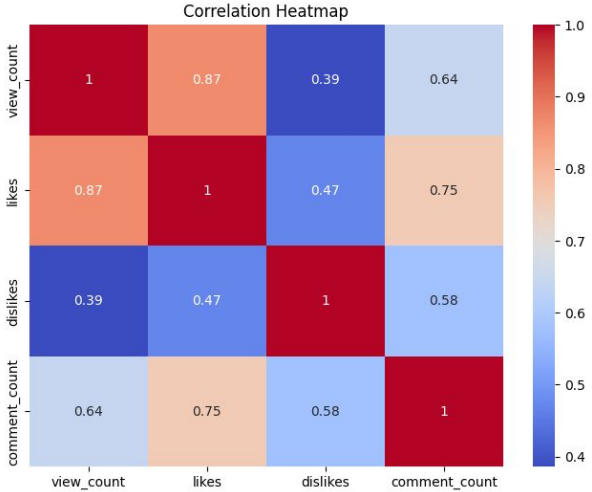
- Views, likes, and comments are all **heavily right-skewed**
- Vast majority of videos get **modest engagement**, while a tiny fraction explode (outliers)
- Motivates::
  - log transforms for regression & top 10% viral label for classification

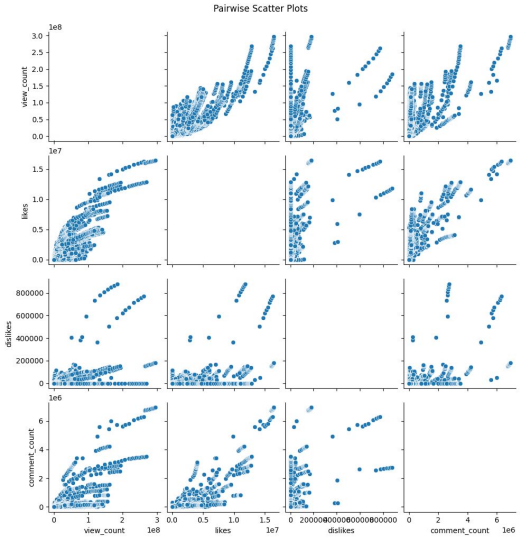






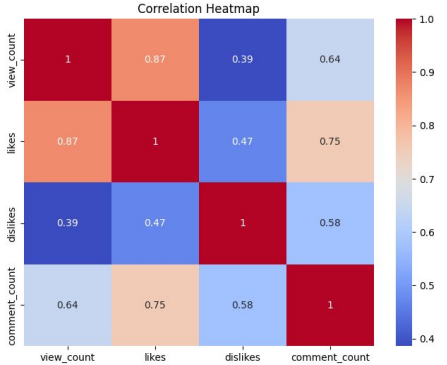
Engagement Features Move Together

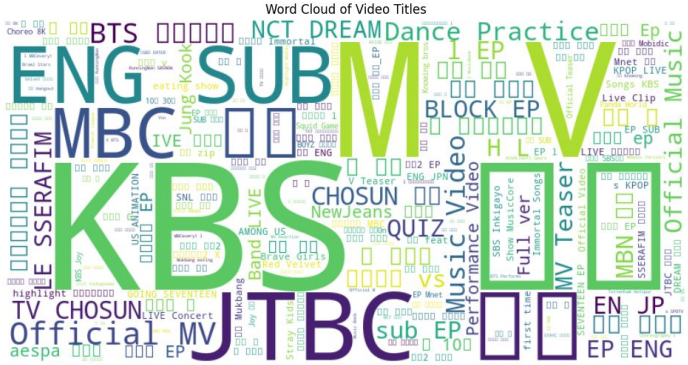


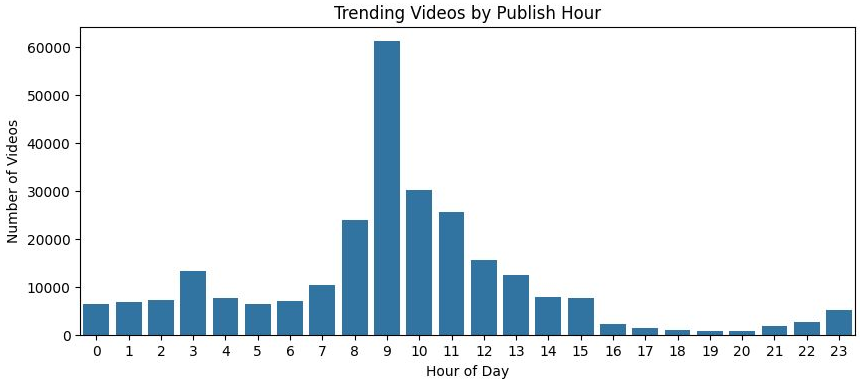


- Strong positive correlations between **views, likes, and comments**
- Scatter shows a tight increasing trend: higher likes → higher views, but with noise
- This validates using early engagement as primary predictors

Example: Korea (KR) EDA Highlights





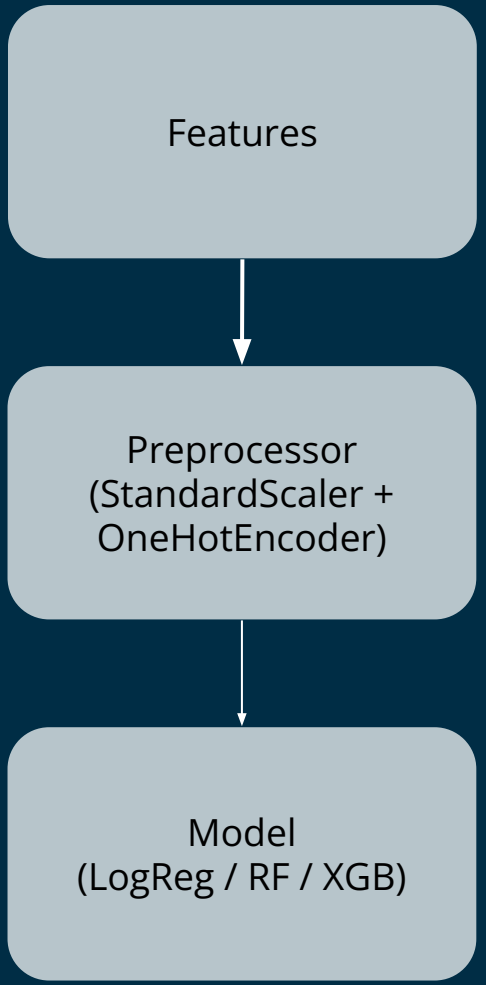


- Same centralized cleaning pipeline applied per-country
- In KR, trending content is dominated by **entertainment & K-pop**, with extreme BTS/BLACKPINK outliers
- Strong correlations mirror US/global patterns → suggests a **global model is feasible**

Modeling and Evaluation

Predicting Virality: Modeling Setup

- Features:
  - Numeric: likes, dislikes, comment_count, publish_hour, publish_dayofweek, time_to_trending_days, title_length, tag_count
  - Categorical: categoryId, country (one-hot encoded)
- Train/test split:
  - 80/20 with **stratification** on viral label (10% positives)
- Models compared:
  - Logistic Regression (baseline)
  - Random Forest Classifier
  - XGBoost Classifier



Global Viral Prediction Performance

	Model	Accuracy	Precision	Recall	F1	ROC-AUC
1	Random Forest	0.9883	0.9766	0.9046	0.9392	0.9980
2	XGBoost	0.9615	0.8621	0.7317	0.7916	0.9806
0	Logistic Regression	0.9474	0.8365	0.5896	0.6917	0.9521

- we used only early engagement + metadata
- viral label is roughly top 10% → strong signal, not a super noisy label
- Performance stays high when we slice by country and region (next slide)
- We compared simpler models: Logistic Regression gets ROC-AUC ≈ 0.95 – so RF isn't *inventing* patterns; it's adding nonlinearity on top of a strong baseline.

