Claim Classification Project

Executive Summary - Regression Modeling

Project Overview

The TikTok data team seeks to develop an accurate predictive model that determines whether a video contains a claim or an opinion. The team built a regression model to investigate how variables are related to verified_status. This step is important because the end goal is to classify claims and opinions and it would be beneficial to investigate why verified users are more likely to post opinions.

Key Insights

- There are some strongly correlated variables which might lead to multicollinearity. (excluded video_like_count)

- Confirmed that opinions videos are more likely to be posted by verified users.

- Every second increase in video duration is associated with a 0.009% increase in probability of user having a verified status.

- Videos with high views, downloads, and/or comments are likely to be posted by unverified users.

- Model results:
Accuracy: 0.65
Precision: 0.61
Recall: 0.85

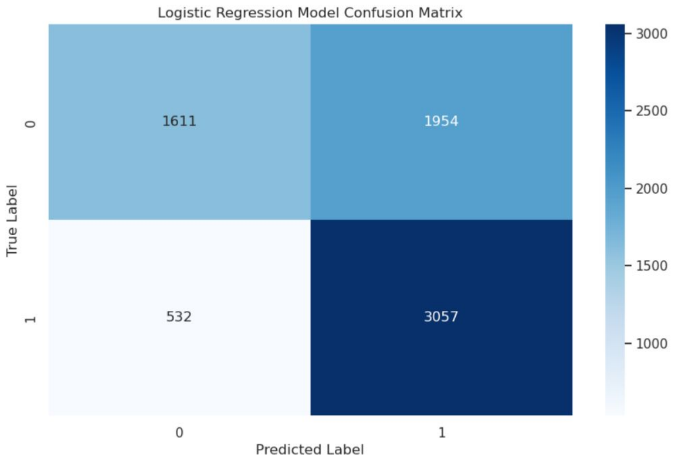
Details

Logistic Regression Model Assumptions (statements about the data that must be true in order to justify the use of a particular modeling technique):

- Linearly
- Independent Observations
- No Multicollinearity
- No Extreme Outliers

We checked against these assumptions and adjusted the data accordingly to meet the assumptions.

Confusion Matrix Labels:
0 = “Not Verified”
1 = “Verified”



Next Steps

Now that we have gathered enough information about user behavior and variable associations, we can move onto the final part of the project:

Construct a **classification model** that will predict whether a video contains a claim or an opinion.