Executive Summary: Regression Analysis

TikTok claims classification project

OVERVIEW

The TikTok team wants to develop a machine learning model to help classify the claim_status of a specific video, as this factor is important for the TikTok team to decide whether to manually review a video. Since we have already found that verified users tend to post more opinion videos, in this part of the project, the data team plans to develop a logistic regression model that predicts verified_status.

PROJECT STATUS

Since the target variable for this prediction task is the discrete variable verified_status (unverified, verified, under review), this prediction task is a classification problem. We decided to use a logistic regression model to complete this task.
The logistic regression model has a precision of 68%, a recall of 65% (weighted average), and an f1 accuracy of 64%.

NEXT STEPS

The final goal for the data team is building a classification model to predict the claim_status of videos.

KEY INSIGHTS

According to the logistic regression model predicting verified author status, the video_duration_sec is the strongest column for predicting.
The model had decent predictive power with an overall accuracy of 65%. Based on the model coefficients estimated by the logistic regression, video_duration_sec had the greatest impact on determining verification status, with longer videos tending to be associated with a higher chance of the user being verified. The other video features had smaller estimated coefficients in the model, so they appear to be less associated with verification status.