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

Executive Summary for Prior Examining

ISSUE / PROBLEM

The TikTok team wants to develop a machine learning model to classify claims for user submissions. To begin, the data team is required to organize the raw dataset and prepare it for EDA.

RESPONSE

The team conducted a preliminary investigation of the claims classification dataset with the goal of uncovering critical relations between variables.

Given the ask for a classification of user claims, the data team looked at the counts of claims and opinions in order to understand the count of each type of video content.

IMPACT

To understand the impact of user videos, the team identified two critical variables to consider: video_duration and video_view_count. Both variables are important factors to consider for future prediction models.

UNDERSTANDING THE DATA

After initial review of the dataset, claim_status variable is seemed particularly useful for the project goal. The following images display crucial points of analysis needed to understand claim_status.

```
data['claim_status'].value_counts()
```

```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

Note: Each claim status counts are quite balanced. There are 9,608 claims and 9,476 opinions.

ENGAGEMENT TRENDS

The team considered viewer engagement of each video in the claim and opinion categories. To understand viewer engagement, the view count is considered. The mean and median view count demonstrated the impact of each category; particularly, the mean and median view counts for both categories show the association between claim/opinion and the video views.

Claims:

Mean view count claims: 501029.4527477102
Median view count claims: 501555.0

Opinions:

Mean view count opinions: 4956.43224989447
Median view count opinions: 4953.0

KEY INSIGHTS

- There is a near equal balance of opinions versus claims.
- With the key variables identified and the initial investigation of the claims classification dataset, the process of exploratory data analysis can begin.

Video numbers of claim and opinion are balanced in the dataset.

