# TikEngage

## TikTok User Engagement Analysis & Prediction

*Tian Lan*

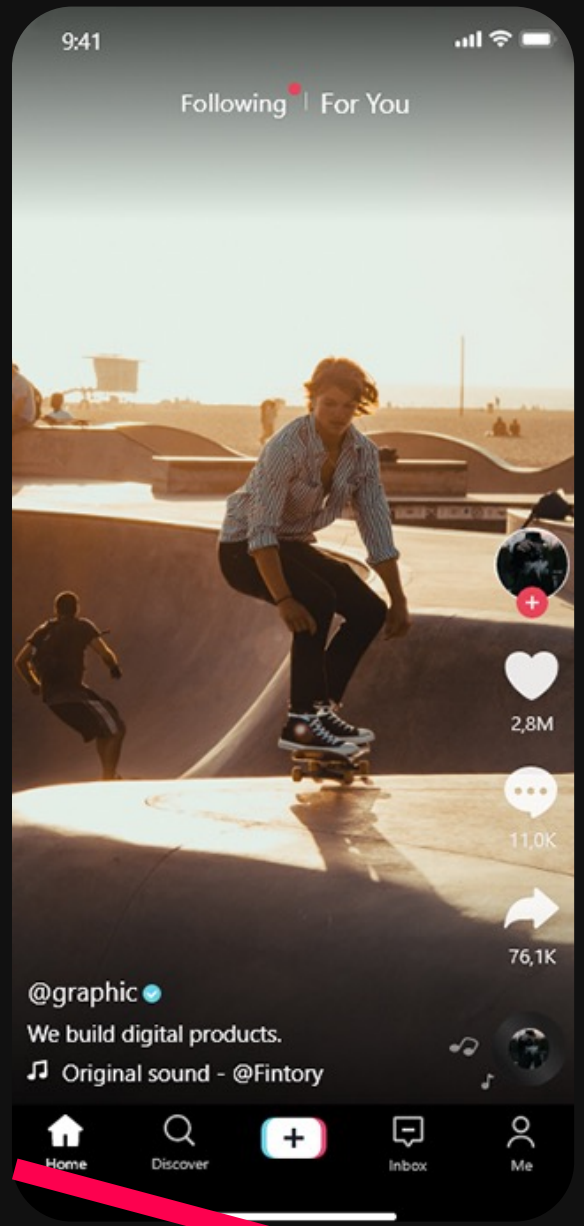
"TikTok has 1.06 Billion active users worldwide"

"83% of TikTok users have posted a video"

Content Creator

Social Media Influencer

Brand Marketing




Best way to ***go viral***?

Which of them really matter?

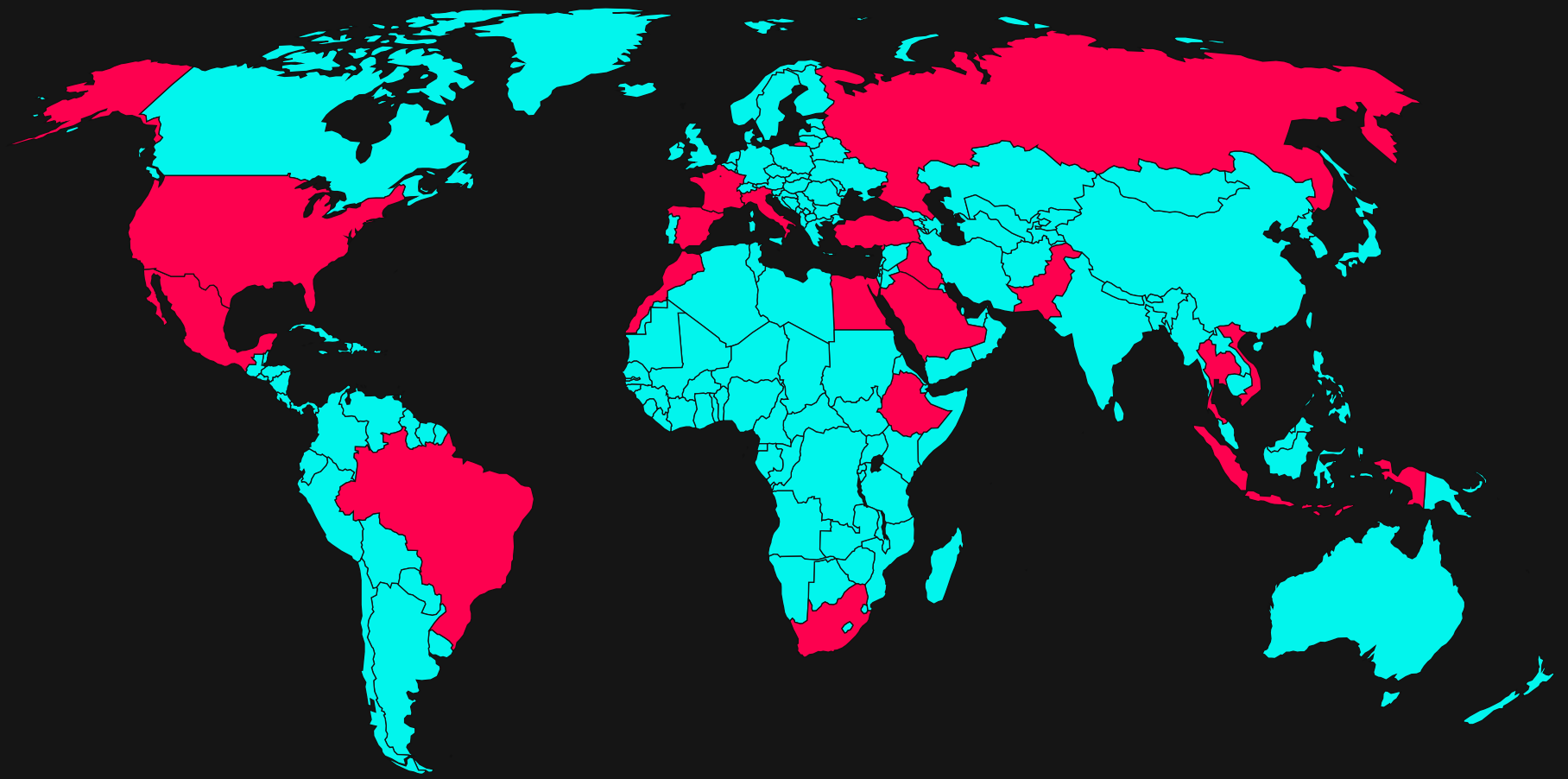
- *Hashtags*
- *Video length*
- *Upload timing*
- *Background sound*
- *Total followers*
- … …

