TikTok Classification Project

Ting-Chun Adam Wang

Executive Summary

- **Objective:** This project developed a predictive model to determine whether a TikTok video contains a claim or expresses an opinion.
- **Data:** A dataset of 19,084 TikTok videos was analyzed, focusing on attributes such as view count, transcription text length, and more.
- **Statistical Test:** The analysis revealed significant differences in engagement metrics (e.g., view count, like count, and others) between claim and opinion videos.

Executive Summary

- **Model:** A tuned Random Forest model was constructed to predict video type, achieving a 99.2% recall score and a 99.6% accuracy score on the test data.

- **Key Feature:** View count emerged as the most predictive attribute in determining whether a video contains claim or opinion content.

- **Application:** This model can assist the content moderation team in more efficiently identifying and mitigating misinformation on the TikTok platform.

Introduction

Objective of the project

Claim videos are more likely to contain content that violates the platform’s terms of service.

We aim to develop a predictive model to determine whether a video contains a claim or an opinion.

Claims
Information from an unverified source.

“It is said that drinking coffee every day can reduce the risk of heart disease by 40%.”

Opinions
Personal beliefs or thoughts of a group or an individual.

“I believe that reading fiction is the best way to develop empathy.”

Introduction

- Dataset from Kaggle
- Analyze Using Python
- Information of 19,048 TikTok Videos

Exploratory Data Analysis (EDA)

Numpy, Pandas, Matplotlib, Seaborn

Statistical Test

Scipy

Machine Learning Model Selection

Sklearn

Introduction

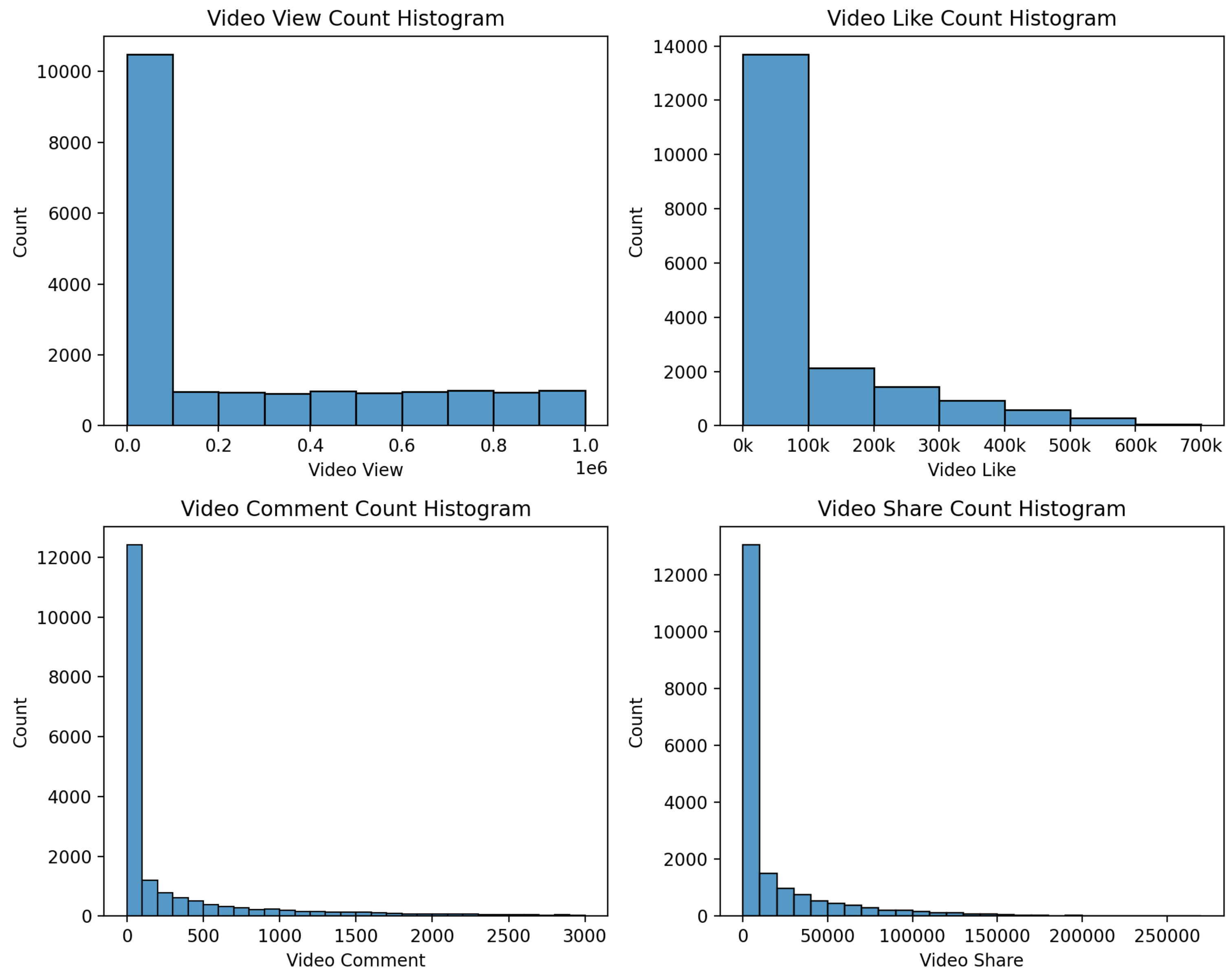
- Dataset from Kaggle
- Analyze Using Python
- Information of 19,048 TikTok Videos

