TikTok: Claims Classification Project

Preparing Data for Predictive Model Development

Project Overview

Our objective is to support TikTok's effort in reducing the backlog of reported videos by preparing the claims classification dataset for analysis. This work is foundational for building a predictive model that will help distinguish between content that contains claims and content that offers opinions.

Key Insights

- The dataset contains 19,382 rows and 12 columns, with 298 missing values.
- Of the 19,382 records, 9,608 (just under 50%) are **claims** and 9,476 (just over 49%) are **opinions**, with 298 missing claim status entries.
- There are 298 **missing values** in critical fields such as claim_status and engagement metrics, which must be handled before analysis.
- Claim videos consistently outperform opinion videos in engagement per view (***likes***, ***comments***, ***shares***), especially when posted by **banned** or **under-review** authors, who also generate the highest absolute engagement.
- Descriptive Statistics reveal notable **outliers** in video engagement, suggesting a skewed distribution that may require transformation or filtering during analysis.
- Video_duration_sec and video_view_count were identified as **key predictors** for the claim classification model

Details

- Variables like video_view_count, video_like_count and video_share_count, display extremely high maximum values compared to their means and medians.
- Banned authors, despite being a minority, show the **highest average engagement per view**, indicating their content may be more attention-grabbing or controversial.
- The structure of the dataset and the consistency of the column types confirm that the data is prepared for cleaning and EDA.

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
```

Next Steps

- Address **missing values** in key columns and validate how to handle them.
- Investigate and possibly transform **outliers** in engagement metrics to avoid skewed model predictions.
- Conduct **EDA** to explore relationships between claim status and engagement.
- Begin **feature engineering** for predictive modelling using variables like video_duration, author_ban_status, and per-view engagement metrics.