Claims Classification Project | Milestone Four

Executive summary prepared for TikTok leadership by the TikTok data team

Overview

TikTok leadership asked the TikTok data team to develop a ML model that will help us better understand user submissions and reduce the backlog of user reports. In this part of the project, the data team needs to conduct statistical analysis on the provided data set and carry out hypothesis tests about the relationship between the variables of 'verified_status' and 'video_view_count' in the sample data.

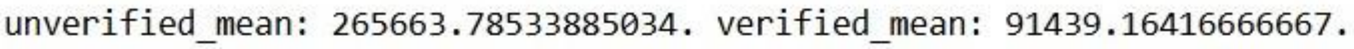
Objective

🎯 **Target Goal:** Develop a two-sample hypothesis test to analyze and determine whether there is a statistically significant difference between mean value of `video_view_count` for each group of `verified_status` in the sample data.

🎯 **Impact:** Statistical tests, such as the one conducted for Milestone Four, enable the TikTok data team to make inferences about the populations from which the data was drawn and help them learn more about user submissions.

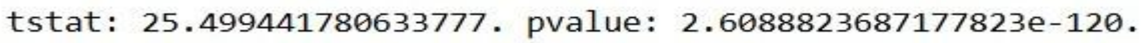
Results

- **Using descriptive statistics to conduct Exploratory Data Analysis (EDA) during data exploration,** we see that the mean for each different verified_status are different to each other. The following screenshot indicates this:



```
unverified_mean: 265663.78533885034. verified_mean: 91439.16416666667.
```

- **Conducting a two-sample t-test to determine whether there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts**, we formulate a null hypothesis stating of "no difference" and a two-sided alternative hypothesis stating of a difference. We choose a significance level of 5% and calculate the p-value. The following screenshot indicates this:



```
tstat: 25.499441780633777. pvalue: 2.6088823687177823e-120.
```

- **Based on the tests, we reject the null hypothesis, meaning one can conclude that the mean video_view_count are different for each group in `verified_status`**. Thus, there is a statistically significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts.

Next Steps

- **The data team recommends building a regression model on verified_status that helps us analyze user behavior.** We need to build a logistic regression model since our goal is to make predictions on the categorical variable of claim status.
- **The data team recommends investigating further the root cause of the behavioral difference between verified accounts and unverified accounts.** For example, are unverified accounts associated with bots helping inflate view counts.