How to evaluate?

TikTok for developers

# TikTok Research API (unofficial version) – captured on Nov.7th, 2020

- 95,000+ videos
- 1,700+ user
- 10+ countries



*https://github.com/datares/TikTok_Famous/tree/main*

# Data Description

| Column Name | Description |
|---|---|
| **ID** | Video identification number |
| **Create Time** | Unix datetime for the upload of the video to the TikTok app |
| **User** | Creator username |
| **Hashtags** | Hash keywords applied to the video description to influence TikTok algorithm |
| **Song Title** | Sound applied to the video |
| **Length** | Length of the video in seconds |
| **Likes & Shares & Comments & Views** | Number of Likes & Shares & Comments & Views the video received from other users |
| **Followers** | Number of TikTok users who follow the creator's account |
| **Total Likes** | Total likes from other users on all creator's videos |
| **Total Videos** | Total number of videos uploaded by the creator |

Each datapoint: TikTok video metadata

Features Engineering:

- Number of hashtags
- Bag of words for hashtags
- Total Engagement
- Upload year/month/weekday/period
- …

Target:

User Engagement Rate

= (Likes + Shares + Comments) / Followers

(Social Media Industry Standard)

# Data Preprocessing

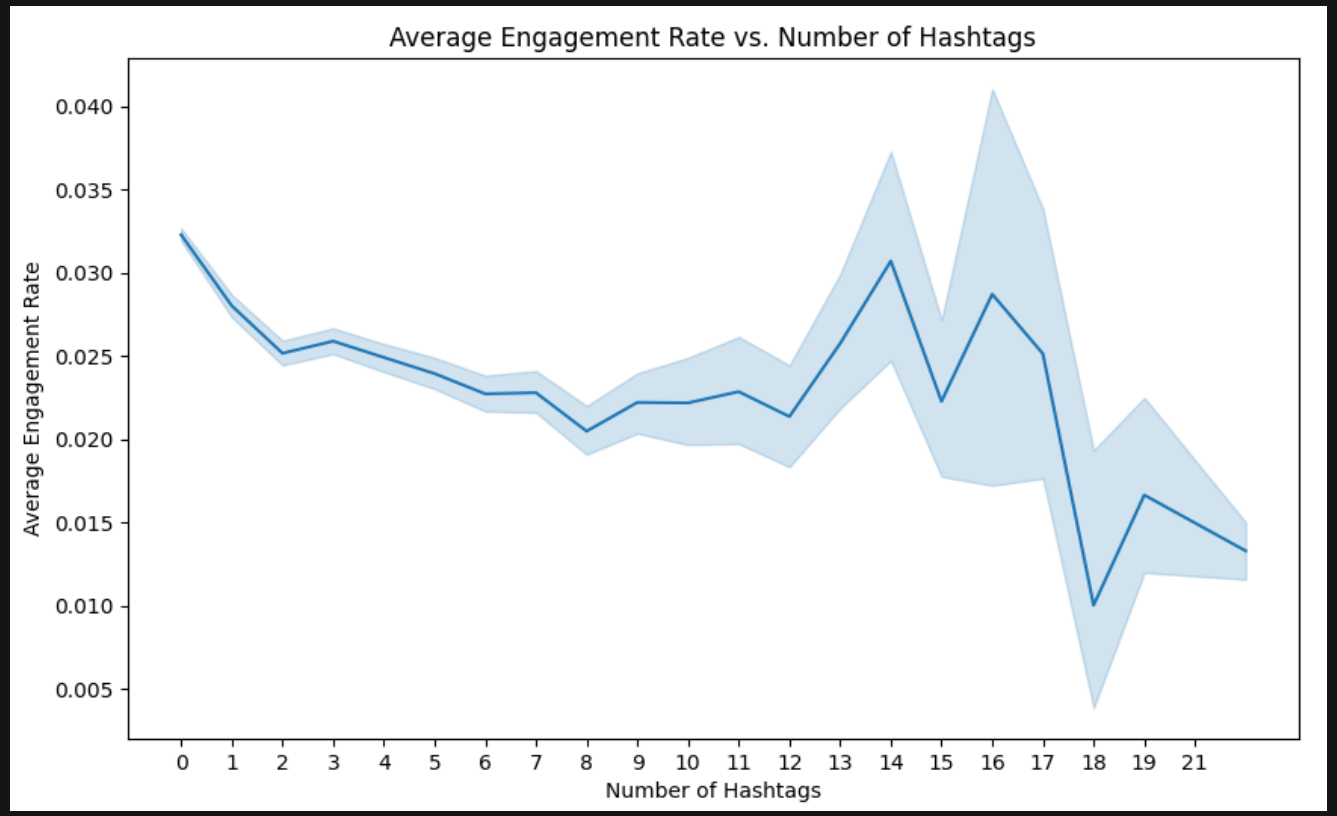
**Original** Shape: 95,963 rows X 13 columns

