Tiktok Claims Classification Project

Exploratory Data Analysis - Executive Summary

ISSUE / PROBLEM

The data team is trying to make a machine learning model that could predict the claim status of a submitted video. EDA is being conducted to understand the data before model training.

RESPONSE

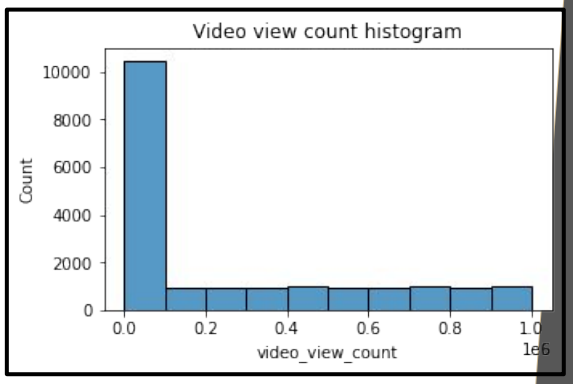
In this stage, the team conducted EDA to explore the data in depth understanding the structure and distribution of data. Also the user engagement was studied and their relation with the claim status.

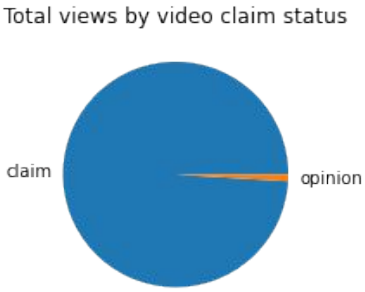
IMPACT

During the initial look of the data, around 200 null values were identified. The number of claim and opinion videos is around 50% which is a good distribution.

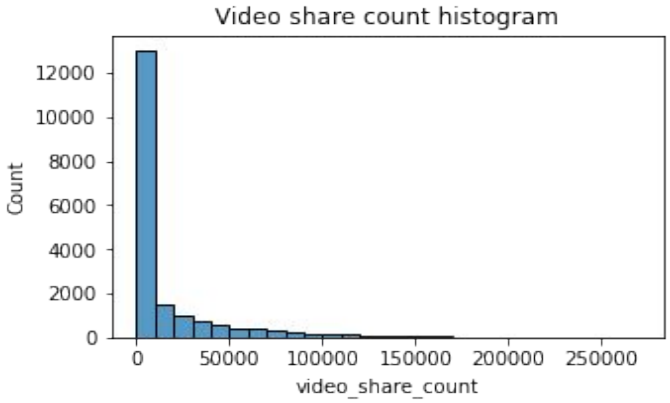
The engagement statistics (i.e. likes, comments, shares and downloads) all follow a heavily right skewed distribution. Most of the videos lie within the lowest range of values.

Views of videos also fall into the lowest range with all other ranges following a uniform pattern.





Views of the video are very unevenly distributed with claims having the highest views even though their counts are almost same.



KEY INSIGHTS

The EDA provided few key insights that need to be considered before model training.

- Data contains null values that need to be considered, for the model not to assume complete data.
- Uneven distribution with most values concentrated in the lower ranges would impact the model.
- Few values such as number of views, ban status and verified status show a high correlation with claim status, which can be considered as potential predictors.