Claim Status	Video id	Video Duration	Transcription Text	Verified Status	Author status	View count	Like Count	Share Count	Download Count	Comment Count
claim	7017666017	59	someone shared with me that drone deliveries a...	not verified	under review	343296	19425	241	1.0	0.0

Exploratory Data Analysis

The distribution of views, likes, comments, and shares is uneven across videos.

Most of the videos have lower numbers of views, likes, comments, and shares.”

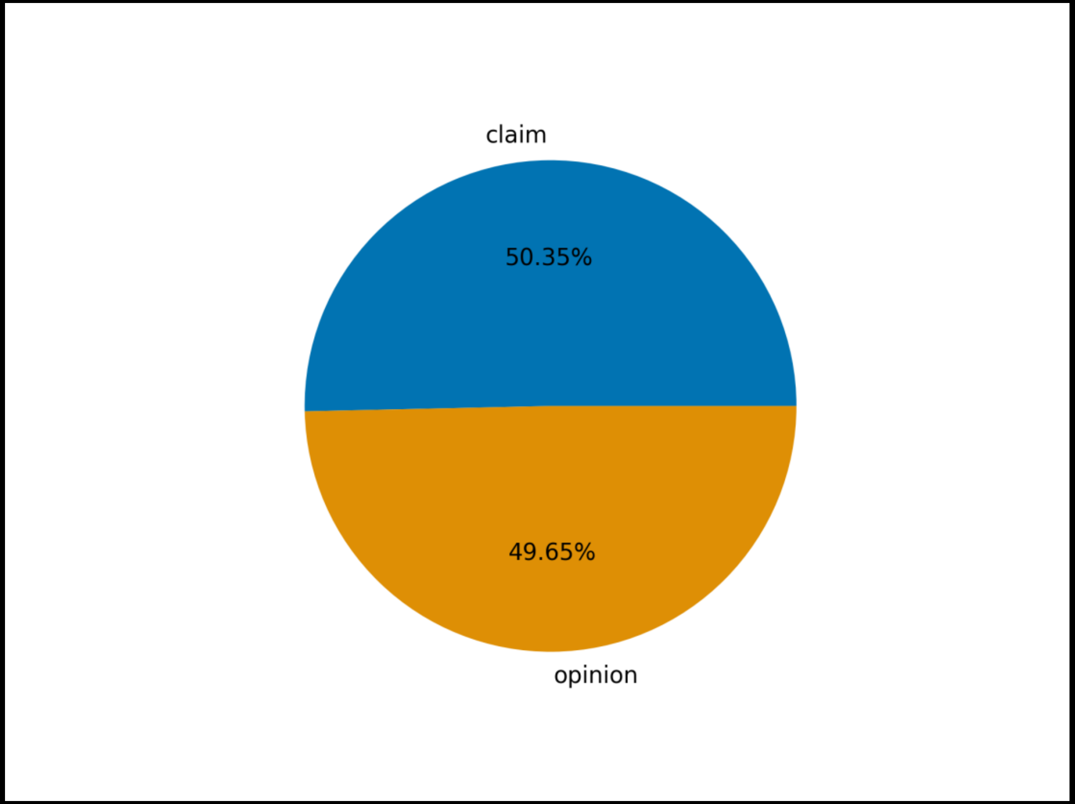


Exploratory Data Analysis

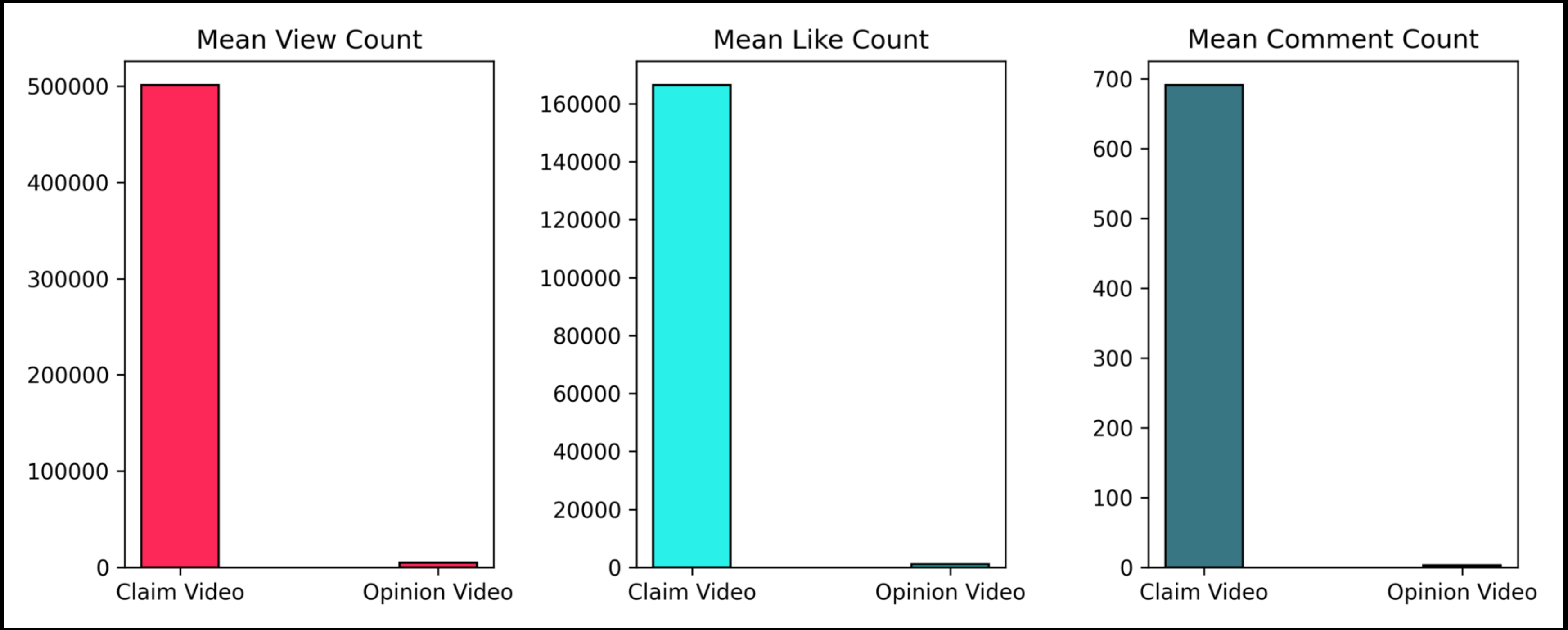
The proportions of both types of videos are similar.

