XGBoost Predictive Model for TikTok's Claim Classification: EDA, Hypothesis Testing, Logistic Regression, Tree-Based Models

Executive Summary for Statistical Testing

Overview

The data team seeks to develop a machine learning model to facilitate the classification of claims for user submissions. For this section of project the data team will test hypothesis to examine the relationship between verified_status and video_view_count.

Key Insights

- The analysis displays that there is a difference in views between videos posted by verified accounts and videos posted by unverified accounts.
- Findings suggest there can be fundamental behavioral differences between these two account groups: verified and unverified.
- It would be interesting to analyze the cause of this behavioral difference. For example, consider:
    - Do unverified accounts tend to post more engaging videos? Is that engaging content a claim or opinion?
    - Or, are unverified accounts associated with spam bots that help inflate view counts?

Details

The data team considered the relationship between verified_status and video_view_count.

One approach was examining the mean values of video_view_count for each group of verified_status. Unverified accounts have a mean of 265,663 views vs. 91,439 views for verified accounts

```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64
```

The second approach was a two-sample hypothesis test. This statistical analysis concluded that any observed difference in the sample data is due to an actual difference in the corresponding population means, aligning with preliminary findings from the mean values.

Next Steps

The team advices moving forward and building a **regression model** on verified status.

A regression model for verified_status can be helpful to illustrate user behavior of verified users. Later, this context can be utilized to consider results from a claim classification model that will be created at downstream.