Video Status Project | Regression Modeling Results

Prepared For: TikTok Leadership Team

OVERVIEW

The TikTok data team is currently developing a machine learning model to classify videos submitted to the platform as containing either a claim or an opinion.

For milestone #5, a binomial logistic regression model is being developed to identify the variables most correlated with "verified_status," which was found to be a significant indicator of whether a video contained a claim or an opinion.

PROJECT STATUS

The "verified_status" variable was chosen because of its observed relationship with video content type. A binomial logistic regression model was determined to be ideal due to the distribution of the data and the binary nature of the outcome variable, "verified_status."

MODEL RESULTS

The model achieved a weighted accuracy of 64% (the ratio of correctly predicted observations to the total number of observations). It also achieved a precision of 60%, a recall of 85%, and an F1 score of 70%.
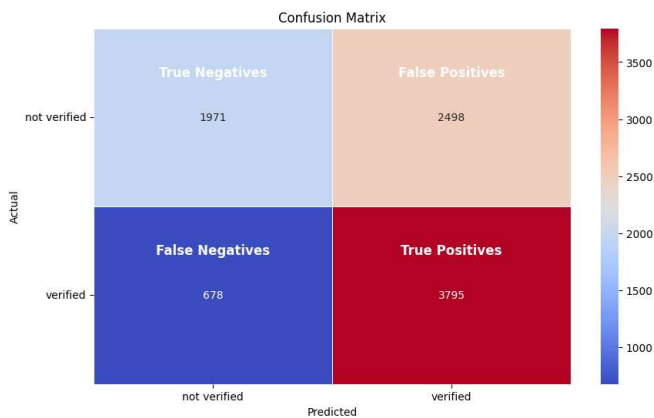
NEXT STEPS

The next step will be to create a predictive model using the relationships identified between the features and the outcome variable, as well as the results from the logistic regression model. This classification model aims to predict whether a video contains a claim or an opinion, which is the ultimate goal of the project.

KEY INSIGHTS

Based on the model's coefficients, longer videos are more likely to be from verified users.

Due to their relatively small coefficients, the other predictor variables do not appear to have a strong correlation with verified status.

Confusion Matrix

OUTCOME AND NEXT STEPS

The model demonstrated acceptable predictive capability, with a recall of 85% (the proportion of actual "not verified" users correctly identified) and a precision of 60% (the proportion of predicted "not verified" users who are actually not verified).

However, there is likely room for further improvement, and another iteration with newly engineered variables or feature selection may help achieve this.