Claim: 9,608

Opinion: 9,476



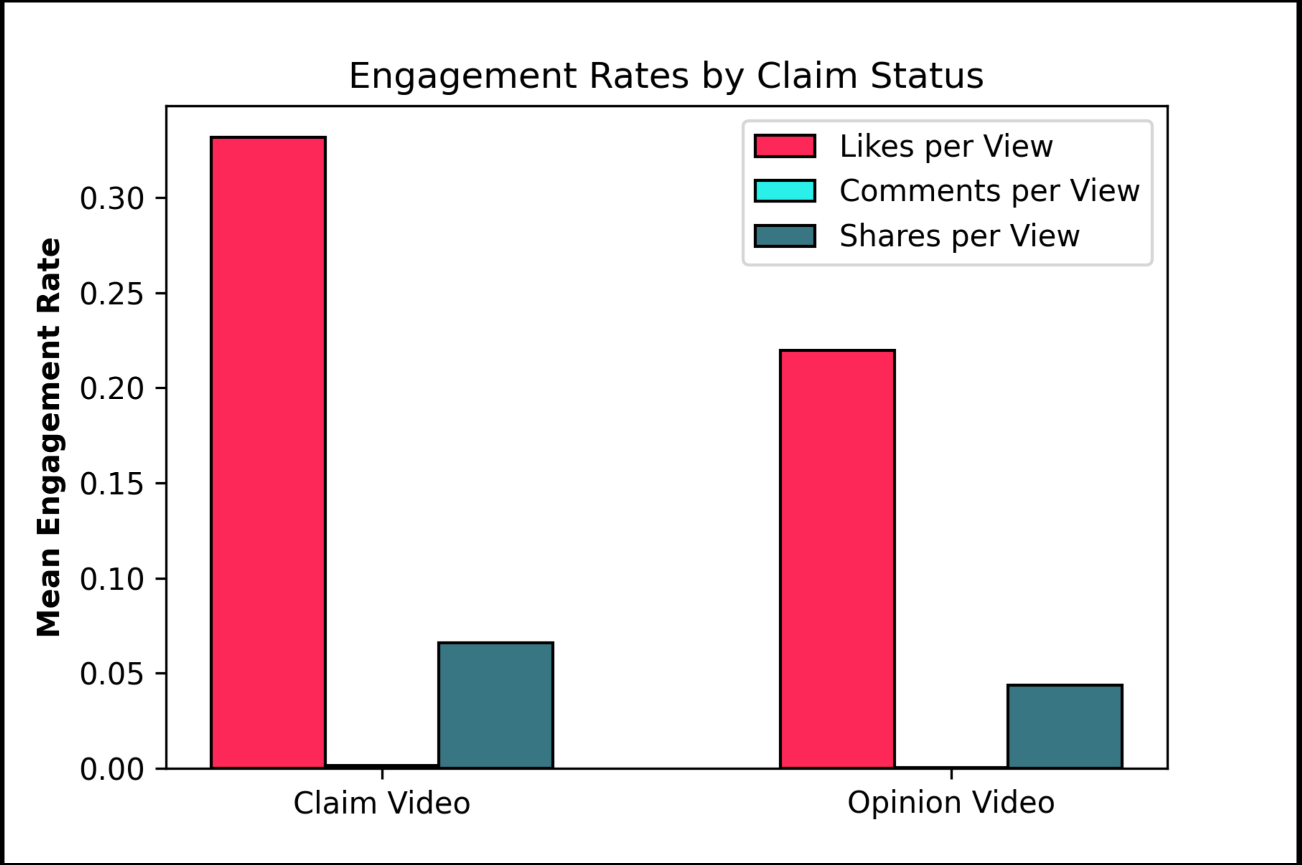
Exploratory Data Analysis



Comparison of engagement metrics:

Claim videos have higher view, like, and comment counts.

