Fessler 2024

TikTok Machine Learning Project

Logistic Regression Analysis of Verified Users

Project Overview

Our goal is to develop a machine learning model to correctly categorize user submissions as ‘claim’ or opinion. As we move closer to this goal, we must review and complete exploratory data analysis on the data. Through taking a look at the relationship of user verification and the other features captured in the data, we come a step closer to confidently building the machine learning model with the correct variable inputs and an understanding of the data that will allow for proper interpretation.

Key Insights

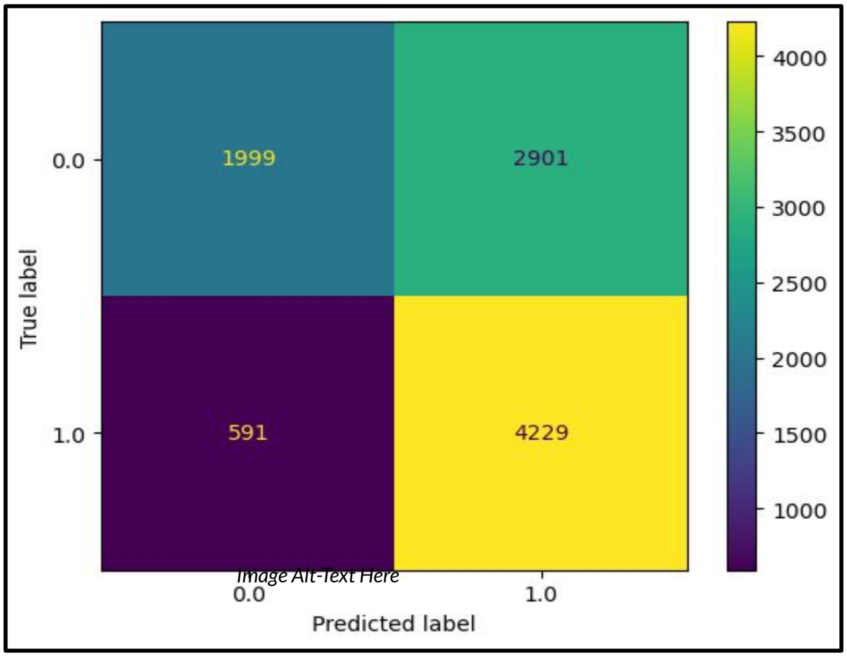
> Many of the features have definite to strong correlation, which may cause collinearity conflicts with the model. The most correlative feature is video_like_count with correlation values of as high as 0.85 with other features. It has been left out of the model's training.

> Video duration has a beta of 0.0079, meaning for each second there is 0.0079 increase in log-odds probability of the user being verified. Claim status has a beta of 0.00036, meaning for a value of 1 (an opinion) there is a 0.00036 increase in log-odds probability of the user being verified. The video download count's beta is -0.00023, meaning for each additional download there is a -0.00023 decrease in log-odds probability of the user being verified.

> Precision: 59%,Max recall: 88%, Max f1-score: 71%, Max accuracy: 64%, the weighted and macro averages are all down in the 60s. The model did perform; however, it does feel like there is room for improvement, since there was a lot misclassification.

> We were able to develop a logistic regression model to predict the verification of users with reasonable accuracy. The major features that contribute understanding are video_duration_sec, claim_status, and video_download_count. The duration of the video points towards longer videos will be more likely to be from a verified user. Claim status points towards more opinion focus videos are more likely to be from verified accounts. Finally, more downloads is most likely on videos from non-verfied users.

Details



We see 2901 False Positives and 591 False Negatives. True Positive is the correct categorizing of a verified user and true negative is that for an unverified user.

Next Steps

- A secondary look at logistic regression can be taken with the inclusion of ‘video_like_count’ while leaving out the variables that it is highly correlated with. Perhaps this will develop a more accurate regression model.
- Moving forward we can use our understanding of the data to better select variables for the machine learning model, and we can more confidently interpret our future findings