XGBoost Predictive Model for TikTok's Claim Classification: EDA, Hypothesis Testing, Logistic Regression, Tree-Based Models

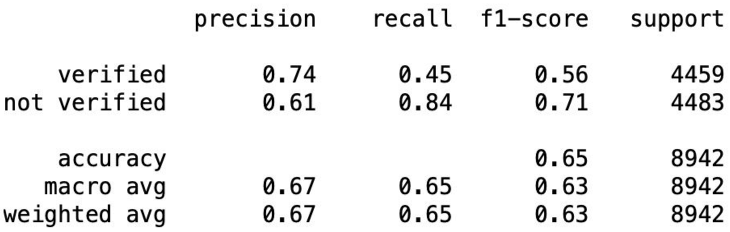
Executive Summary for Regression Analysis

OVERVIEW

The data team aims to develop a machine learning model to assist in the classification of claims for user submissions. Previously, the data team showed that if a user is verified, they are much more likely to post opinions. Hence, ultimate purpose is to classifying claims and opinions, it's crucial to build a model that demonstrates how to predict the behavior of the account type (verified) that likely to post opinions. Thus, a logistic regression model that predicts verified_status was built.

PROJECT STATUS

The 'verified_status' variable was analyzed in this regression model since the relationship between account type and the video content was indicated previously. A logistic regression model was selected because of the data type and distribution.



```
              precision    recall  f1-score   support

    verified       0.74      0.45      0.56      4459
not verified       0.61      0.84      0.71      4483

    accuracy                           0.65      8942
   macro avg       0.67      0.65      0.63      8942
weighted avg       0.67      0.65      0.63      8942
```

The logistic regression model yielded precision of 67% and a recall of 65% (weighted averages). This model had F1 score of 63%.

The logistic regression model had decent predictive power.

NEXT STEPS

Constructing a classification model that will predict the claim status made by users is next. That is the original expectation from the TikTok project.

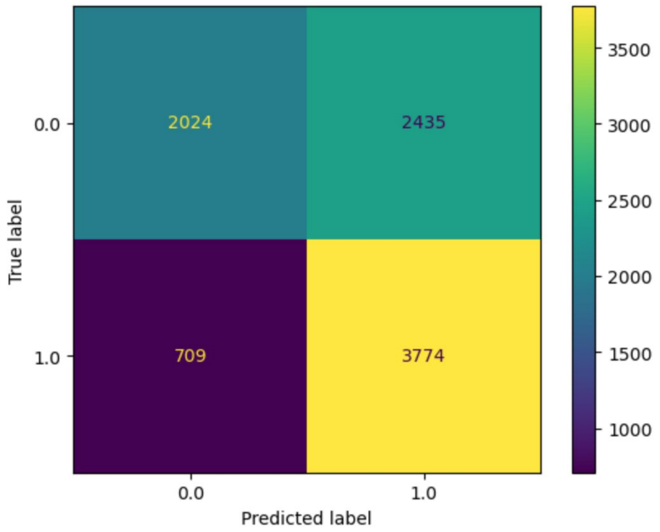
KEY INSIGHTS

Based on the estimated model coefficients from the logistic regression, longer videos tend to be associated with higher odds of the user being verified.

Other video features have small estimated coefficients in the model, so their association with verified status seems to be small.

All in all, other video features do not seem to be associated with verified status except video length.

Confusion Matrix for Logistic Regression Model



0: Not verified account & 1: Verified account.
Upper-left: true negatives. Upper-right: false positives.
Lower-left: false negatives. Lower-right: true positives.