Trend Engagement on TikTok: Regression Forecasting

V. Brad Culbertson
Data Scientist

Virality has been the subject of great interest over the past two decades. From YouTube and Vine to Reels and TikTok, created and curated video content has made everyday people and relatively unknown performers into global sensations and made millions of dollars for creators and host platforms alike. The purpose of this analysis is to generate a forecasting model for trend-driven video engagement using data science techniques. This work was performed in the Python coding language using Jupyter Notebook. All relevant datafiles are included.

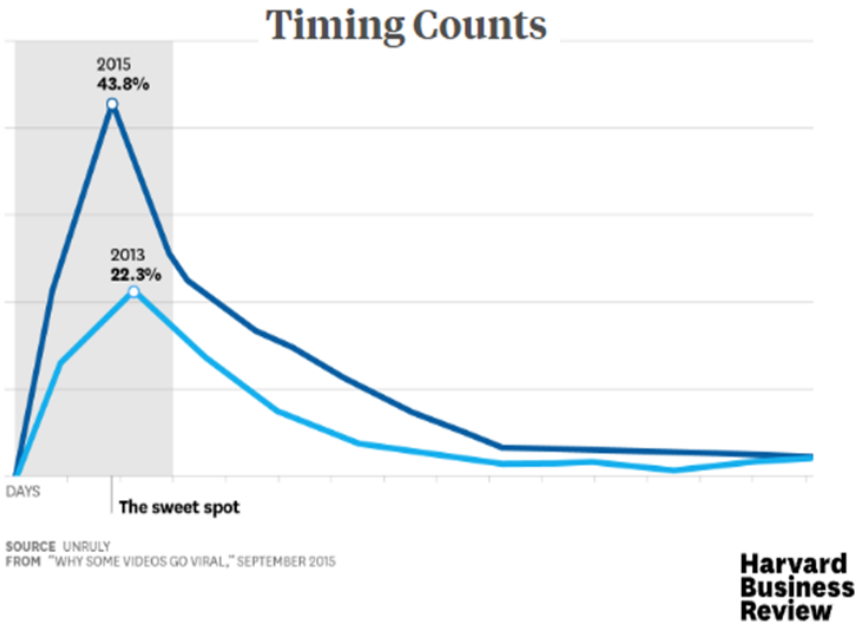
Business Value

Current marketing strategies on social media content platforms continue to be professionally produced commercials. Production of these short-form videos takes time from conceptualization to final cut. Production time depends on the scale of the project. It can take anywhere from six days to six months. (Braun Film and Video) In the world of social media, even six days can miss the life cycle of a trend. Metric Marketing suggests it can take months to begin seeing the results of digital marketing efforts. With real-time data availability, This kind of delay may be unnecessary. Digital marketing is measured most frequently by CTR (click-through rate), which counts the number of users who click through the advertisement to reach the content page. In social media, 14% is considered a good CTR and 16% is high (Tik Tok). These results are achieved with the standard produced video content. Rarely do they align with viral trends such as trending song/sounds or online challenges.

The virality of trending video content by its nature yields much higher-than-average engagement rates. Produced media campaigns miss these opportunities due to the slowness of campaign creation and production. Forecasting trend engagement may lead to new marketing strategies which focus on virality and near-immediate production time to leverage the engagement of viral trends.

Scope & Data

Acquiring data for this analysis required finding social media metadata filtered by trending engagement. The original dataset was collected by Ivan Tran in a GitHub repository titled “How to be TikTok Famous”. The data was scraped using the TikTok API (which has since been updated and is now much more difficult to obtain credentials). The data consisted of six .csv files containing the features described later in the Data Dictionary.



It was important to find data pulled from the Trending API because of its significance to the problem statement. This importance is underscored by a 2015 analysis by Unruly (reposed by Harvard Business Review). Virality is short-lived, and data from trends would need to be narrowly focused.

Exploratory Data Analysis

The original concatenated dataframe consisted of 95,963 rows and 13 columns. There was a large number of duplicated records likely resulting from the way TikTok API calls the data and when the data was scraped from the service initially. Fortunately, the ID column is a unique video ID number and could easily be leveraged to remove true duplicates. No duplicate columns were noted. Three columns were object data types: 'user_name', 'hashtags', and 'song'. Each record in the 'hashtags' column was a list of strings. All other columns were integers. The 'create_time' column was unusual and required research. The integers were identified as a Unix timecode. The target variable was included in parts spread across four different columns: 'n_likes', 'n_shares', 'n_comments', and 'n_plays'. These were combined in 'Engagement'.

A correlation matrix was run on the variables (sans objects) with the following findings:

- Positive correlations between Views and Likes/Shares/Comments.
- Weak positive correlation between Followers and Views.
- Weak positive correlation between Followers and Likes/Shares/Comments.