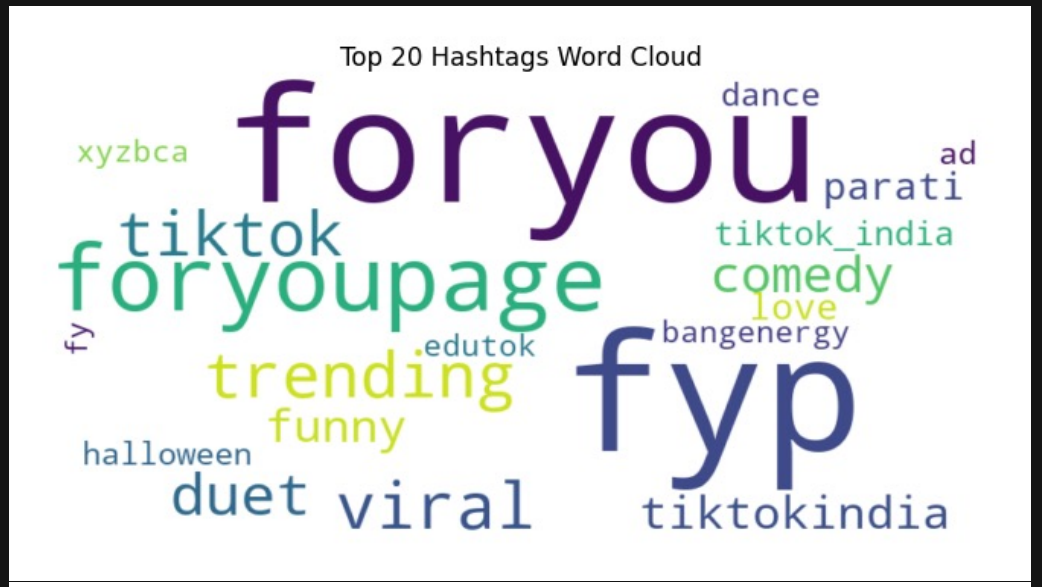
- Duplication: 43,119 rows
- Missing value: 42 NaN value in Song column
- Refine data type
- Remove outliers
- High correlated columns
- New calculated features

