TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok team wanted to do some EDA and 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. In this part of the project, the data needs to be analysed, explored, cleaned, and structured before further data analysis and model building can be performed.

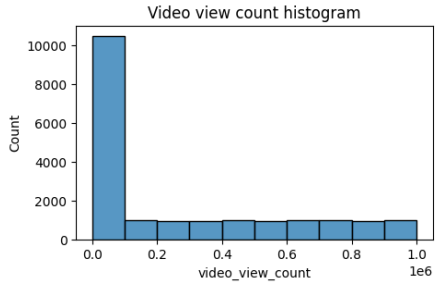
RESPONSE

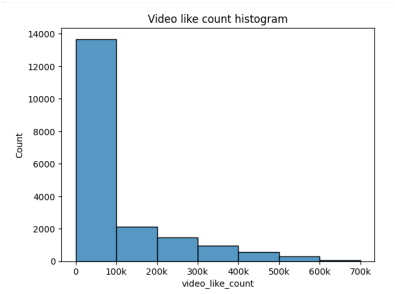
The TikTok data team conducted exploratory data analysis at this stage. The purpose of exploratory data analysis is to understand the impact of videos on TikTok users. To do this, the TikTok data team analysed variables that can show user engagement: views, likes, and comments.

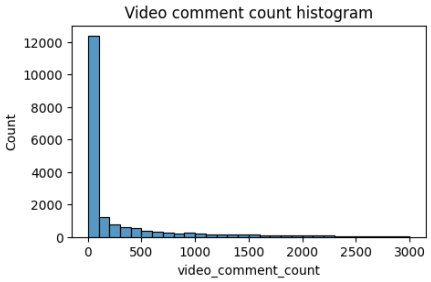
IMPACT

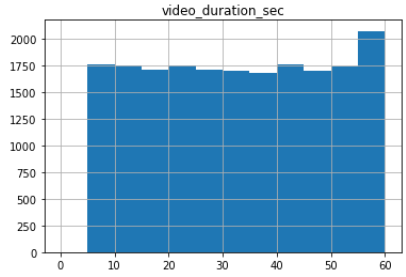
Future claim status classification models need to incorporate the imbalance of null values and opinion video counts into the model parameters.

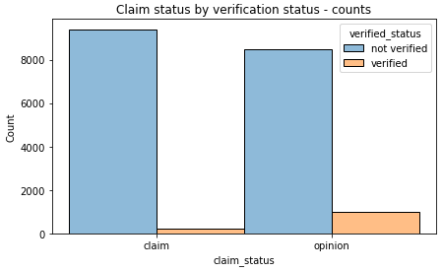
The distribution of video_duration_second is relatively balanced, while the distribution of other video engagement indicators is not balanced. Most of them are concentrated at the low end of the value, and the rest of the data tilts to the right, showing a large downward trend.
For both claim_status, the number of unverified users is significantly larger than verified users.
However, verified users tend to post more opinion videos.
Only a very small fraction of all videos are opinion videos, and most of them are claim videos.

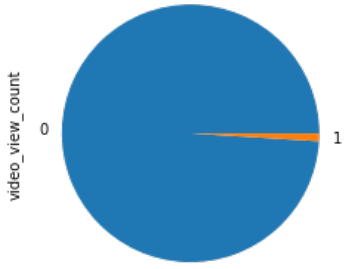












KEY INSIGHTS

A total of 298 records in the dataset contain null values. Therefore, we need to consider and handle null values to avoid drawing conclusions that assume the data is complete. Further analysis is needed to investigate the causes of these null values and their impact on future statistical analysis or model building.
Only a small percentage of all videos are opinion videos, and the majority of videos are advocacy videos.
The distribution of video_duration_second is relatively balanced, while the distribution of other video engagement indicators is not balanced. Most of them are concentrated at the low end of the value, and the rest of the data tilts to the right, showing a large downward trend.