Executive Summary: Regression Analysis

Milestone 5

Overview

The objective of this milestone was to build and evaluate a logistic regression model to predict whether a TikTok user is verified based on video characteristics and account attributes. This analysis aims to uncover which factors are most associated with verified status, support TikTok's broader claims classification project, and provide insights that will guide the development of a final predictive model distinguishing claims from opinions.

Data Preparation

Initial data exploration showed that the dataset was highly imbalanced, with 93.7% of posts from unverified accounts and 6.3% from verified accounts.
To address the class imbalance, we applied resampling techniques, resulting in a balanced dataset with equal representation (50% verified, 50% unverified).
Outliers were identified in **video_like_count** and **video_comment_count**. Using the IQR method, we capped extreme values to minimize their influence on the model.

We developed a **logistic regression model** to predict `verified_status` using video engagement metrics. This model reveals which video characteristics most influence the likelihood of verification and serves as a diagnostic tool in the larger claims detection framework.

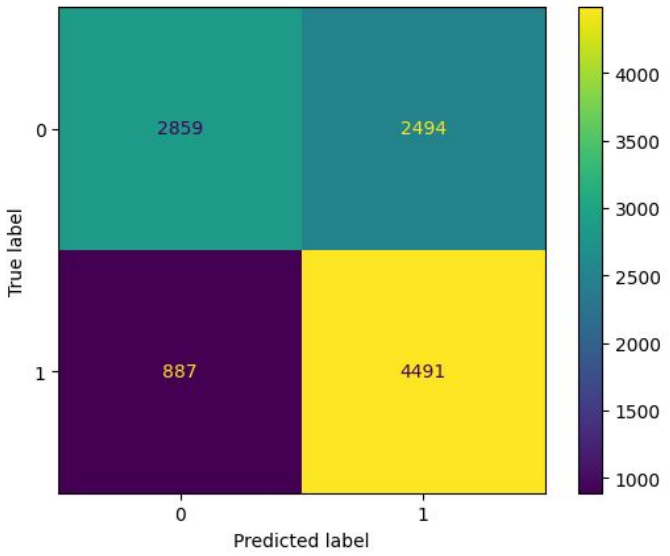
Details

Features Used: `video_duration_sec`, `video_view_count`, `transcription_length`, `claim_status`, `author_ban_status`

Key Correlations: Strong correlation observed between engagement metrics (e.g., likes, shares, downloads)

Top Predictors (Log-Odds Coefficients):

- `claim_status_opinion`: **+1.47506**
- `author_ban_status_banned`: **–0.43319**
- `author_ban_status_under_review`: **–0.37031**
- `transcription_length`: **–0.00192**
- `video_duration_sec`: **–0.00174**



Model Performance:

- **Accuracy:** 68.5%
- **Precision:** 64.3%
- **Recall:** 83.5%
- **F1 Score:** 72.7%

Next Steps

The next step is to build the final classification model to predict claim status. With insights from the regression analysis, we are now better equipped to interpret user behavior and improve model performance.