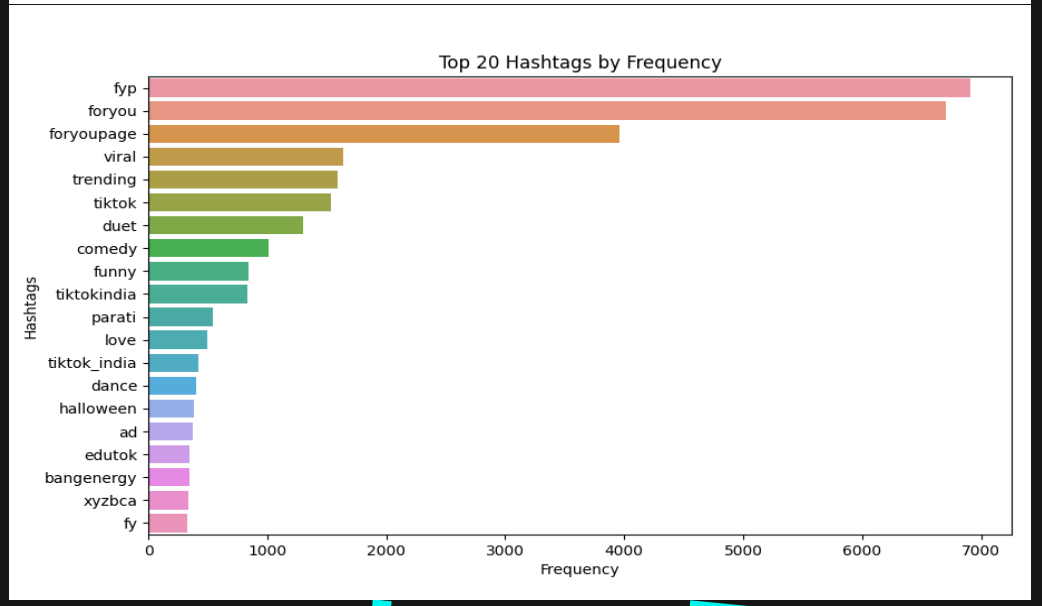
**Cleaned** dataframe: (45,804, 20)

```
Index: 45804 entries, 0 to 12558
Data columns (total 20 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Video ID             45804 non-null  int64
 1   User Name            45804 non-null  object
 2   Hashtags             45804 non-null  object
 3   Song Title           45804 non-null  object
 4   Video Length         45804 non-null  int64
 5   Likes                45804 non-null  int64
 6   Shares               45804 non-null  int64
 7   Comments             45804 non-null  int64
 8   Views                45804 non-null  int64
 9   Followers            45804 non-null  int64
 10  Total Likes          45804 non-null  int64
 11  Total Videos         45804 non-null  int64
 12  Upload Year          45804 non-null  int32
 13  Upload Month         45804 non-null  int32
 14  Upload Day           45804 non-null  int32
 15  Upload Weekday       45804 non-null  int32
 16  Upload Period        45804 non-null  object
 17  Total Engagement     45804 non-null  int64
 18  Engagement Rate      45804 non-null  float64
 19  Number of Hashtags   45804 non-null  int64
dtypes: float64(1), int32(4), int64(11), object(4)
```

