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.

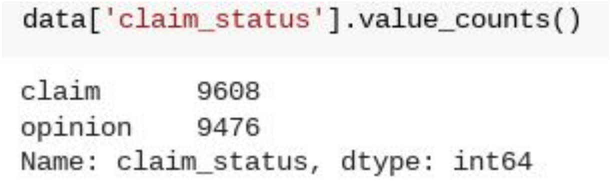
Given the ask for a classification of user claims, the data team looked at the counts of claims and opinions in order to understand the count of each type of video content.

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

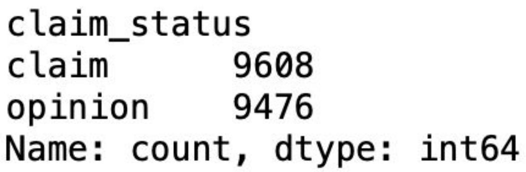
Opinions:

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.

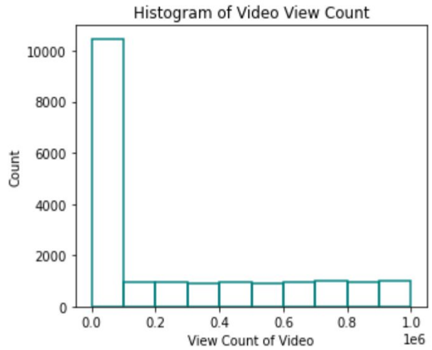
RESPONSE

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.

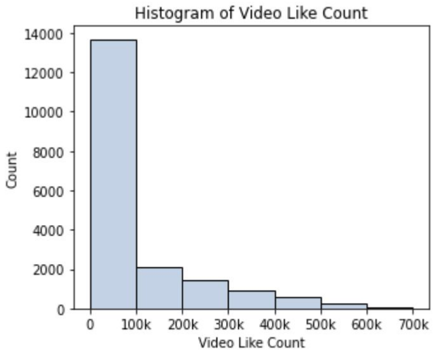
IMPACT

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.

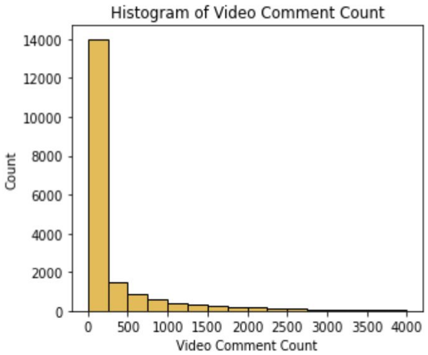
A essential component of this project's EDA is visualizing the data. Following histograms illustrates that vast majority of videos are grouped at the bottom of value range for three variables of interest (video view count, video like count, video comment count – related to user engagement).



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.