TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok analytics team aims to create a machine learning algorithm to help categorize user-submitted claims. This stage of the project involves examining, understanding, tidying, and organizing the data before starting to construct the model.

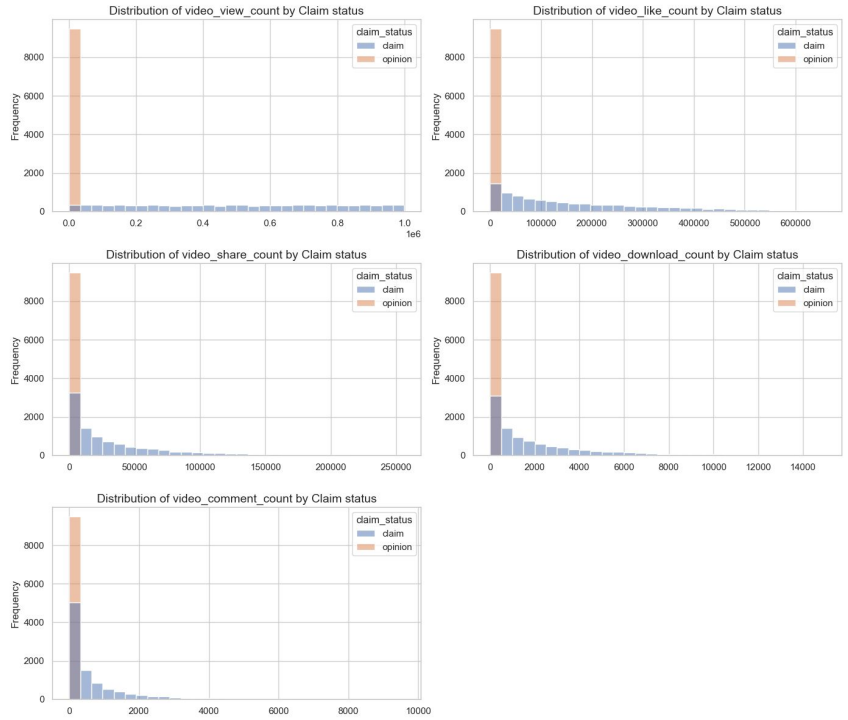
RESPONSE

At this phase, the TikTok data team undertook an exploratory data analysis to comprehend how videos influence TikTok users. They aimed to grasp user engagement by examining metrics such as views, likes, and comment counts.

IMPACT

The insights from the exploratory data analysis indicated that the upcoming claim classification model must consider null values and uneven distributions of opinion video counts by integrating these factors into the model's parameters.

A key component of this project's exploratory data analysis involves visualizing the data. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase TikTok users (video viewers') engagement with the videos included in this dataset.



The view count variable has a very uneven distribution, with more than half the videos receiving fewer than 100,000 views. Distribution of view counts > 100,000 views is uniform.

Similar to view count, there are far more videos with < 100,000 likes than there are videos with more.

Again, the vast majority of videos are grouped at the bottom of the range of values for video comment count. Most videos have fewer than 100 comments. The distribution is very right-skewed.

KEY INSIGHTS

The exploratory data analysis carried out by TikTok's data team uncovered several important factors for the classification model, such as handling missing values, achieving a balance between "claims" and "opinions," and understanding the overall spread of data variables. From this analysis, two main insights emerged:

Null values

The analysis found over 200 missing entries in seven columns, indicating that future models should incorporate these null values to ensure accuracy. Further exploration is needed to determine the cause and impact of these missing values on model development.

Skewed data distribution

Video view and like counts are all concentrated on low end of 1,000 for opinions. Therefore, the data distribution is right-skewed, which will inform the models and model types that will be built.