Statistical Review for TikTok Claims Prediction Project

Executive summary report

Overview

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform. The team has successfully completed exploratory data analysis on the data for the claims classification project. The team is ready to begin the process of hypothesis testing.

Problem

The leadership team is interested in whether there is a statistical difference in the data between verified and unverified accounts.

Solution

The data team conducted a two-sample hypothesis test of verified versus unverified accounts in terms of video view counts. The videos of unverified accounts tend to have higher view counts than the videos of verified accounts.

Details

Steps conducted in the hypothesis test:

1. Dropped the rows with missing values from the dataset.
2. Calculated the average values of video view counts for each verified status value of the user accounts.
3. Divided the dataset rows into two groups:
    a. Videos of verified accounts.
    b. Videos of unverified accounts.
4. Conducted a two-sample t-test to determine if there is a statistically significant difference in the average video view count between verified and unverified accounts.

Hypothesis test results:

There is a statistically significant difference in the average video view count between videos of unverified accounts and verified accounts.

Next Steps

Consider applying the verification status variable of the dataset into the development of the claims prediction model.