Executive Summary: Regression Analysis

TikTok claims classification project

OVERVIEW

The TikTok data team is working on a machine learning model to better categorize user-submitted content as either claims or opinions. Initially, they noticed that verified users tend to share their opinions more frequently. Given the goal is to predict claims and opinions, it's crucial to build a model that can accurately forecast the behavior of verified accounts, which are more likely to post opinions. To achieve this, the data team created a logistic regression model that identifies whether an account is verified or not.

PROJECT STATUS

The "verified_status" variable was chosen for the regression model because it's closely linked to the type of content found in YouTube videos. A logistic regression model was used because the data is binary (0s and 1s) and categorical.

The results of the model show that it was 67% accurate in identifying correct predictions and 65% accurate in recalling correct predictions. The model's overall f1 score is 63%. These results reveal valuable insights about the characteristics of the video content, which are discussed in the "Key Insights" section.

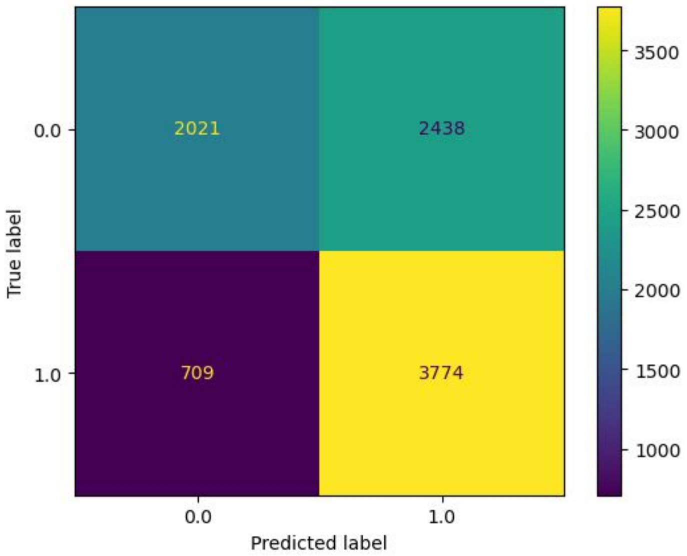
NEXT STEPS

The next step is to develop a classification model that accurately predicts the status of user claims. This is the original goal set by the TikTok team and the ultimate solution to the problem at hand. With the rich user behavior data collected, we're now in a position to evaluate the performance of this model and gain valuable insights into user behavior patterns, enabling us to fine-tune the model and improve its accuracy.

KEY INSIGHTS

According to the estimated coefficients from the logistic regression model, there is a tendency for longer videos to be linked with higher probabilities of the user being verified. In contrast, other video characteristics exhibit small estimated coefficients, indicating that their connection with verified status appears minimal. Consequently, aside from video length, other video features do not seem to be related to verified status.

Confusion matrix for logistic regression model



Upper-left: the number of videos posted by unverified accounts accurately classified as so. Upper-right: the number of videos posted by unverified accounts that the model misclassified. Lower-left: the number of videos posted by verified accounts that the model misclassified Lower-right: the number of videos posted by verified accounts accurately classified as so.