TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok Data team wants to create a model to assist in determining whether videos are claims or opinions. A machine learning model is request of the data team and we're currently working on the EDA of the project.

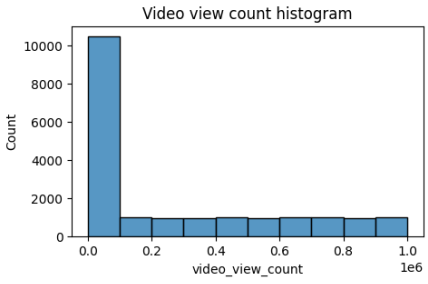
RESPONSE

Our team have conducted an EDA on the provided data. The purpose of the exploratory data analysis was to understand the impact that videos have on TikTok users. Variables such as views, comments, and share counts were analyzed. Correlations were to be found using these variables as well as other descriptive statistics.

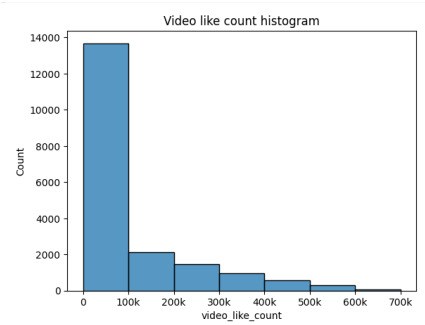
IMPACT

The findings from the EDA has provide insight on what is necessary for the future claims classification model. The model should include to handle many videos that have null values.

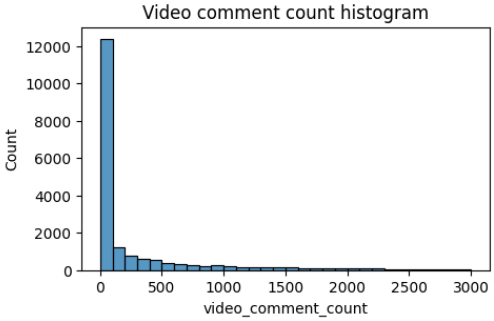
Part of this projects visualizations start with an EDA. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase TikTok users (video viewers') engagement with the videos included in this 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 conducted from TikTok's data team revealed many considerations for the classification model, including missing values, "claims" to "opinions" balance, and overall distribution of data variables. The two key insights from this analysis were:

Null values

Hundreds of null values were present in multiple columns. Future modeling should consider the null values and the impact it has on the model. The reason for these null values will be discovered with further analysis. Further investigation will be necessary to determine the reason.

Skewed data distribution

As shown in the visualizations above, all counts are right-skewed. This needs to be taken into consideration before creating the model.