We chose RF as our primary model because it balances performance and interpretability

Top Global Features:
Engagement Rules Everything

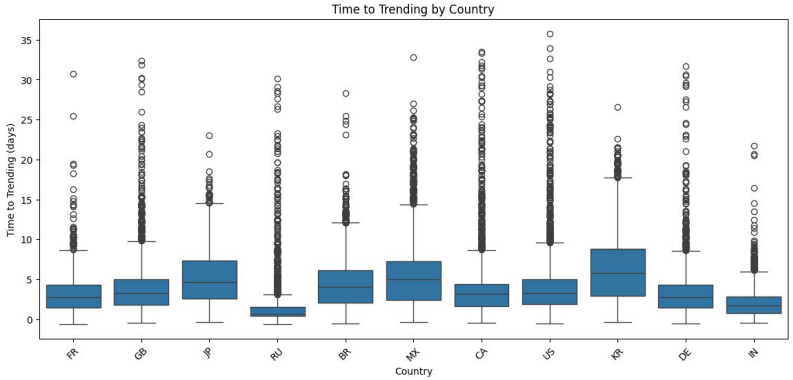
- **Likes** and **comments** are the top two features, accounting for most of the model's signal.
- Temporal & lightweight text features (time_to_trending, title length, tag count) matter, but less.
- Category and country add small, but non-zero, lift — virality is **mostly behavior,** not just geography or category

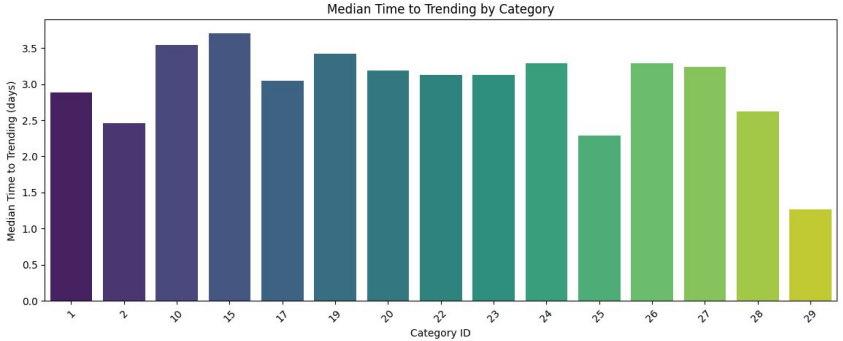
*** BAR CHART OF TOP 10**

RF IMPORTANCES FROM "fi_df"

	feature	importance
0	likes	0.3819
2	comment_count	0.2060
1	dislikes	0.0922
5	time_to_trending_days	0.0653
6	title_length	0.0555
7	tag_count	0.0492
3	publish_hour	0.0446
4	publish_dayofweek	0.0279
12	categoryId_17	0.0103
10	categoryId_10	0.0100

Time to Trending Varies by Country & Category





- Most videos hit Trending within the first **~5 days,** but there's widespread.
- **Fastest trending** markets: RU, IN, DE (lower medians, tighter boxes)
- Some categories (News & Politics, Sports) trend faster; others (Education, Travel) are slower
- This motivates predicting not just *whether* a video will be viral, but **how fast**

Model Generalizes Across Countries and Regions

	country	n_samples	viral_rate	accuracy	precision	recall	f1	roc_auc
3	FR	53696	0.0525	0.9959	0.9872	0.9340	0.9599	0.9997
7	KR	53037	0.0666	0.9953	0.9887	0.9400	0.9637	0.9995
2	DE	53643	0.0881	0.9938	0.9900	0.9393	0.9640	0.9991
4	GB	53452	0.1264	0.9899	0.9859	0.9334	0.9589	0.9990
6	JP	53797	0.0464	0.9954	0.9853	0.9155	0.9491	0.9990
10	US	53897	0.1424	0.9860	0.9861	0.9149	0.9492	0.9983
1	CA	53926	0.1382	0.9870	0.9845	0.9205	0.9514	0.9983
0	BR	53695	0.0751	0.9920	0.9816	0.9110	0.9449	0.9982
8	MX	53648	0.1423	0.9852	0.9711	0.9235	0.9467	0.9974
9	RU	47764	0.0448	0.9879	0.9373	0.7826	0.8530	0.9945
5	IN	50336	0.1753	0.9650	0.9348	0.8606	0.8962	0.9907

