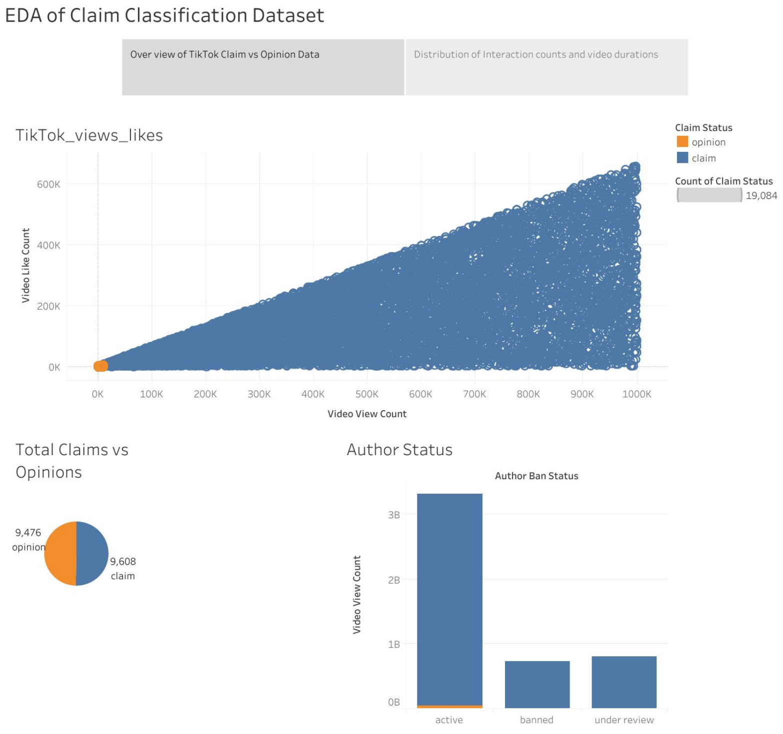
# TikTok Claims Classification

# Problem + Data Overview

**<u>Problem:</u>** With the number of submissions and interactions on TikTok each day, it’s challenging to manually review each video, user comment, and content claim. Thus, the goal of this project is to develop a predictive model that can determine whether a video contains a claim or offers an opinion.

**<u>Dataset:</u>** The dataset contain ~19000 records of reported TikTok videos and whether they are express a Claim or an Opinion. Included in the dataset is information regarding user interactions (e.g. likes, shares, etc). Some notable points regarding the data:

- About 300 records with Missing data
- Almost perfectly balanced between claims and opinions

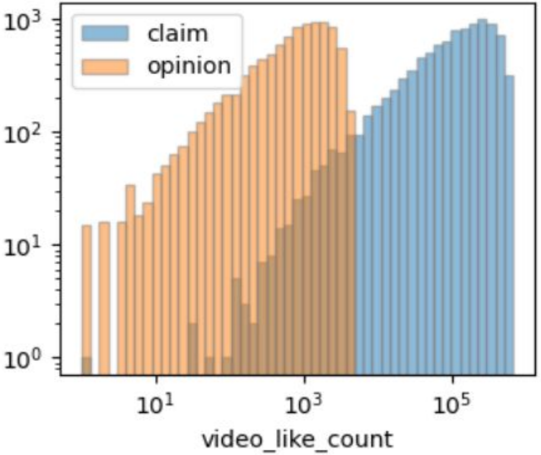


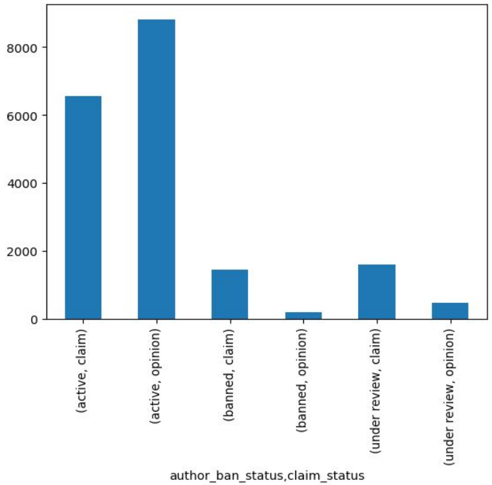

**[Tableau dashboard Available](#)**

# EDA Summary

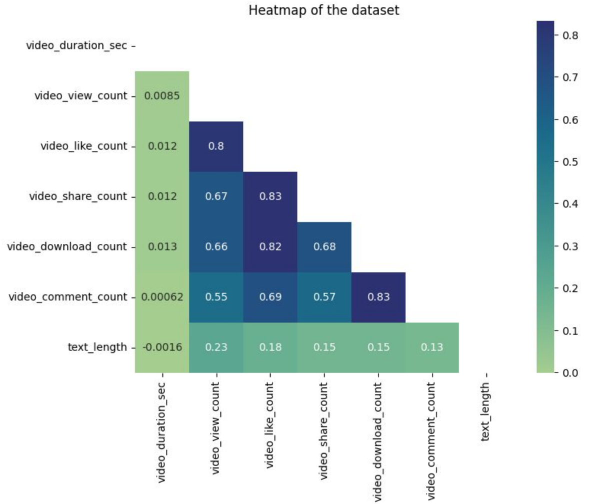
Clear difference in distribution of Claim/Opinion for each of the user interaction features (ex on left).

Also, author ban status show a trend as well (banned/under_review authors post more claims and vice versa)




Cross-correlation matrix shows clearly correlation between user interaction parameters (video likes, shares, views, downloads)



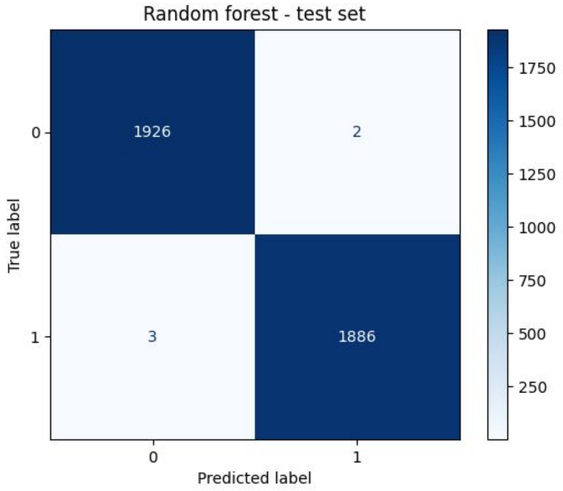
# Model Approach

**Metric of interest:** While the goal is to minimize the manual labor required in the review process with automation, we want to ensure that we are careful to review as many claims as possible. Therefore, we are less concerned with false positives and a suitable primary metric of interest is the **Recall.**

**Model Type:** For this classification task we will use decision tree ensemble methods. In particular we test  a Random forest as well as an XGBoost model.

# Results

Bothe the Random forest and XGBoost models perform extremely well, with recall at ~99% and ~100% for the XGBoost and RF models respectively. Since the RF performed slightly better, the **Random Forest** will be our final model.



# Final thoughts

Our models perform extremely well, with little/no feature engineering. However, this is very likely due to the artificial nature of the dataset, and application to real world data would be much different

The most predictive features are the interaction type features, especially view count. This is in line with the insights from the EDA

In the case that we wanted to improve the dataset, having more information about the author could improve our understanding of likely claim status for videos. Other information such as number of reports for the video would halo as well.