# Top Hashtags

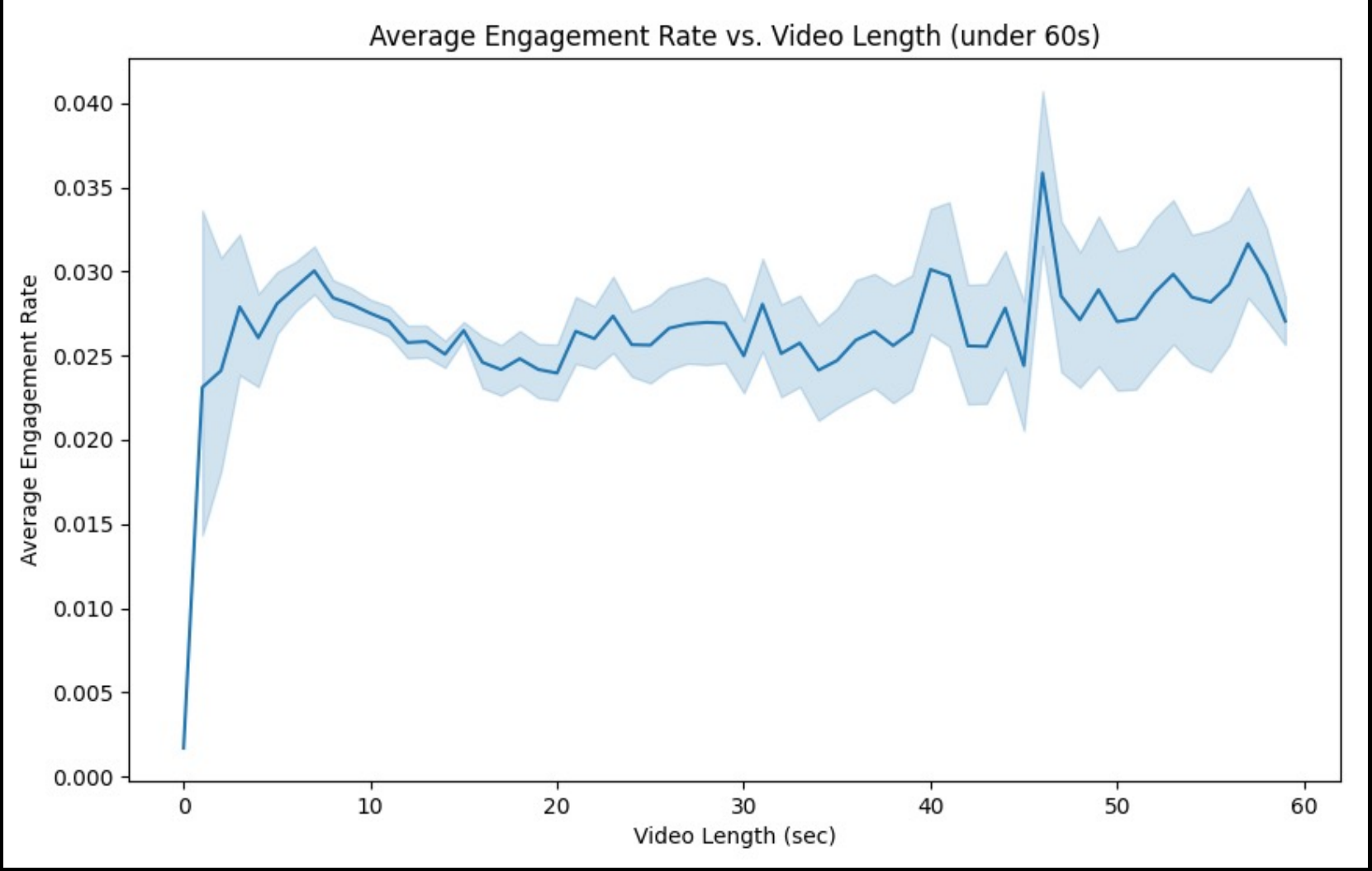



- ✓ Increase discoverability
- ✓ Collect user-generated content
- ✓ Jump on trends
- ✓ Encourage engagement

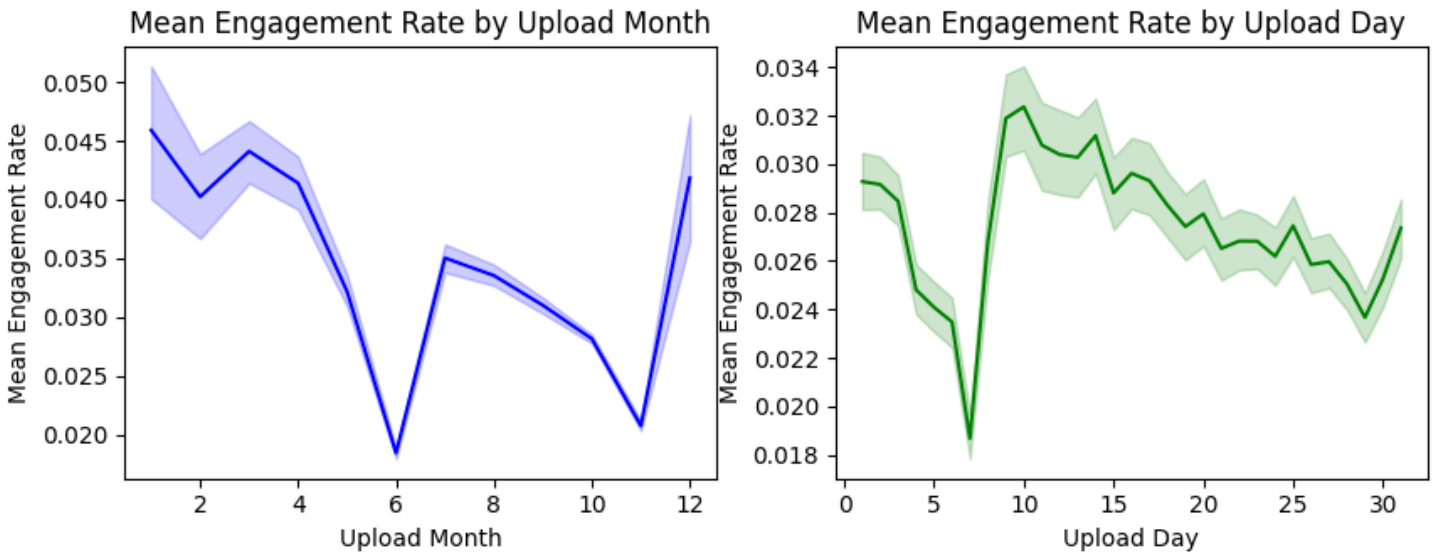
# *Video Length*




- TikTok video length limit:
  Up to 3 mins/10 mins
- TikTok recommendation:
  21 - 24 secs ?
- Best video length on average:
  10 - 15 secs ?
- … …

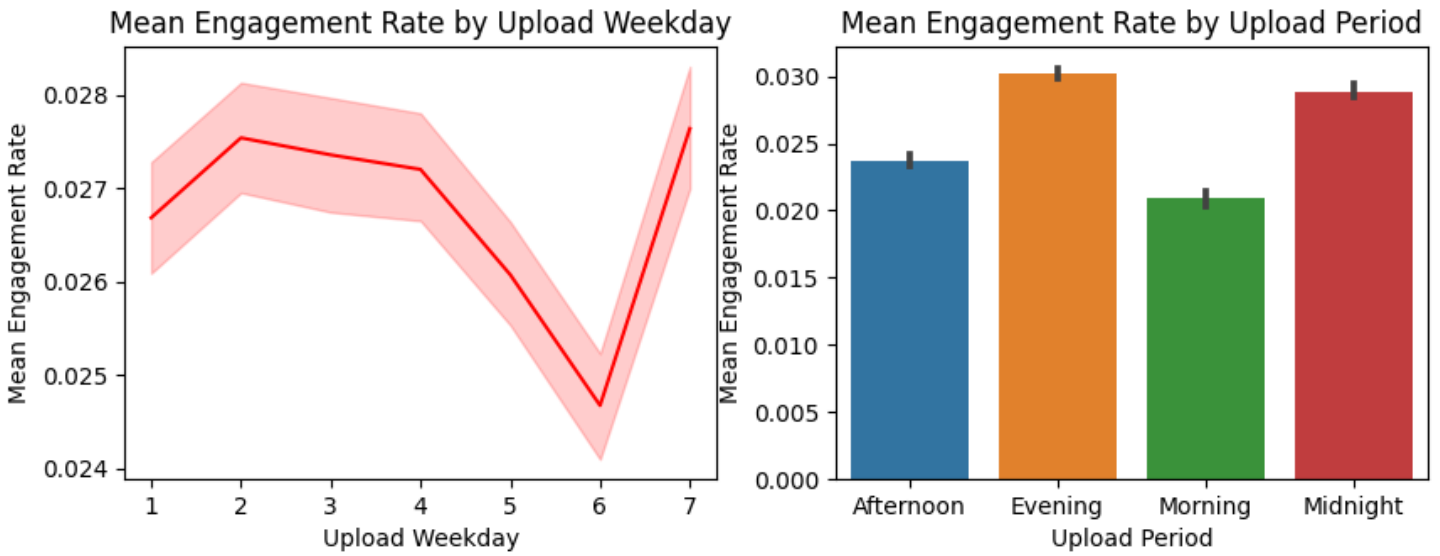
NOT significantly correlated to *Engagement Rate*

