Executive Summary: Hypothesis Test: Verified/Unverified Accounts

Statistical Analysis of Number of Views

Overview

At this stage in the project, the TikTok Data Team has completed important steps in preparing the data for a claims classifier. We've completed an initial data discovery and a deeper exploratory data analysis that revealed correlations that needed to be further studied from a statistical standpoint. A statistical analysis was done to explore the significance between certain engagement activity and verified status.

Objective

The TikTok data teams aims at determining if there is a statistical significance between the number of views of verified accounts and the number of views of non-verified accounts. A hypothesis test will be conducted on the average number of views between verified and not verified accounts with a significance level of 5% using a two-tailed t-test.

Results

- Our experimental set up was as follows:
  - H_0 - There is no difference in the average number of views between verified and non-verified accounts.
  - H_A - There is a difference in the average number of views between verified and non-verified accounts.
  - Since we are comparing the means of two independent samples, we used a two-tailed ttest (in both directions).
  - We chose a significance level of 5% which is appropriate for out objective.
- Our results showed that we had t-score: -25.49, pvalue: 2.60e-120, indicating that we can reject the null hypothesis and that the difference in views is highly significant.
- We can say with 5% significant level that there would be an absolute difference of the observed means as or more extreme than what was observed if the null hypothesis were true.

Next Steps

The extremely low p-value indicates there is some root cause analysis that should be explored further into the behavior of these groups before making a final regression model.

Next steps is to build a regression model on verified status as a natural next step, before the machine learning modeling on claim status.