TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok data team is working on creating a machine learning model to help categorize claims from user submissions. In this phase of the project, it is essential to examine, explore, tidy up, and organize the data before proceeding with the development of the model.

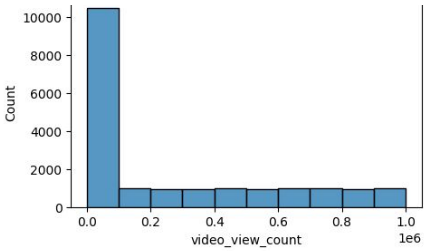
RESPONSE

In this stage, the TikTok data team conducted exploratory data analysis to understand how videos impact TikTok users. To accomplish this, the team examined variables that offer insights into user engagement, with a particular emphasis on metrics such as views, likes, and comment counts.

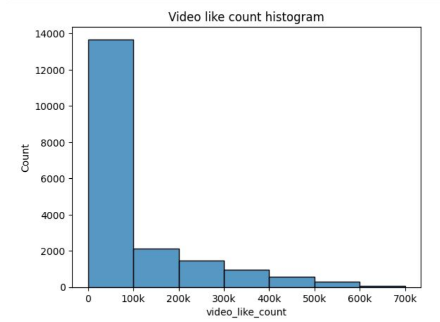
IMPACT

Drawing from the findings of the exploratory data analysis, it is clear that the upcoming claim classification model needs to tackle both null values and the uneven distribution of opinion video counts. This necessitates the incorporation of these considerations into the model parameters.

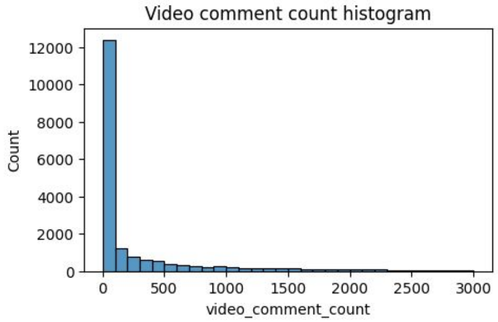
An essential element of the exploratory data analysis in this project involves visualizing the data. The histograms displayed below vividly demonstrate that the bulk of videos are centered at the lower range of values for three variables representing the engagement of TikTok users (video viewers) with the videos contained in this dataset.



The distribution of the view count variable is notably uneven, with over half of the videos receiving fewer than 100,000 views. Conversely, the distribution of view counts exceeding 100,000 is more uniform.



Similar to the view count, there is a notable imbalance in the distribution of likes, with a considerably higher number of videos having fewer than 100,000 likes compared to those with more.



Once more, the overwhelming majority of videos cluster towards the lower end of the value range for video comment count. Most videos have fewer than 100 comments, and the distribution is heavily right-skewed.

KEY INSIGHTS

The exploratory data analysis conducted by TikTok's data team unveiled several crucial considerations for the classification model, encompassing aspects such as missing values, the balance between "claims" and "opinions," and the overall distribution of data variables. The two primary insights derived from this analysis were:

Null values

More than 200 null values were identified across seven distinct columns. Consequently, forthcoming modeling efforts should account for these null values to prevent making assumptions based on incomplete data. Further analysis is imperative to explore the reasons behind these null values and assess their potential impact on future statistical analysis or model building.

Skewed data distribution

View and like counts for opinion videos are predominantly concentrated at the lower end, around 1,000. This indicates a right-skewed data distribution, providing valuable insights for the construction of models and the selection of appropriate model types.