# Temporal Pattern













Highest Engagement Rate on average:

- Month - January
- Day - 10$^{th}$ ~ 15$^{th}$
- Weekday - Sunday
- Period - Evening

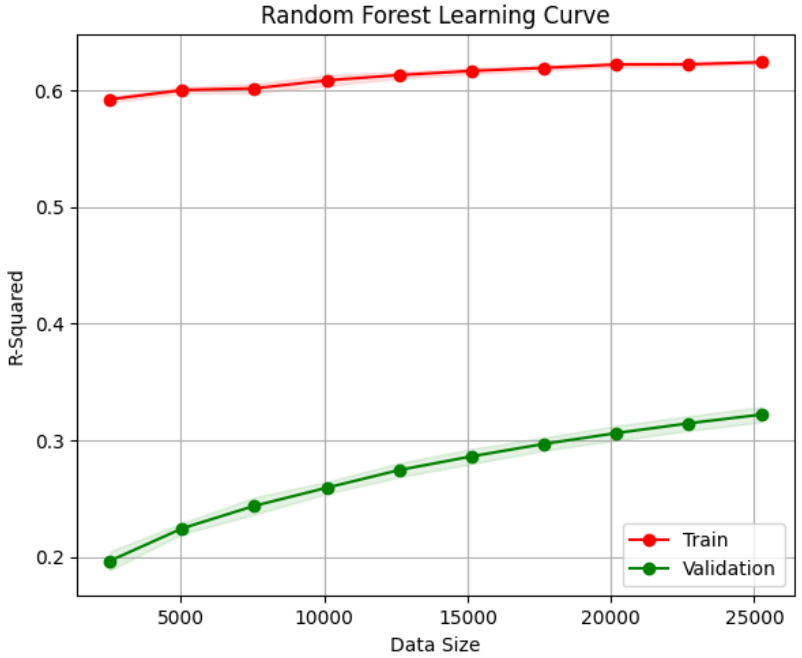
# Model Evaluation

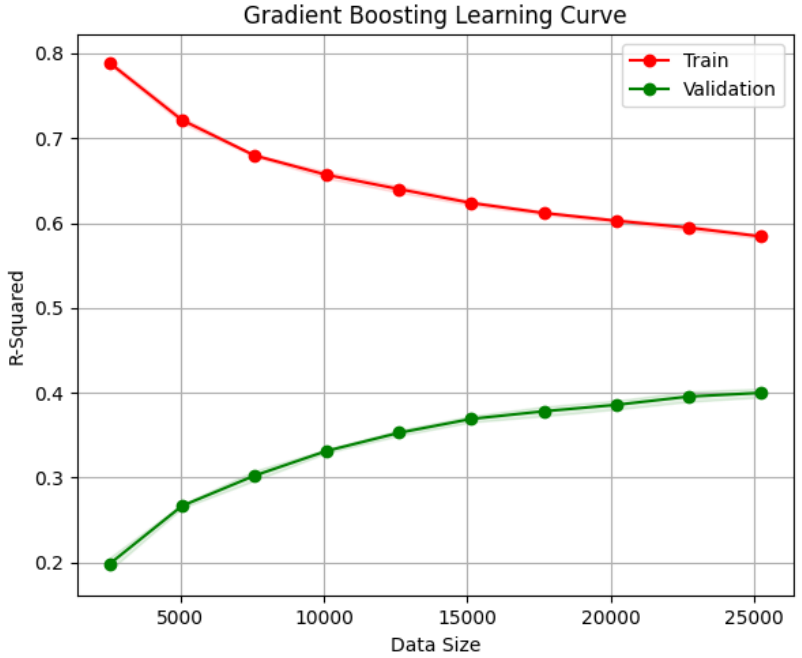
| Model | R-Squared | MSE | RMSE | MAE |
|---|---|---|---|---|
| Linear Regression | -0.0040 | | | |
| Random Forest Regressor | 0.3397 | 0.0004 | 0.0197 | 0.0149 |
| Gradient Boosting Regressor | 0.4167 | 0.0003 | 0.0185 | 0.0139 |
| XGBoost Regressor | 0.4409 | 0.0003 | 0.0182 | 0.0133 |
| LightGBM Regressor | 0.3361 | 0.0004 | 0.0198 | 0.0151 |

