Executive Summary: Regression Analysis

Predicting User Verification on TikTok Using Logistic Regression

OVERVIEW

The TikTok data team aims to build a machine learning model to support the classification of claims in user-generated video content. Previous analysis revealed that verified users are more likely to post opinions rather than factual claims. As a result, understanding the characteristics of verified accounts became a critical step. In this phase of the project, a logistic regression model was developed to predict the verification status of a user based on features extracted from their videos. This model provides foundational insight to help improve the efficiency of claim classification downstream.

PROJECT STATUS

The verified_status variable was selected as the target for this logistic regression model based on prior observations that verified accounts are more likely to post opinions. Logistic regression was chosen due to the binary nature of the outcome variable and the interpretability of the model coefficients.The model was trained and evaluated on a cleaned and balanced dataset. It achieved a precision of 61%, recall of 84%, and an accuracy of 65%. These results are considered acceptable and useful for supporting downstream tasks such as identifying opinion-based content. The model's strongest performance lies in its ability to correctly identify verified accounts, which is aligned with the project's overall goal.

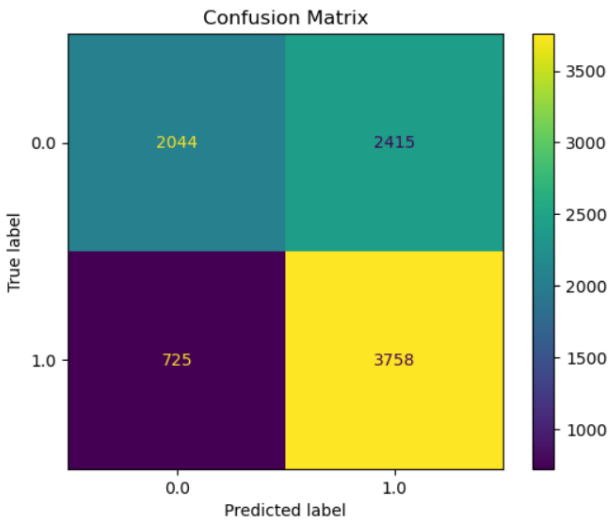
NEXT STEPS

The next phase of the project will focus on developing a classification model to predict the claim status of videos, determining whether a given submission represents a factual claim or a personal opinion. This is the ultimate goal outlined by the TikTok data team. Now that we have a better understanding of the characteristics of verified users, who are more likely to post opinions, we can integrate this contextual information into the final model. The verified_status model provides valuable groundwork for enhancing the performance of the claim classification task by incorporating behavioral insights about the type of users generating content.

KEY INSIGHTS

Based on the estimated model coefficients from the logistic regression, video duration is the feature most strongly associated with verified status. Longer videos tend to slightly increase the odds of a user being verified. Other features, such as video view count, download count, comment count, and author ban status, had small coefficients close to zero. This suggests that these features have a weak or negligible association with whether a user is verified. As a result, video duration appears to be the only meaningful predictor of verified status among the video-level features included in this model.

Confusion matrix for logistic regression model



Upper-left (2044): *videos correctly classified as posted by unverified accounts.****Upper-right (2415):*** *videos posted by unverified accounts but incorrectly predicted as verified.****Lower-left (725):*** *videos posted by verified accounts but incorrectly predicted as unverified.****Lower-right (3758):*** *videos correctly classified as posted by verified accounts.*