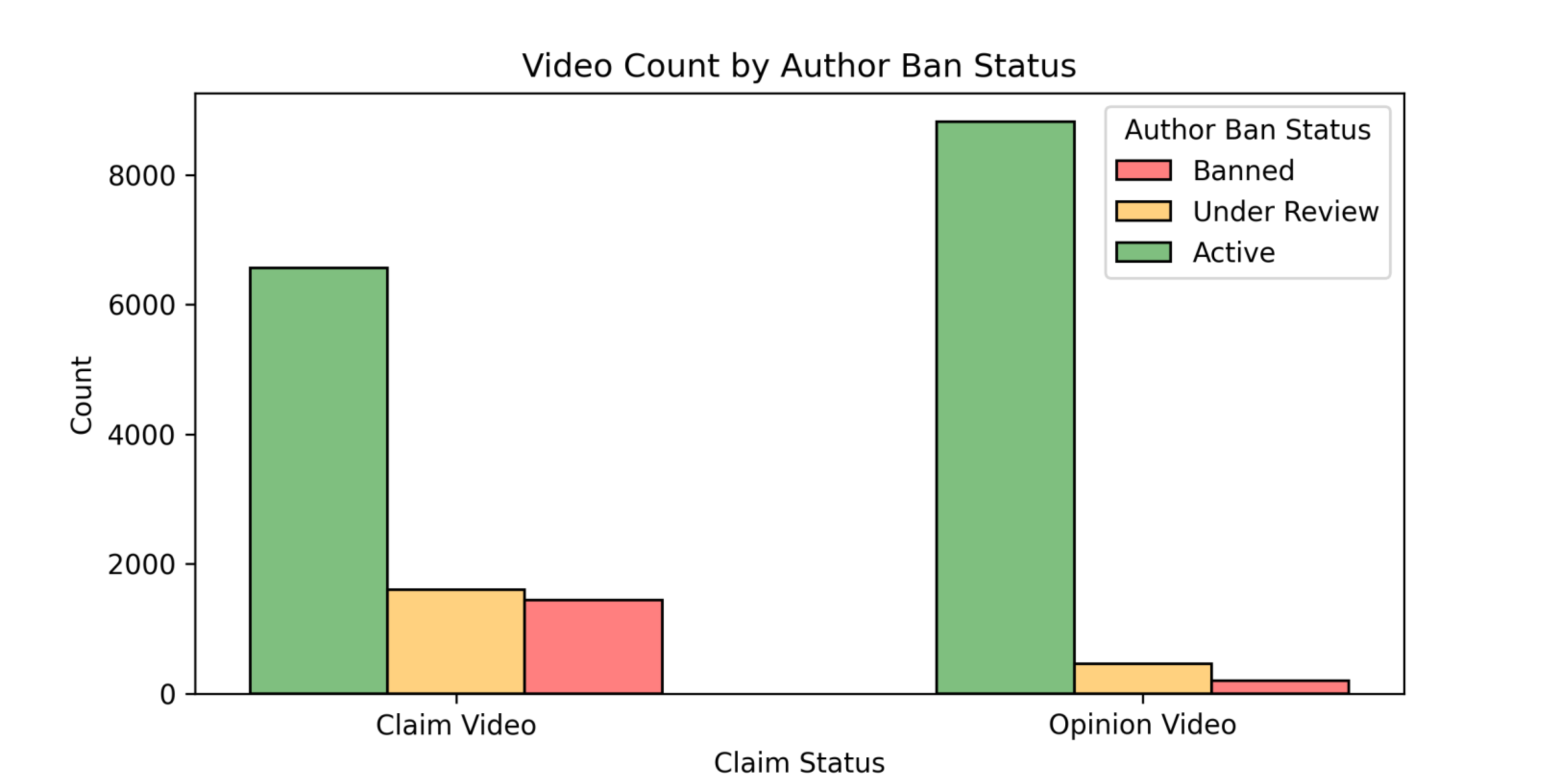
Hypothesis Test



Claim videos have higher engagement rates,

with more likes, shares, and comments per view.