```
claim_status
claim      9608
opinion    9476
Name: count, dtype: int64
```

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

Executive Summary for Exploratory Data Analysis (EDA)

ISSUE / PROBLEM

The TikTok project aims to develop a machine learning model for the classification of claims for user submissions. In this part, the dataset required to be analyzed, explored, cleaned, and structured prior to any model building.

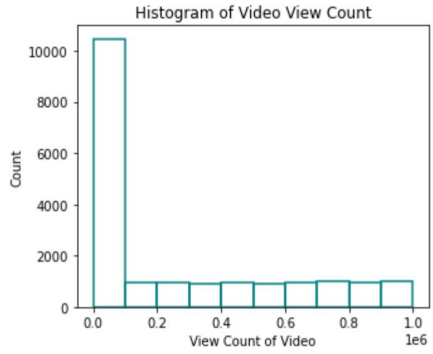
RESPONSE

The data team carried out exploratory data analysis on the dataset. The goal of EDA was to investigate the impact that videos have on TikTok users. Therefore, the data team analyzed variables that would indicate user engagement: view, like, and comment count.

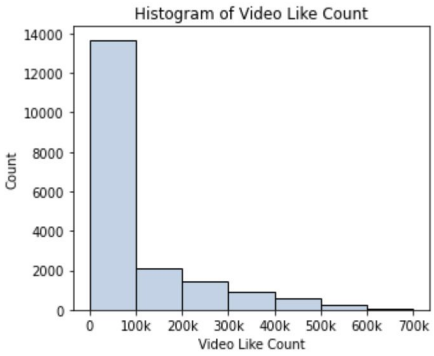
IMPACT

Based on conclusions from the EDA, the claim classification model should consider null values and imbalance in opinion video counts by incorporating them into the model parameters.

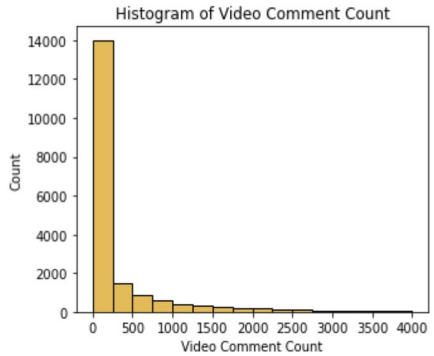
A essential component of this project's EDA is visualizing the data. Following histograms illustrates that vast majority of videos are grouped at the bottom of value range for three variables of interest (video view count, video like count, video comment count – related to user engagement).



The view count variable has uneven distribution. More than half of the videos receives fewer than 100,000 views.
Distribution of view counts > 100K views is uniform.



There are far more videos with < 100K likes than there are videos with more.



The majority of videos are grouped at the bottom of range for video comment count. Most videos have fewer than 100 comments. The distribution is very right-skewed.

KEY INSIGHTS

The completed EDA on the Tiktok data Project revealed various considerations for the classification model, including missing values, "claims" to "opinions" balance, and overall distribution of data variables. The two fundamental insights from this analysis were:

Null values
Null values were found in 7 different columns. Therefore, future modeling should consider null values prior to generate insights. Further analysis is advised for understanding reasons for these null values, and their possible impact on statistical analysis or model building in the downstream.

Skewed data distribution
Video view, like, and comment counts are all concentrated on low end value range. The data distribution is right-skewed, which will impact the models and model types that will be built.

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.

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

Executive Summary for Regression Analysis

OVERVIEW

The data team aims to develop a machine learning model to assist in the classification of claims for user submissions. Previously, the data team showed that if a user is verified, they are much more likely to post opinions. Hence, ultimate purpose is to classifying claims and opinions, it's crucial to build a model that demonstrates how to predict the behavior of the account type (verified) that likely to post opinions. Thus, a logistic regression model that predicts verified_status was built.

PROJECT STATUS

The 'verified_status' variable was analyzed in this regression model since the relationship between account type and the video content was indicated previously. A logistic regression model was selected because of the data type and distribution.

	precision	recall	f1-score	support
verified	0.74	0.45	0.56	4459
not verified	0.61	0.84	0.71	4483
accuracy			0.65	8942
macro avg	0.67	0.65	0.63	8942
weighted avg	0.67	0.65	0.63	8942

The logistic regression model yielded precision of 67% and a recall of 65% (weighted averages). This model had F1 score of 63%.

The logistic regression model had decent predictive power.

NEXT STEPS

Constructing a classification model that will predict the claim status made by users is next. That is the original expectation from the TikTok project.

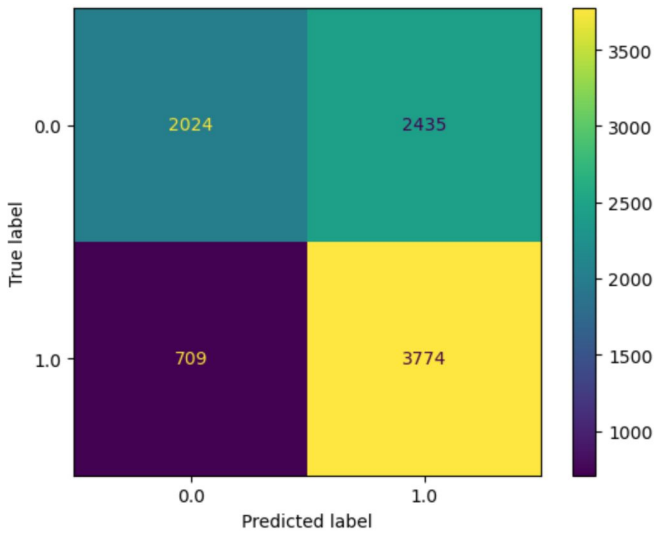
KEY INSIGHTS

Based on the estimated model coefficients from the logistic regression, longer videos tend to be associated with higher odds of the user being verified.

Other video features have small estimated coefficients in the model, so their association with verified status seems to be small.

All in all, other video features do not seem to be associated with verified status except video length.

Confusion Matrix for Logistic Regression Model



0: Not verified account & 1: Verified account.
Upper-left: true negatives. Upper-right: false positives.
Lower-left: false negatives. Lower-right: true positives.

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

Executive Summary for Tree-Based Classification Models

Overview

The TikTok team wants to develop a machine learning model to assist in the classification of videos as either claims or opinions. Previous investigation into the dataset indicated that video engagement levels were highly indicative of claim status.

Problem

TikTok videos receive a large number of user reports for many different reasons. Not all reported videos can be reviewed by a human moderator. Videos that make claims are much more likely to contain content that violates the TikTok's terms of service. The goal is developing method to identify videos that make claims to prioritize them for review.

Solution

The data team built two tree-based classification models. Both models were used to predict on validation dataset, and final model selection was determined by the model with the best recall score. The final model was then used to score a test dataset to estimate future performance.

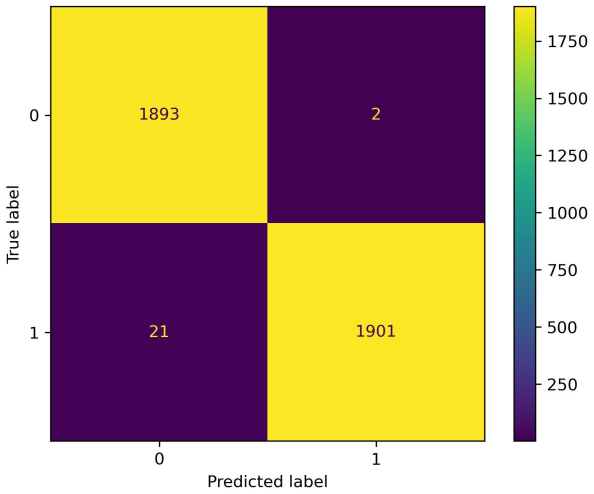
Details

Both random forest and XGBoost models performed exceptionally well. The XGBoost had a better precision and recall score and was selected as the champion.

Performance on the test holdout data yielded near perfect scores, with only 23 misclassified samples out of 3,817.

Subsequent analysis indicated that the primary predictor was 'video view count' -- related to video engagement levels. In conclusion, videos with higher user engagement levels were much more likely to be claims. In fact, no opinion video had more than 10,000 views.

Confusion matrix for the champion XGBoost model on test holdout data shows only 23 misclassified samples out of 3817.



Next Steps

Before deploying the model, further evaluation using additional subsets of user data is recommended. Moreover, the data team suggests monitoring the distributions of video engagement levels to ensure that the model remains robust to fluctuations in its most predictive features.