TikTok Share Prediction Through Machine Learning (Logistic Regression)

Kevin Conaty, Justin Derenthal, Dr. Nur Dean
Farmingdale State College

FARMINGDALE STATE COLLEGE IGNITE SYMPOSIUM

Introduction

Social media platforms offer both advantages and disadvantages to users. While they facilitate connection and information sharing, they also present challenges such as misinformation and privacy concerns. This study concentrates on the positive aspects of social media and seeks to understand the factors contributing to user engagement.

In particular, we investigate the phenomenon of user obsession and the desire for attention within social media communities. User engagement metrics, including likes, comments, and shares, serve as indicators of this obsession. Our research aims to analyze and identify the key features influencing video share counts, thereby enabling the prediction of the number of times a video will be shared.

Tik Tok

Dataset and Features

Metadata collected and used for this study was gathered from a data scraper used in a previous study to gather user engagement from 1200 TikTok videos. The data initially included over 19,000 videos from both verified and unverified accounts, it was narrowed down to only verified due to the higher impact verified accounts have on the platform.

Video Duration	How long the video is in seconds
Video Views	The number of times a video was viewed
Video Likes	The number of likes a video receives
Video Downloads	The number of times a video was saved to a device
Video Comments	Number of comments left on a given video
Video Shares	The number of times a video was shared to other users

Methods

Variable X: The variable "X" was used to hold the data found within the video: duration, view, like, download, and comment count.

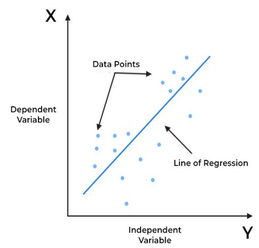
Variable Y: The variable "Y" was used to hold the data found within the manipulated "Binary Classifications" column.

Binary Classification: The mean value [6951] of shares was used to create a binary classification. This is a method of representing a question for the models to execute. In this example, the binary classification was defined as 1 [true] if the value of a rows shares are above the mean, and 0 [false] for values less than the mean.

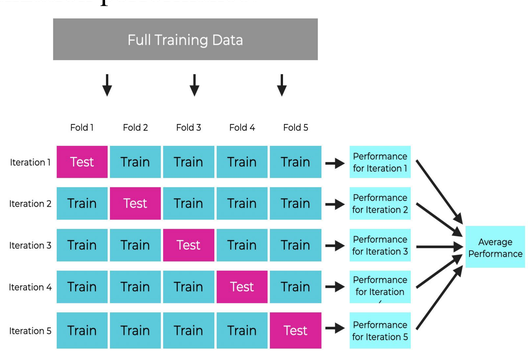
Sklearn: This library was implemented to split the training and testing sets for the models use. 80% of the dataset was allocated towards the training set, while 20% of the dataset was allocated towards the testing set.

Standardization: StandardScaler was imported from Sklearn to allow the mean to be represented as 0, and find/implement the standard deviation, as this helps reduce overfitting

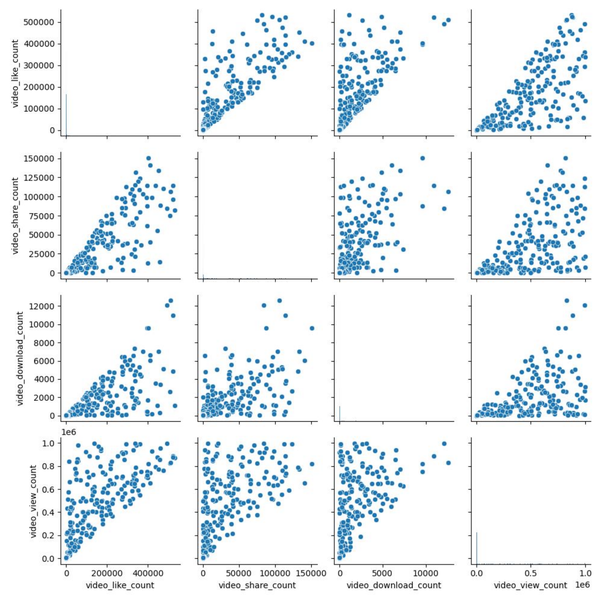
Logistic Regression: Used to represent the relationship between the dependent variables [duration, like, downloads, views, and comments] and independent variable [video_share_count] better known as the "Binary classification".



K-fold Cross Validation: Utilized to divide the dataset into k subsets or folds. This model was evaluated 5 times, with a different fold each time. The metrics from each fold are then averaged to estimate the models generalization performance.



Results



Model	Accuracy Score
Logistic Regression Model without Binary Classification	0.04%
Logistic Regression with Binary Classification	96%
Logistic Regression with Binary Classification and Standardization	87.92%

Future Research

This is a preliminary study on building machine learning models using TikTok data. Eventually our goal is to create our own scraper that we can use and personalize to fetch any data from TikTok that we see as valuable to our research and further tests. Tests such as creating more binary classification groups for different share values. This will enable us to enhance our accuracy, allowing us to find the optimal engagement to increase the number of shares a video receives and consequently boost its profitability.

References

Carta, S., Podda, A.S., Recupero, D.R., Saia, R. and Usai, G., 2020. Popularity prediction of instagram posts. *Information*, *11*(9), p.453.

Yakhyojon. "TikTok User Engagement Data." *Kaggle*, 18 Oct. 2023, www.kaggle.com/datasets/yakhyojon/tiktok.