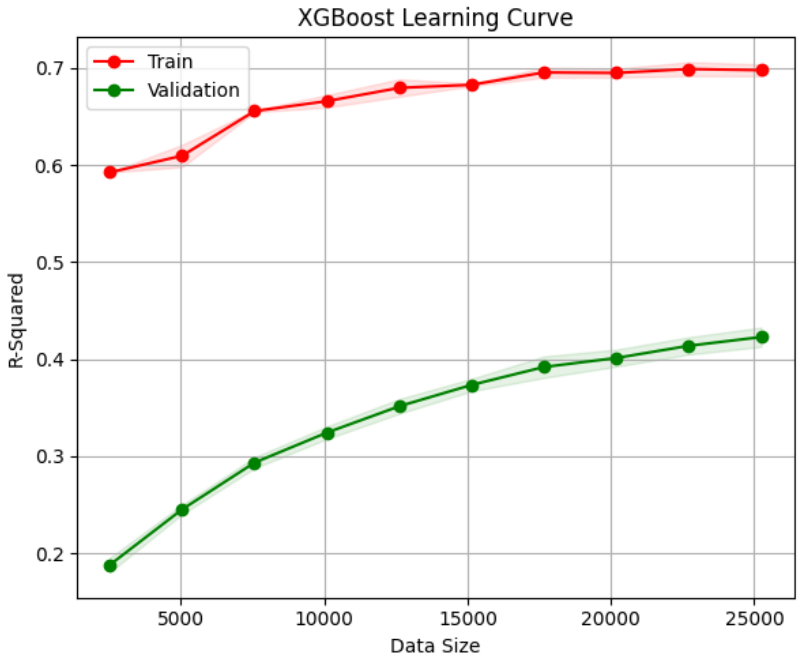
Model score trained by others:

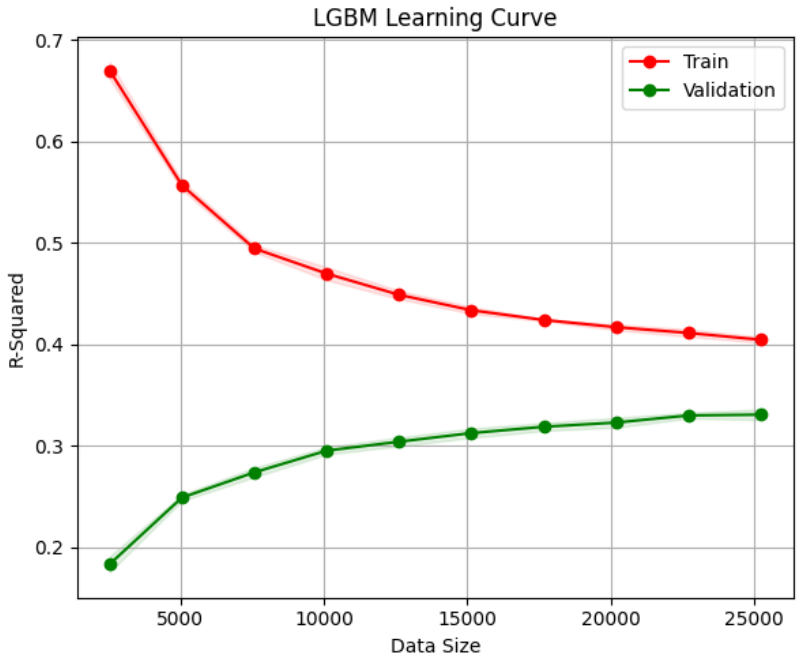
- CatBoost Regressor: -0.1874
- XGB Regressor: -0.3519
- Random Forest Regressor: 0.5667

*$R^2$: Coefficient of Determination *MSE: Mean Squared Error
*RMSE: Root Mean Squared Error *MAE: Mean Absolute Error



