TikTok Machine Learning Model Project

Statistical Testing and Review

Overview

The machine learning model that is the target for development of this project. We must understand the correlations between variables of the dataset so that we can better train a model to help our moderators for determining 'claim' versus 'opinion'.

Objective

- Determine the statistical significance of video view counts and the verified status of the user.
- Review descriptive statistics for better interpretation.

Results

- The number of views by non-verified accounts is on average nearly triple to that of the view counts of verified accounts.
- The null hypothesis is that there is no major difference between the view counts of verified and non-verified accounts. The alternative hypothesis is that there is a statistically significant differences between the two data groups.
- With a p-value of $2.6088823687177823e-120$ we easily reject the null hypothesis

Next Steps

- Implement testing to better understand this major difference in view counts, potentially through collecting new controlled data and running another A/B testing.
- Build a regression model for verified status, keeping in mind that the data is skewed toward the more viral 'not verified' data group.