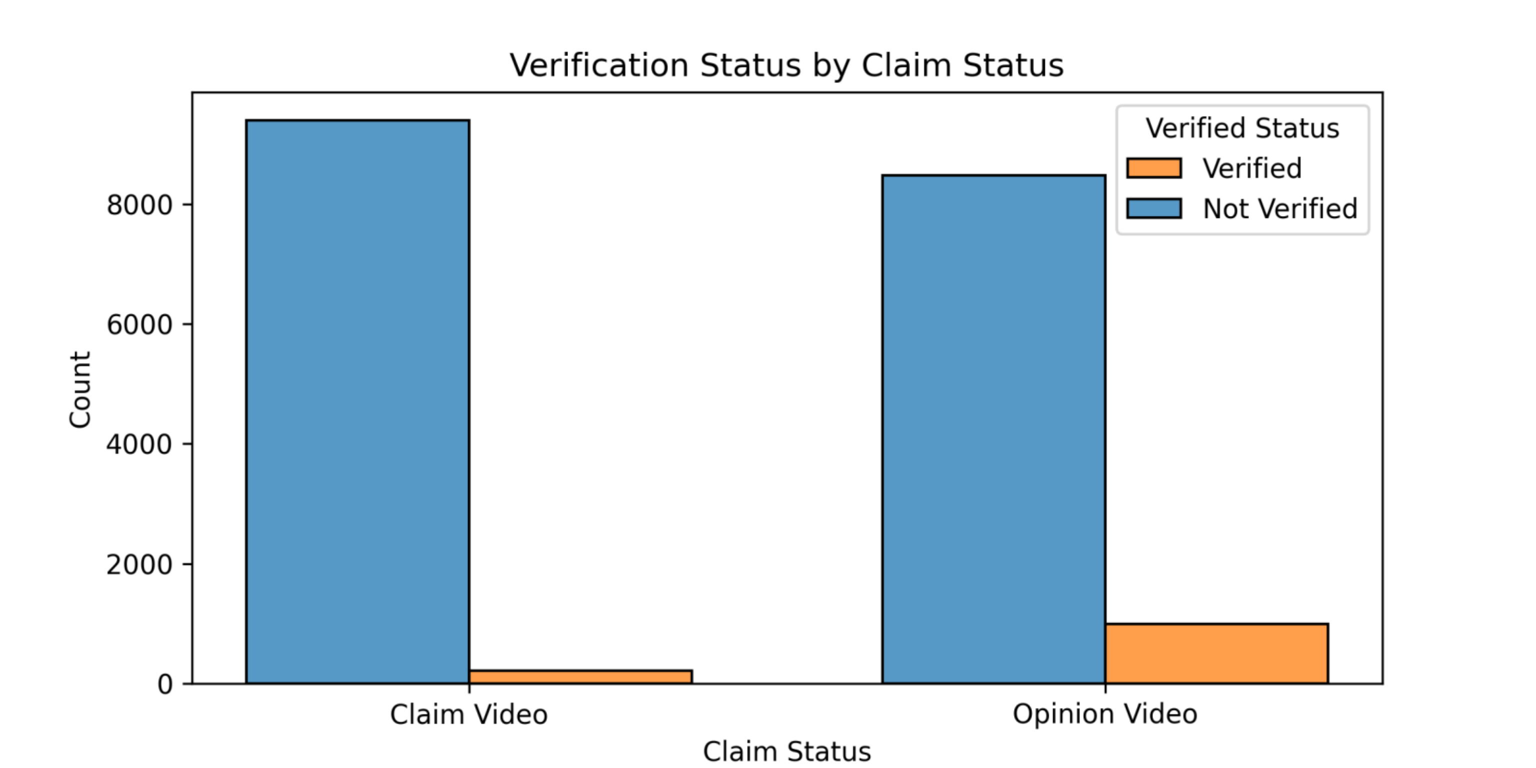
Exploratory Data Analysis



There are fewer active authors for claim videos.

However, claim videos outnumber opinion videos when it