Executive Summary

Regression Analysis

ISSUE / PROBLEM

Previously we learned that a user being verified significantly increased the probability of posting opinions. Since our final goal is to create a machine learning model that will predict whether a tiktok contains a claim or opinion, the predictors of verified_status could be great indicators

RESPONSE

Since a user is either verified or unverified, we will be using a logistic regression model. The dataset had class imbalance for verified_status, so we performed upsampling on the minority 'verified' class. It is important to consider this because of the possibility that the model, after upsampling the minority class, disproportionately identifies tiktoks as verified.

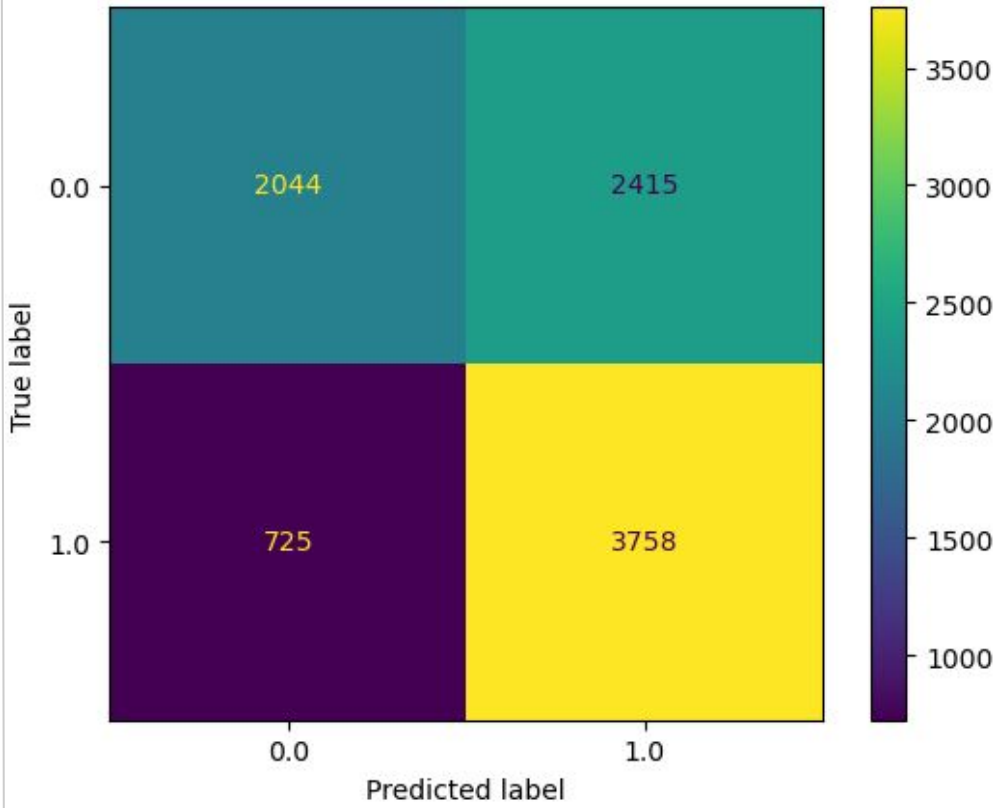
IMPACT

If our model finds that a given continuous variable is a well performing predictor for verified_status, then given the relationship between claim_status and verified_status, we will then have both a categorical and continuous variable for predictions.

Significant Class Imbalance:

```
verified_status
not verified    0.93712
verified        0.06288
```

verified_status Confusion Matrix



Top-left: unverified accounts predicted correctly
Top-right: unverified accounts predicted incorrectly
Bottom-left: verified accounts predicted incorrectly
Bottom-right: verified accounts predicted correctly

KEY INSIGHTS

The model performed adequately, predicting verified accounts with a precision of .61, recall of .84, and overall accuracy of .65. With the added context that longer tiktoks tend to be paired with higher odds of a user being verified, we can proceed with the machine learning model under the assumption that video_duration_sec will likely be a indicator for higher verified_status tiktoks, which we already established have significant correlation with claim_status.

	Feature Name	Model Coefficient
0	video_comment_count	-6.404235e-04
1	video_download_count	-1.099775e-05
2	video_duration_sec	8.607893e-03
3	video_share_count	5.930971e-06
4	video_view_count	-2.132079e-06
5	claim_status_opinion	3.908384e-04
6	author_ban_status_banned	-1.781741e-05
7	author_ban_status_under review	-9.682447e-07