TikTok Claims Classification Project

Exploratory Data Analysis

ISSUE / PROBLEM

An analysis of the claims classification project dataset is required to identify key variables and the relationship between them.

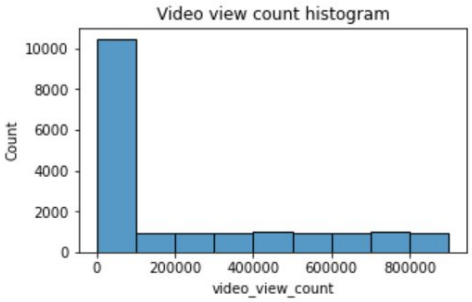
RESPONSE

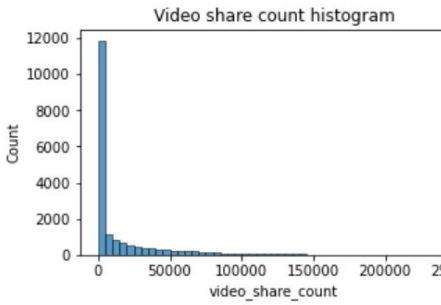
An exploratory data analysis (EDA) of the dataset was carried out. Distribution of variables were observed as box plots and histograms. Bar graphs and scatter plots were used to examine the relationship between variables.

IMPACT

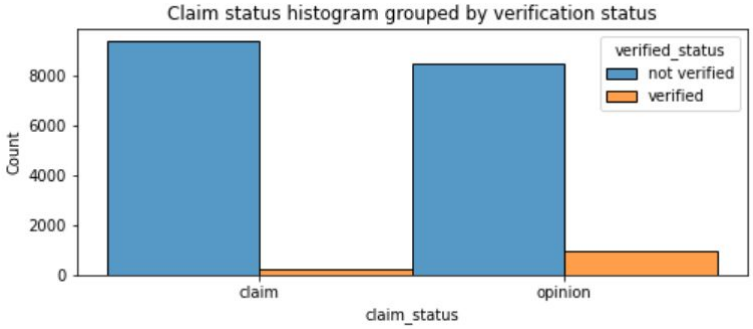
Key variables were identified as those representing user engagement with the video content (likes, comments etc.).

The histograms show the skew in the data i.e., more than half the videos have low user engagement (views, shares etc.)

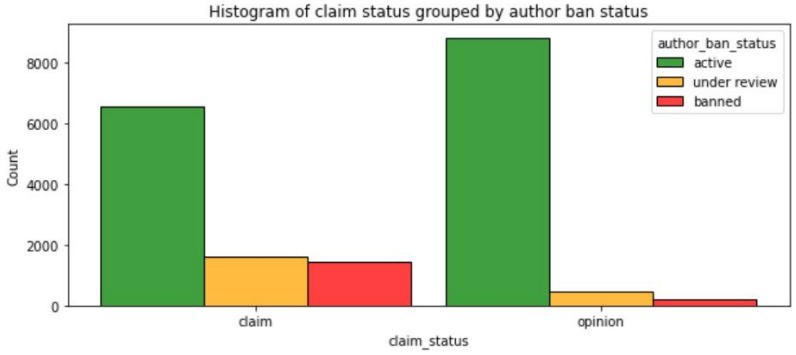




Majority accounts are unverified, and they are more likely to post claim videos.



Banned users are more likely to post claim videos whereas active users are more likely to post opinions.



KEY INSIGHTS

- The dataset is balanced i.e., number of claim videos = number of opinion videos.
- Variables representing user engagement (video_view_count, video_like_count, video_share_count, video_comment_count, video_download_count) were identified as key variables useful for model development.