TikTok Machine Learning Decision Model

Opinion vs Claim Exploratory Data Analysis

Overview

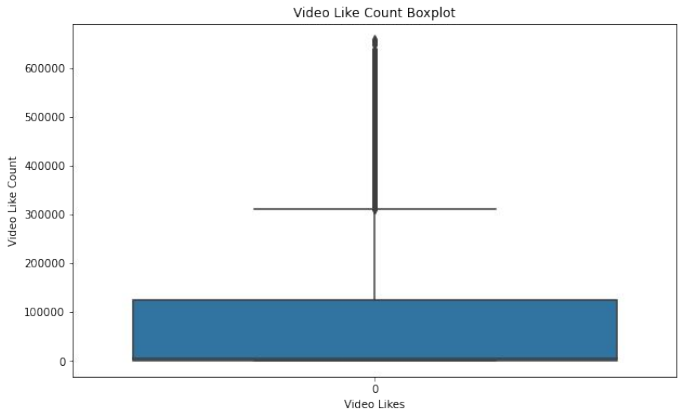
Exploratory Data Analysis (EDA) was completed on the data set, it was observed that there are multiple examples of correlation indicating differences between 'opinion' and 'claim' submissions.

Problem

EDA was faced with null values and many extreme outliers

Solution

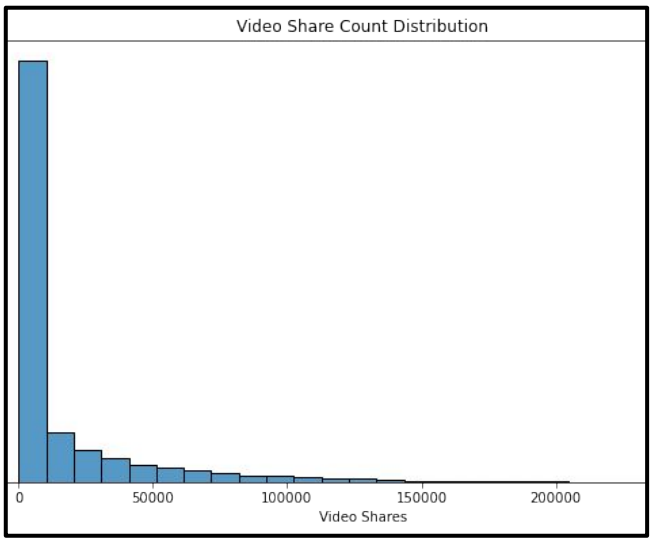
The null values were a minor subset of the data and were ignored. Additionally, the outliers are potentially necessary for the correct telling of the data stories, since they are artifacts of social media viralism



The extreme number of outliers are seen in all the public metrics, including "Like Count"

Details

- A comprehensive review of the data was completed through the EDA process, looking at the author and public generated data.
- There is a distinct increase in views, shares, comments, likes, and downloads for 'claim' submissions.
- Active and Verified users much less commonly post videos that are marked as 'claim', while authors that are under review or banned and non-verified are very commonly submitting videos marked as 'claim'.
- The length of submissions does not have any correlation across all submission lengths.
- Major Takeaway: there does appear to be a real binary situation between 'opinion' and 'claim' submissions, where all of the major metrics are heavily skewed towards the claim category.



The large number of outliers is also observable from the strong right distribution. This also really shows the effect of the low activity submissions.

Next Steps

- Address outliers. Where should we cut off outliers? Potential after the 90th quartile? Additionally, as much as upper outliers and big numbers are common with social media viral content, is there a way to address the staggering amount of data points with very few or zero public interaction? Should we set a lower bound?
- Move forward with developing the Machine Learning model, incorporating many of public engagement quantitative variables and the author qualitative variables.