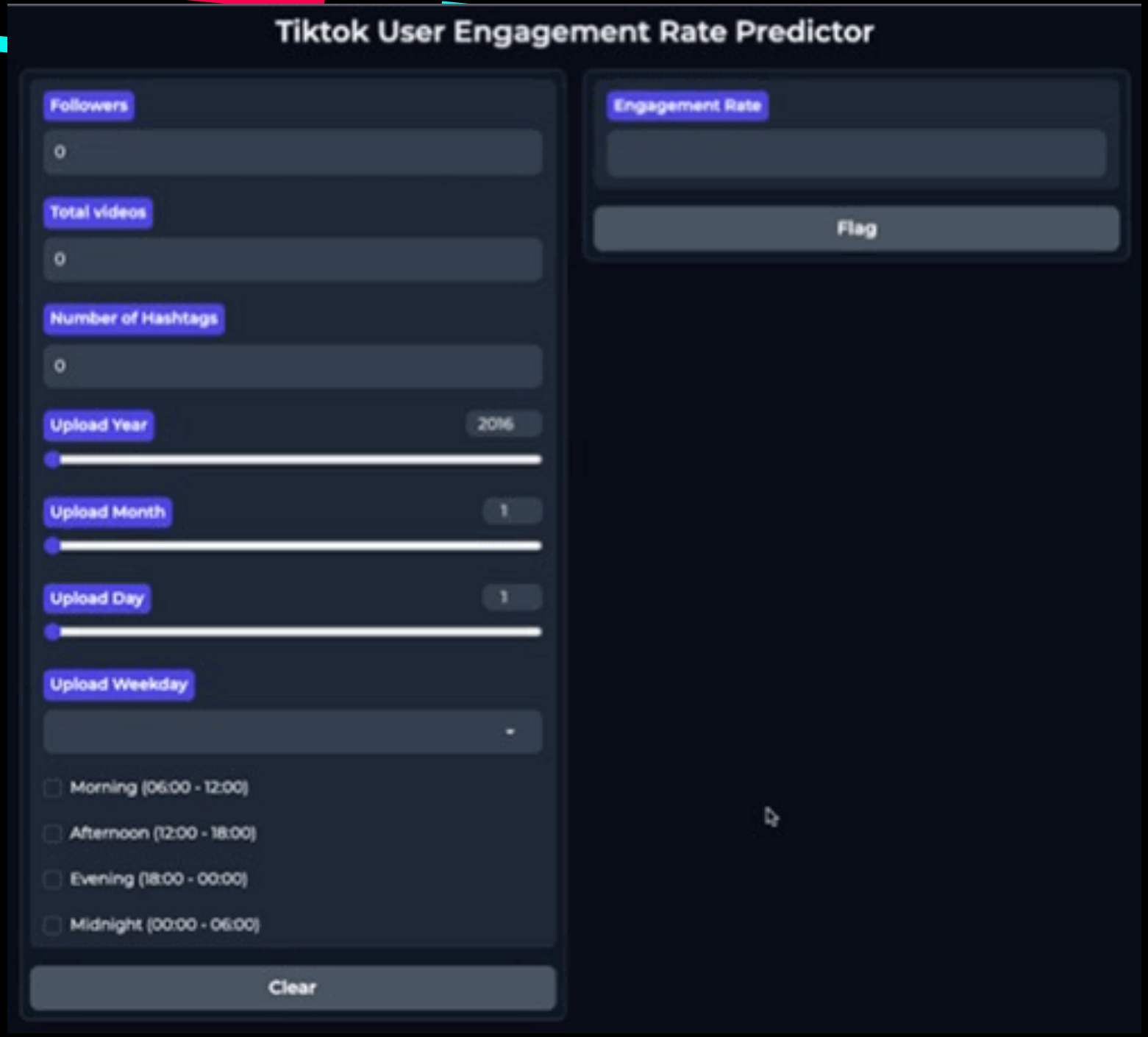




# *Model Evaluation*

## Best Model:
## Gradient Boosting Regressor

{'subsample': 0.7, 'n_estimators': 150, 'min_samples_split': 3, 'min_samples_leaf': 5, 'max_depth': 10, 'learning_rate': 0.03}

# Model Demo




Running on local URL:
http://127.0.0.1:7864
Running on public URL:
https://db758d8c1f97a91b93.gradio.live

# Next Step


- Segment analysis on hashtags content
- More features – country, category


- Real-time video data from TikTok
- Time-series models for forecasting
- Cross-platform

# Thank you!

# Appendix

Code:

- https://github.com/tianlan8/TikTok_Engagement_Rate_Analysis_Prediction

Original Data Source:

- https://github.com/datares/TikTok_Famous/tree/main/Analysis/TikTok%20Videos

Reference:


- <Machine Learning Yearning> by Andrew Ng
- https://www.kaggle.com/code/julienjta/tiktok-popularity-prediction-stackblend-etc/notebook#2.-Prediction-of-the-popularity
- https://www.kaggle.com/code/antoniosabatini/tiktok-popularity-track-eda-ml-models/notebook#notebook-container
- https://www.kaggle.com/code/dataranch/tiktok-popular-songs-feature-importance/notebook#**Overall-Feature-Importance**
- https://github.com/datares/TikTok_Famous/tree/main/Analysis/TikTok%20Videos
- https://www.kaggle.com/code/vbradculbertson/tiktok-engagement-forecasting-and-analysis/notebook#Random-Forest
- https://www.kaggle.com/code/erikvdven/tiktok-some-python-magic-in-a-notebook/notebook#Some-first-Analysis-%F0%9F%93%88
- https://www.researchgate.net/figure/Learning-curves-can-be-used-to-examine-the-behavior-of-a-neural-network-model-during_fig9_349898718
- https://theinfluencermarketingfactory.com/how-to-calculate-tiktok-engagement-rate/