TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok data team aims to create a machine learning model to help classify user-submitted claims. For this phase of the project, the data must be analyzed, explored, cleaned, and organized before proceeding with model development.

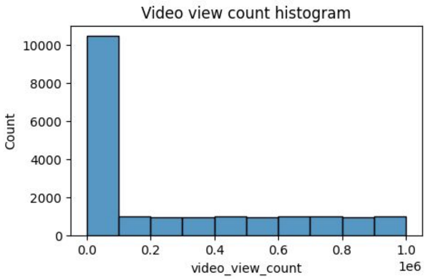
RESPONSE

Visualizing the data played a crucial role in the exploratory data analysis phase of this project. The histograms provided clearly demonstrate that the vast majority of videos in this dataset have low values for three variables that represent TikTok users' engagement with those videos. Most videos are clustered towards the bottom end of the range of values for these engagement metrics.

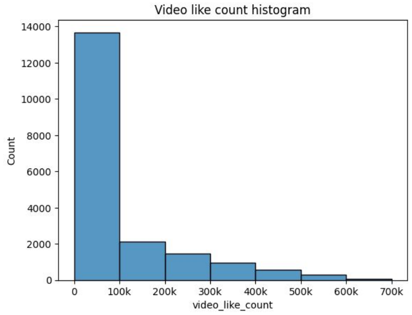
IMPACT

Based on the exploratory data analysis findings, the upcoming claim classification model must address null values and the imbalance in the counts of opinion videos by incorporating these factors into the model parameters.

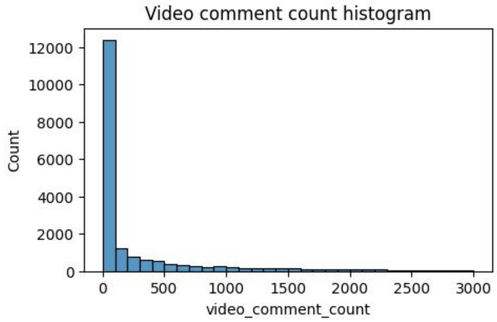
An essential part of this project's exploratory data analysis is visualizing the data. As shown in the following histograms, it’s evident that most videos are clustered at the lower end of the value range for three variables that represent TikTok users' (video viewers') engagement with the videos in this dataset.



The view count variable shows a highly uneven distribution, with over half of the videos receiving fewer than 100,000 views. For videos with view counts exceeding 100,000, the distribution is uniform.



Much like with view count, there are significantly more videos with fewer than 100,000 likes than those with more.



Once again, most videos fall at the lower end of the spectrum for comment count, with the majority having fewer than 100 comments. The distribution is highly right-skewed.

KEY INSIGHTS

The exploratory data analysis performed by TikTok's data team uncovered several important considerations for the classification model, including missing values and uneven data distributions. Two key takeaways from this analysis are:

Missing values
Approximately 200 null values were identified in 7 different columns, which should be taken into account to avoid making assumptions about complete data. Further investigation is necessary to determine the cause of these missing values and assess their potential impact on future statistical analysis or model development.

Skewed data distribution
The data shows a significant skew to the right, particularly in video view and like counts, which are concentrated on the lower end of the scale for opinion videos. This insight will inform the selection of suitable models and model types that can effectively handle this type of data distribution.