comes to both banned and under-review authors.

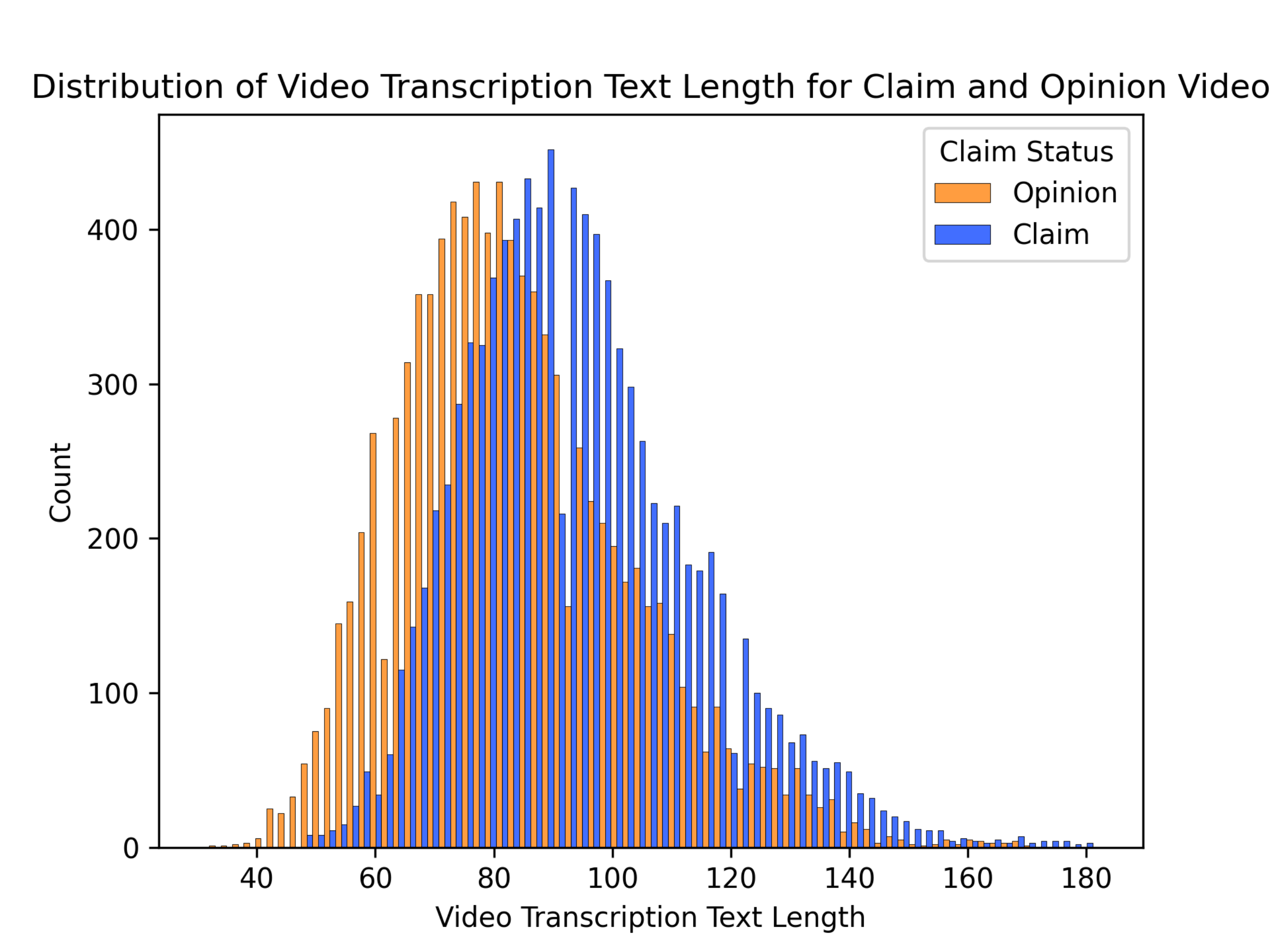
Exploratory Data Analysis



Regardless of claim status, most videos are not verified.