	region	n_samples	viral_rate	accuracy	precision	recall	f1	roc_auc
0	Europe	208555	0.0788	0.9920	0.9817	0.9156	0.9475	0.9986
1	Americas	215166	0.1245	0.9876	0.9806	0.9183	0.9485	0.9981
2	Asia	157170	0.0945	0.9857	0.9564	0.8887	0.9213	0.9969

- ROC-AUC for each country: **0.99+**

  → model ranks viral vs non-viral consistently across 11 markets
- Viral rate differs:
  - Europe ~7-8%
  - Americas ~12%
  - Asia ~9-10%
- One global models works well, but region-specific **decision thresholds** might be tuned in production (e.g. lower threshold in Europe where viral events are rare)

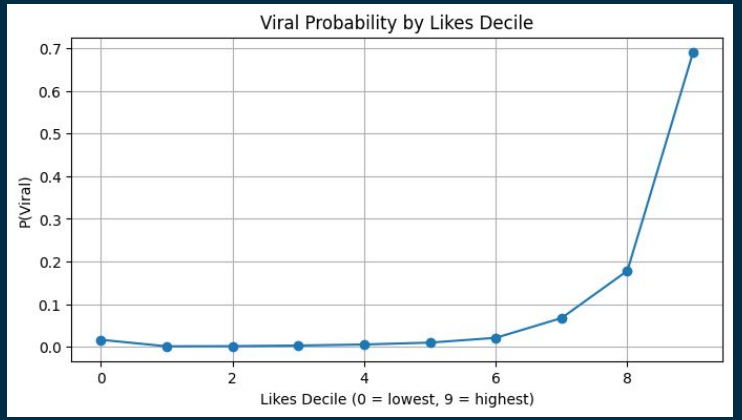
* This also directly answers “does 97% generalize?” → yes, across countries & regions.

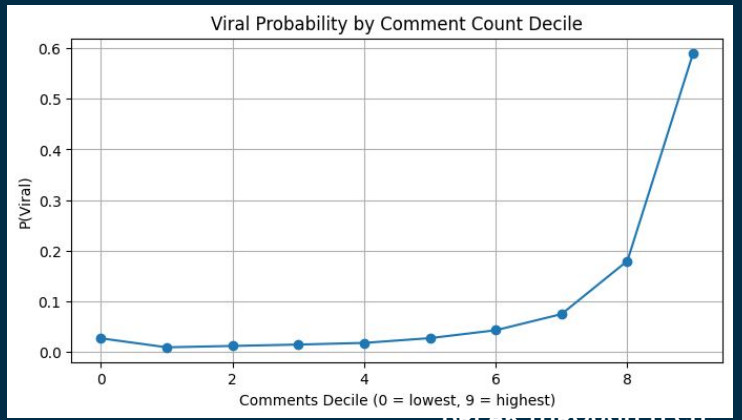
Viral Probability Jumps at High Engagement Levels