Data Munging (Cleaning)

Column names were standardized into a more readable format. 'Create Time' was converted from Unix to standard datetime format. The '@' handle identifier was removed from User names. Engagement Rate and Engagement columns were created and calculated appropriately. Exploring the songs/sounds used in the videos further limited the dataset for modeling. Nearly half the videos in the dataset used *original sound*, which means the creator's own voice was used in the video. This is unhelpful when exploring song/sound virality across videos of different users. These were accordingly removed from the modeling dataset. The number of uses for each song/sound were calculated as were the min and max dates for the use of each song/sound. The number of days the song/sound was in use was calculated as well as the relative window position (where in the window of activity was this video placed).

The Window mean was 115 days of activity. The mean window placement was .87, which implies that most of the videos using a given song/sound were created near the end of the window of activity. After removing *original sound* videos, this mean shifted to .59. This implies a more even distribution of videos created over the time of activity. The dataset was also limited to Engagement values of 5 million and below to limit the influence of a small number of significant outliers (200 M + Engagement).

Preprocessing

With the identified target variable being continuous Engagement, regression models were selected for predicting and forecasting. Sci-Kit Learn's Multi-Label Binarizer was used to one-hot encode song titles and hashtags. Separate X_sets were created from the parent dataframe. Train, validation, and test splitting was performed. A standard scaler was applied to X train and X validation data.

Secondary preprocessing was performed after initial modeling to evaluate an hypothesis. Four X_sets were ultimately created from the parent data. Set one included binarized song titles and hashtags. Set two only binarized song titles, dropping hashtags from the set. Set three binarized only hashtags, dropping song titles. Set four dropped both song titles and hashtags. Each was evaluated in an iteration of the complete modeling battery.

Refreshing our view of the problem statement, predicting Engagement would be done without the knowledge of how many users viewed, liked, shared, or commented on a video prior to its posting. To eliminate the inclusion of "crystal ball" features, they were removed from all datasets.

Modeling

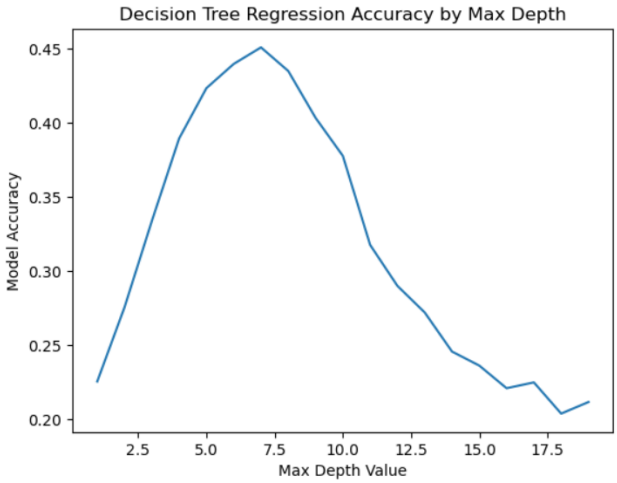
Linear regression was initially performed (without binarized variables) with R2 values of .267 on training data and 251 on validation data. Beta coefficients from initial modeling are shown here. Indications are that Followers and Total Likes have strong influence over Engagement. Length of video also has a positive effect. Total Videos (higher numbers implying spam posting) has a strong negative effect on Engagement. The longer a song has been in use (Days Since Debut) also has a strong negative effect.

A Decision Tree Regressor was optimized over max depth values of 1 to 20. Accuracy was plotted and optimum max depth was identified at 7 with an R2 value of .451. This marked significant improvement in forecast accuracy over the linear regression model.

A Random Forest Regressor was applied next to attempt greater precision. Random Forest Regression yielded at maximum R2 value of .57. Another significant improvement over the Decision Tree Regressor.

Finally simple neural networks were applied with significant hyperparameter tweaking with hidden layer sizes (30, 20, 10, 10, 10, 5) and max iterations of 1000 required to achieve just .390 R2 score.

TensorFlow Neural Network was then applied with similar results, despite 9X40 (layers x nodes). This model complexity might typically lead to overfitting, however, the model may be exposing the limits of this dataset and features.



Results

Initial modeling (and its varied iterations) yielded a question of significance relating to some of the features. Would the inclusion or exclusion of binarized variables (songs and hashtags) exert influence on model performance/accuracy?

An iterator was created to run each of the four variations of the dataset through each of the models and record the R2 values. The results were captured in the dataframe shows on the right.

Despite initial observations during model iteration and optimization efforts, the best model performance as measured by R2 score, was achieved by including binarized hashtags and applying the Random Forest Regressor (~.57).

Compared to a naïve prediction in a range of 0-5000000, the model performs significantly better.

Findings

Engagement is difficult to predict. A small number of high-engagement creators exert significant influence over forecasting. Tightening the range of analyzed engagement greatly increases model accuracy. There was a surprisingly low correlation between a creator's number of followers and the resulting total engagement. Total number of videos created has a negative impact on video engagement, suggesting a quality over quantity approach to be more effective.

Be Aware

Trend lifecycles are short-lived. Significantly fewer song/sounds appear in-use longer than 100 days. Most of the engagement for those is held in the first 14 to 26 days. Nearly all trends terminate within 200 days.

With an average production time of 4 to 8 weeks, even fast turn-around from conceptualization to release misses the peak of trending engagement windows. Trends must be actively monitored to be leveraged within the peak engagement window.