Claims Classification Project (Milestone 3)

Exploratory Data Analysis (EDA) – Prepared by PhD Aleksandar Osmanli

ISSUE / PROBLEM

The video company client seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, the data needs to be analyzed, explored, cleaned, and structured prior to any model building.

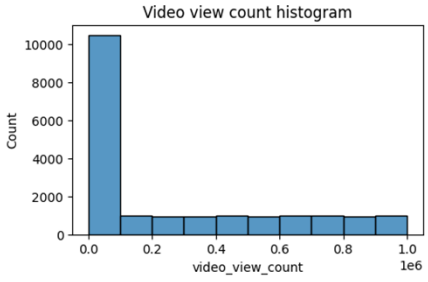
RESPONSE

I conducted exploratory data analysis at this project milestone. The purpose of the EDA analysis was to understand the impact that videos have on the users. To do so, I analyzed variables that would showcase user engagement: view, like, and comment count.

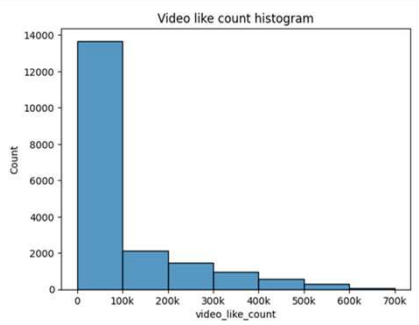
IMPACT

According to the findings from the exploratory data analysis, the future claim classification model will need to account for null values and imbalance in opinion video counts by incorporating them into the model parameters.

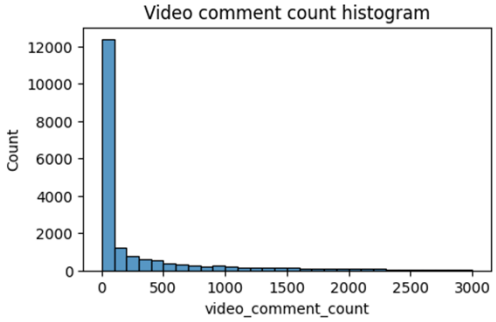
A key component of this project's EDA analysis involves visualizing the data. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase the users (video viewers') engagement with the videos 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 conducted EDA analysis 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

Over 200 null values were found in 7 different columns. As a result, future modeling should consider the null values to avoid making insights that would assume complete data. Further analysis is necessary to investigate the reason for these null values, and their impact on future statistical analysis or model building.

Skewed data distribution

Video view and like counts are all concentrated on low end of 1,000 for opinions. Therefore, the data distribution is right-skewed, which will inform the models and model types that will be built.