Executive Summary

Milestone 2 of the TikTok Claims Classification Project

ISSUE / PROBLEM

The TikTok team wants to do some EDA and develop a machine learning model to help classify the claim_status of a specific video, as this factor is very important for the TikTok team to decide whether to manually review the video. At the beginning of this task, the data team must obtain some basic statistics about the dataset and prepare for future EDA and machine learning model building.

IMPACT

Based on the analysis in this assignment, we can check how different features of the video affect the claim status classification of a particular video. For example, video_view_count and video_like_count are two variables that are highly correlated with claim status.
In addition to this, we can also check whether the dataset distribution needs to be balanced, as an overly imbalanced dataset may lead to potential bias in the training model

RESPONSE

The data team examined basic statistics of the dataset and performed an initial investigation of the claim classification dataset with the goal of understanding important relationships between variables.
They examined the distribution of claim_status and author_ban_status and showed their relationship with certain engagement metrics of the video, such as video_share_count and video_like_count.

KEY INSIGHTS

* Videos of banned users receive the most engagement, such as video_view_count, video_like_count, and video_share_count. Videos of users under review also receive more engagement than active users, and less than banned users.
* Because we consider engagement levels per view, for videos that have been viewed, a video's author_ban_status does not significantly affect the video's engagement, so we can conclude that video's engagement is less correlated to author_ban_status and more related to its claim status.
* There are in total 19,382 video records in this dataset, 9608 of them are claim videos, which is 49,6%, and the count of opinion videos is 9476 - 48.9%
* Engagement levels like the video_view_count and the video_like_count are clearly correlated with claim status.