- Relationship is **highly nonlinear:**

  nothing happens for a while, then virality explodes after a threshold

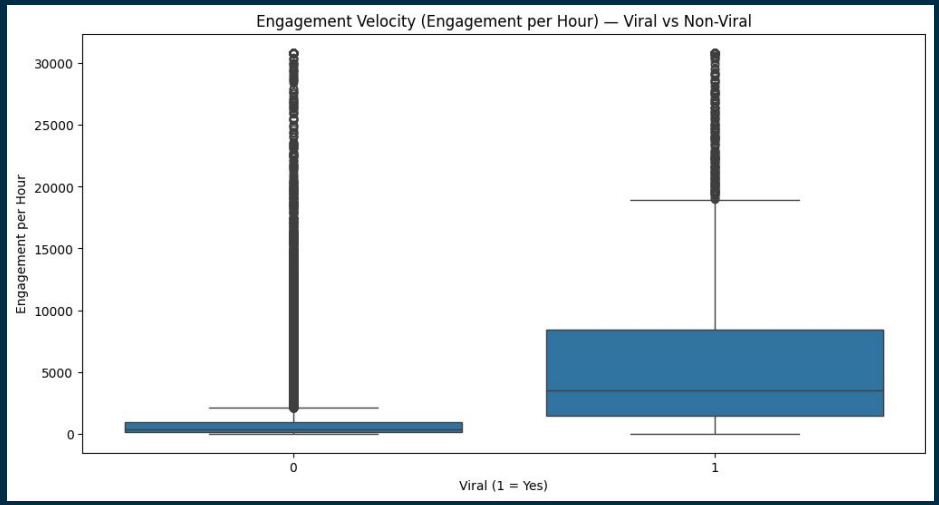
- For likes, viral probability stays near zero until ~50k+ likes, then jumps sharply
- This supports the idea of a **“viral tipping point”** in early engagement





Viral Videos Get Engagement ~6x Faster

- Viral videos get:
  - ~6x higher average likes per hour
  - ~6x higher comments per hour
  - corresponding ~6x higher engagement_per_hour
- Velocity separates viral vs non-viral *much more* than raw counts
- **How quickly** the audience reacts is crucial for Trending, not just total engagement



Boxplot (viral vs non-viral) of "engagement_per_hour"

(likes + comments per hour)

Predicting Trending Speed (Regression)

```
=== Regression Performance (Predicting Time to Trending) ===
MAE:  1.7859 days
RMSE: 2.3954 days
R²:   0.3307
```

- We reframed the task:

  “Given a video that will trend, how fast will it trend?”

- RF Regression achieves:
  - Avg error ≈ **1.8 days,** R^2 ≈ 0.33 → moderate but meaningful predictive power
- Same story: early engagement + country strongly shape trending speed
- However, there’s still lot of **randomness** and missing information (subscriber count, creator history, embeddings)

	feature	importance
0	likes	0.1720
2	comment_count	0.1476
5	title_length	0.0981
29	country_KR	0.0929
1	dislikes	0.0788
6	tag_count	0.0751
31	country_RU	0.0650
3	publish_hour	0.0624
4	publish_dayofweek	0.0430

False Positives & False Negatives

- **False Positives:**
  - High engagement but never cross viral threshold
  - Often from IN/MX/RU with huge absolute numbers but slower growth; high competition
- **False Negatives:**
  - True viral videos with *moderate* early engagement that blow up later (“slow burn”)
  - Many from BR/CA/GB and “late-trending” patterns (3-8 days)
- **Takeaway:**
  - Our model is conservative: it rarely calls something viral w/o very strong early signals
  - Next step: incorporate more detailed time-series features & creator-level data

Error Type	Count
False Positives (predicted viral but not viral)	**1,369**
False Negatives (actually viral but predicted non-viral)	**5,230**

Final Thoughts

What Can YouTube & Creators Do With This?

For YouTube / Google:

- Use global RF model for **early viral detection** across markets
- Adjust recommendation **thresholds by region** to account for different viral rates
- Monitor velocity features ("likes_per_hour", "comments_per_hour") as early risk/opportunity signals

For Creators:

- Focus on **early engagement velocity**
  - Encourage likes & comments first 24-48 hours
- Optimize metadata:
  - Clear, descriptive titles and rich tags provide modest but consistent boosts
- Recognize that virality looks similar across countries:
  - High early engagement → similar viral behavior globally, regardless of category/country

Limitations & Future Directions

- Data only includes **videos that already hit Trending,** not the entire YouTube universe
- Missing features:
  - channel subscriber counts
  - creator history
  - video content embeddings (NLP / vision)
- Computational tradeoffs:
  - Full deep NLP / transformers were out of score with ~2.9M rows.
- Future work:
  - Add engagement velocity features to main model
  - Try **per-region thresholds** or fine-tuned models
  - Incorporate multilingual text embeddings for titles/tags

Thank you!

Any Questions?