Claims Classification Project | Milestone Three

Executive summary prepared for TikTok leadership by the TikTok data team

Overview

TikTok leadership asked the TikTok data team to develop a ML model that will help us better understand user submissions and reduce the backlog of user reports. In this part of the project, the data team needs to conduct exploratory data analysis on the provided data set.

Problem

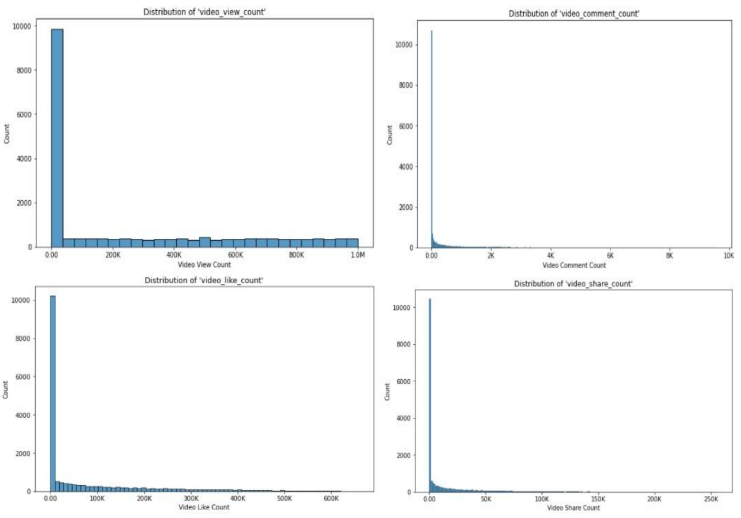
What distinguishes claim videos from opinion videos is of particular interests for TikTok because effective prioritization and management of user reports are crucial for maintaining user satisfaction and allocating the right amount of time and effort to resolve future issues. To build a machine learning model that can streamline the claims process, the data teams needs to conduct the exploratory data analysis phase of the data prior to any modeling of it.

Solution

To obtain a model with the highest predictive power, the TikTok data team in this data analysis phase examined the provide date set and prepared it for further analysis. To learn more about the variables and their relationships, the data team imported the data, access it via data exploration and cleaning, and built visualizations to share with the TikTok team.

Details

- **Performing data cleaning and exploration and building boxplot visualizations to understand variables and outliers.** We examine the `video_like_count`, `video_comment_count`, `video_share_count`, and `video_download_count` variables and discover boxes of the interquartile range are squished at one end because the outliers at the other end take up all the space.



- **Using descriptive statistics to gain a better grasp of the data,** we calculate the interquartile range, the median, the outlier threshold, and the number of outliers for several variables. The following screenshot indicates this:

```
Number of outliers, video_view_count: 2343
Number of outliers, video_like_count: 3468
Number of outliers, video_share_count: 3732
Number of outliers, video_download_count: 3733
Number of outliers, video_comment_count: 3882
```

- **There appears to be positively skewed distributions for most of the variables** as indicative in the histograms.

Next Steps

- **The data team recommends investigating the discrepancies for both claims and opinions between the number of active authors and banned non-active authors.** We need to effectively prepare our data for analysis and modeling, understanding the nature of the outliers and their potential impact and building models with and without them to compare model performance.

- **The data team will use inferential statistics to bridge the gap between the descriptive statistics found in milestone two and three and building a predictive model**. We need to determine the most relevant variables for running regression and statistical analysis.