Executive Summary: Statistical Testing Results

TikTok Claims Classification Project

Project Overview

The TikTok data team is developing a machine learning model to classify user-reported content as claims, aiming to reduce the backlog of user reports. As part of this project, the team is conducting a hypothesis test to analyze the relationship between verified_status and video_view_count. This statistical analysis will provide insights that contribute to the model's development.

Key Insights

The analysis shows a statistically significant difference in video view counts between verified and unverified TikTok accounts. This suggests that verified and unverified accounts may exhibit different behavioral patterns when it comes to content engagement. Further research is needed to explore the reasons behind this difference. For example, consider:

- Do unverified accounts tend to post more engaging or clickbait-style content?
- Is this content opinion-based or classified as claims?
- Are some unverified accounts potentially associated with spam bots that inflate view counts?

Investigating these factors will offer deeper insights into content engagement dynamics and help refine TikTok's content classification model.

Details

The TikTok data team analyzed the relationship between verified_status and video_view_count using two approaches. The first revealed that unverified accounts averaged 265,663 views, while verified accounts averaged 91,439 views.

```
verified_status
not verified    265663.785
verified         91439.164
Name: video_view_count, dtype: float64

verified_status
not verified    17884
verified         1200
Name: count, dtype: int64

pvalue=np.float64(2.6088823687177823e-120)
```

This image shows the mean view counts, group counts, and p-value for verified and non-verified accounts.

The second approach, a two-sample hypothesis test, confirmed that the observed difference in views between the two groups is statistically significant and likely reflects true population differences, rather than random chance.

Next Steps

The next step after this statistical analysis is to build a regression model to further analyze the impact of verified status on video view count.

Verified status is the independent variable, as it may influence video view count, which serves as the dependent variable and the outcome being analyzed. By using regression modeling, the team aims to understand how verified status affects video engagement, providing valuable insights into user behavior for further model development.