Executive Summary

TikTok Claims Classification Project – Data Inspection and Initial Analysis

ISSUE / PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. To begin, the data team needs to organise the raw dataset and prepare it for future exploratory data analysis.

RESPONSE

The data team performed a preliminary investigation of the claims classification dataset with the aim of learning important relationships between variables.

Given the request 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

The impact of this preliminary analysis will be evident in the next steps. In order to understand the impact of user videos, the data team identified two important variables to consider. video_duration (in seconds) and video_view_count are both important factors to consider for future prediction models.

UNDERSTANDING THE DATA

After reviewing the provided dataset, the variable claim_status seemed particularly useful, given the client's proposed project. The following screenshots show important points of analysis required to understand the claim_status variable.

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

claim_status	count
str	u32
"claim"	9608
"opinion"	9476
null	298

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

ENGAGEMENT TRENDS

The data team considered viewer engagement with each video in the claim and opinion categories. In order to understand viewer engagement, the data team considered the view count. The mean and median view counts show the impact of each category of video. Specifically, the mean and median view counts for both categories show the association between content (claim or opinion) and the video views.

Claims:

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

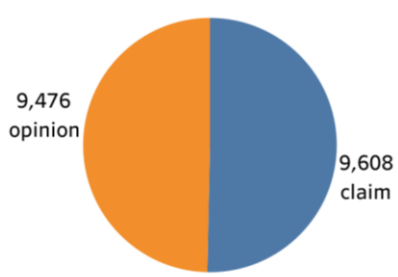
Opinions:

```
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 this understanding, we can proceed with our future analysis knowing that there is a fairly balanced amount of claims and opinions for the videos included within this dataset.
- With the key variables identified and the initial investigation of the claims classification dataset, the process of exploratory data analysis can begin.

Pie chart visualises the comparison of the count of claims and opinions



TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, the data needs to be analysed, explored, cleaned, and structured prior to any model building.

RESPONSE

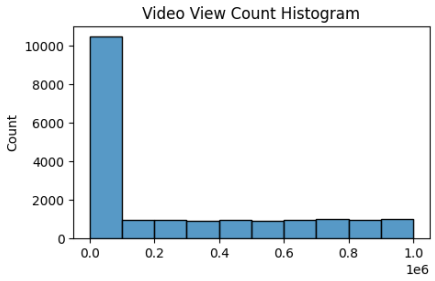
The TikTok data team conducted exploratory data analysis (EDA) at this stage. The purpose of the EDA was to understand the impact that videos have on TikTok users.

To do so, the TikTok data team analysed variables that would showcase user engagement: view, like, and comment counts.

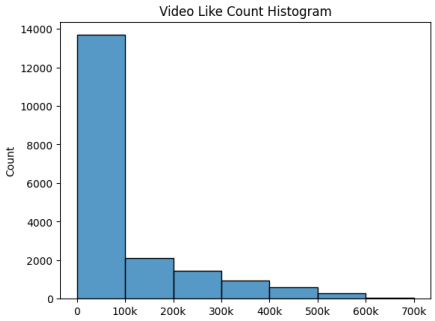
IMPACT

According to the findings from the EDA, the future claim classification model will need to account for null values and imbalance in opinion video counts by incorporating them into the model parameters.

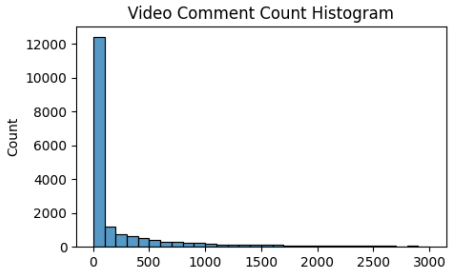
A key component of this project's EDA involves visualising the data. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase TikTok users' (video viewers') engagement with the videos included in this dataset.



The view count variable has a very uneven distribution, with more than half the videos receiving fewer than 100,000 views. Distribution of view counts greater than 100,000 views is uniform.



Similar to view count, there are far more videos with less than 100,000 likes than there are videos with more.



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

KEY INSIGHTS

The EDA conducted from TikTok's data team revealed many considerations for the classification model, including missing values, "claims" to "opinions" balance, and overall distribution of data variables. The two key insights from this analysis were:

Null values

About 300 null values were found in seven different columns. As a result, future modeling should consider the null values to avoid making insights that would assume complete data. Further analysis is necessary to investigate the reason for these null values, and their impact on future statistical analysis or model building.

Skewed data distribution

Video view and like counts are all concentrated on the low end of 1,000 for opinions. Therefore, the data distribution is right-skewed, which will inform the models and model types that will be built.

Executive Summary: Statistical Testing Results

TikTok Claims Classification Project

Project Overview

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, the data team will conduct a hypothesis test to analyse the relationship between verified_status and video_view_count.

Key Insights

- The analysis shows that there is a difference in the number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts.
- As a result, these findings suggest there might be some fundamental behavioural differences between these two groups of accounts: verified and unverified.