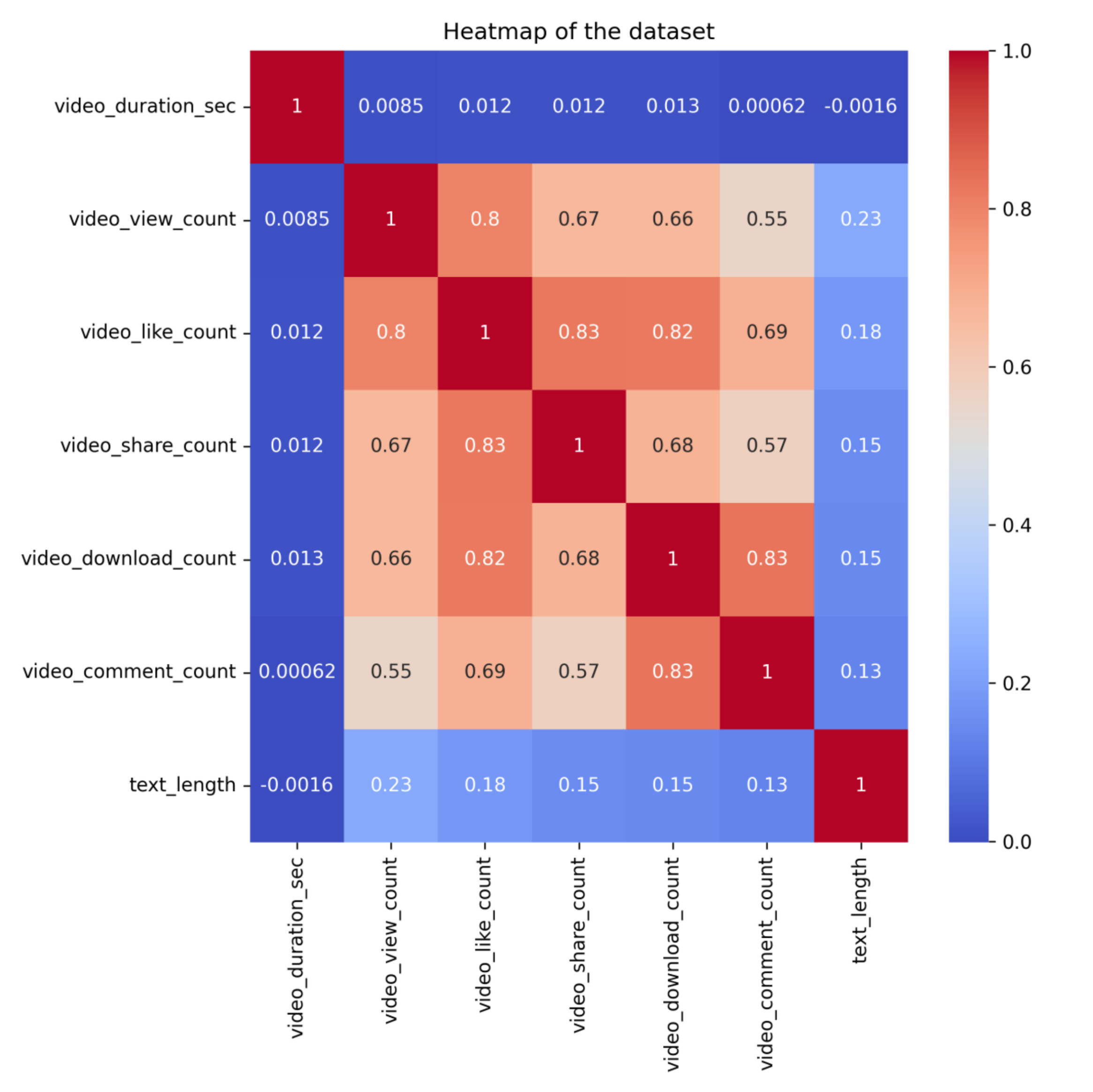
However, if a video is verified, it is more likely to be an opinion video.

Exploratory Data Analysis



Claim videos tend to have longer transcriptions compared to opinion videos.

Exploratory Data Analysis



Correlation among variables:

Engagement metrics, such as view count, like count, and share count, are highly correlated.