- It would be interesting to investigate the root cause of this behavioural 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

verified_status	mean_video_view_count
str	f64
"not verified"	265663.785339
"verified"	91439.164167

The TikTok data team considered the relationship between verified_status and video_view_count.

One approach conducted was to examine the mean values of video_view_count for each group of verified_status in the sample data. The findings showed that unverified accounts have a mean of 265,663 views vs. 91,439 views for verified accounts.

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

Next Steps

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

A regression model for verified_status can help analyse user behavior in this group of verified users. Then, this context can be used to consider results from a claim classification model that will be created afterwards.

Executive Summary: Regression Analysis

TikTok Claims Classification Project

OVERVIEW

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. Earlier, the data team observed that if a user is verified, they are much more likely to post opinions. Since the end goal is to classify claims and opinions, it is important to build a model that shows how to predict the behavior of the account type (verified) that tend to post more opinions. So, in this part of the project, the data team built a logistic regression model that predicts verified_status.

PROJECT STATUS

The variable of verified_status was selected for this regression model because of the relationship seen between the verified account type and the video content. A logistic regression model was selected because of the data type and distribution.

A LOOK AT THE MODEL RESULTS

The logistic regression model achieved a precision of 69% and a recall of 67% (weighted averages). This model achieved an f1 accuracy of 67%. These model results inform key insights on video features, discussed in “key insights.”

NEXT STEPS

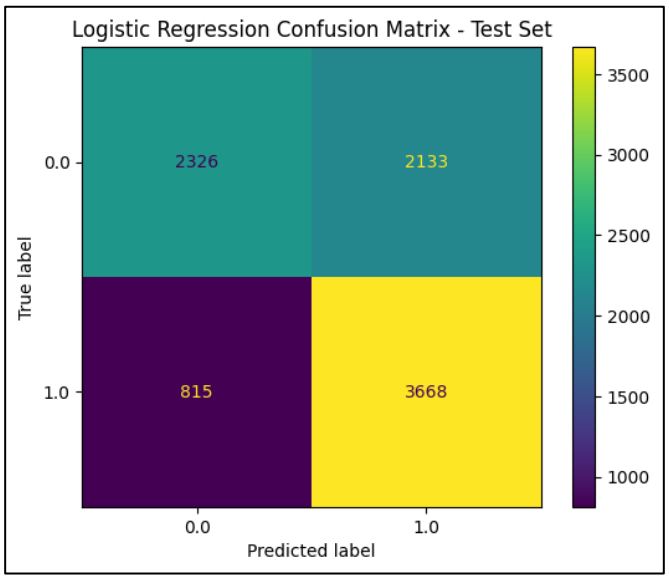
The next step is to construct a classification model that will predict the status of claims made by users. That is the final project and original expectation from the TikTok team. Now, there is enough information to analyse the results of that model with helpful context around user behavior.

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. As a result, other video features besides video length do not seem to be associated with verified status.

Confusion matrix for logistic regression model



Upper-left: the number of videos posted by unverified accounts. Upper-right: the number of videos posted by unverified accounts. Lower-left: the number of videos posted by verified accounts. Lower-right: the number of videos posted by verified accounts.

Machine Learning Model Outcomes

Executive Summary Report for TikTok prepared by the TikTok Data Team

Overview

The TikTok data team seeks to develop a machine learning model to assist in the classification of videos as either claims or opinions. Previous investigation into the available data revealed that video engagement levels were highly indicative of claim status. The team is confident that the resulting model will meet all performance requirements.

Problem

TikTok videos receive a large number of user reports for many different reasons. Not all reported videos can undergo review by a human moderator. Videos that make claims (as opposed to opinions) are much more likely to contain content that violates the platform's terms of service. TikTok seeks a way to identify videos that make claims to prioritize them for review.

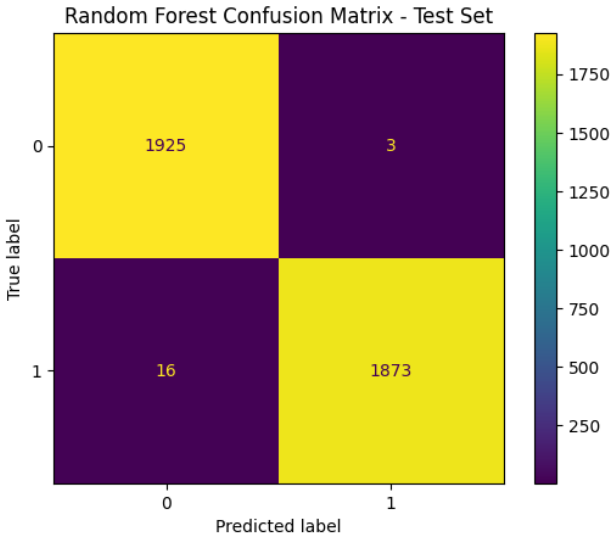
Solution

The data team built two tree-based classification models. Both models were used to predict on a held-out 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 the random forest (RF) and XGBoost models performed exceptionally well. The RF model had a better recall score (0.990) and was selected as champion. Performance on the test holdout data yielded very good scores, with 19 misclassified samples out of 3,817.

Subsequent analysis indicated that, as expected, the primary predictors were all related to video engagement levels, with video view count, like count, share count, and download count accounting for nearly all predictive signal in the data. With these results, we can conclude that videos with higher user engagement levels were much more likely to be claims. In fact, no opinion video had more than 10,000 views.



Next Steps

As noted, the model performed very well on the test holdout data. Before deploying the model, the data team recommends further evaluation using additional subsets of user data. Furthermore, the data team recommends monitoring the distributions of video engagement levels to ensure that the model remains robust to fluctuations in its most predictive features.