Statistical Test

Two Sample T-test

Claim vs. Opinion

Test	T statistic	P-Value
Duration Time	0.540	0.588729
View Count	166.889	<.001
Like Count	109.742	<.001
Comment Count	66.341	<.001
Share Count	82.923	<.001
Text Length	44.385	<.001

Machine Learning

Random

Forest

Model

XGBoost

Model

Test set (60%):

Train,

tune hyperparameters

Validation set (20%):

Determine Champion Model

Test set (20%):

Obtain Final Result:

Machine Learning

Test set (60%)

Random Forest

Hyperparameter	Best Recall Score
Max depth	5, 7, None
Max features	0.3, 0.6
Max samples	0.7
Min samples leaf	1, 2
Min samples split	2, 3
Number of estimators	75, 100, 200

XGBoost

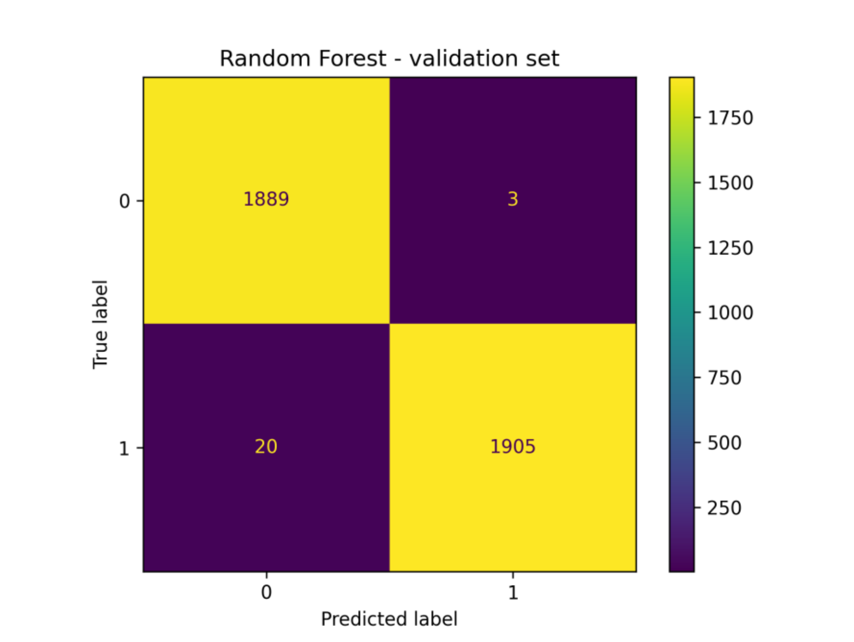
Hyperparameter	Best Recall Score
Max depth	4, 8, 12
Min child weight	3, 5
Learning rate	0.01, 0.1
Number of estimators	300, 500

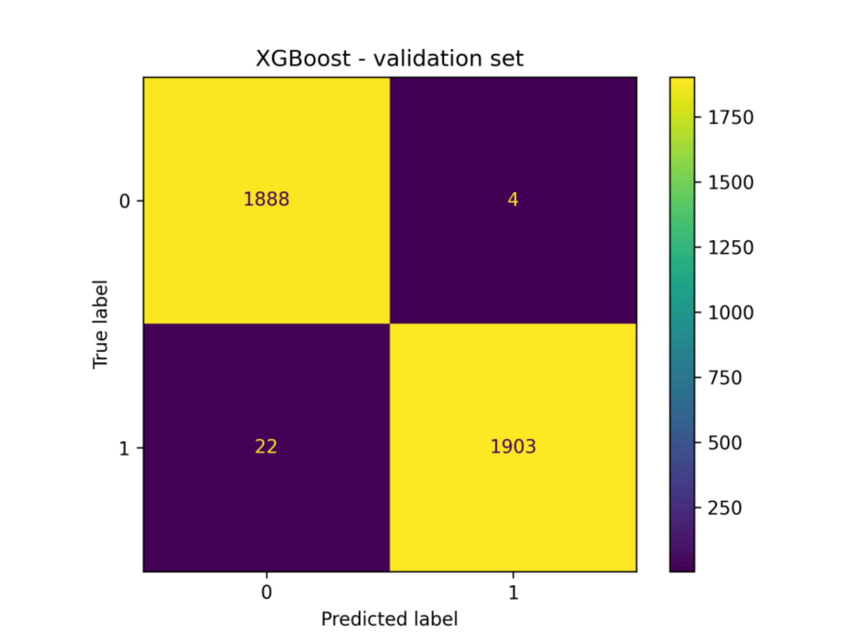
Machine Learning

Random Forest

XGBoost

Validation set (20%)





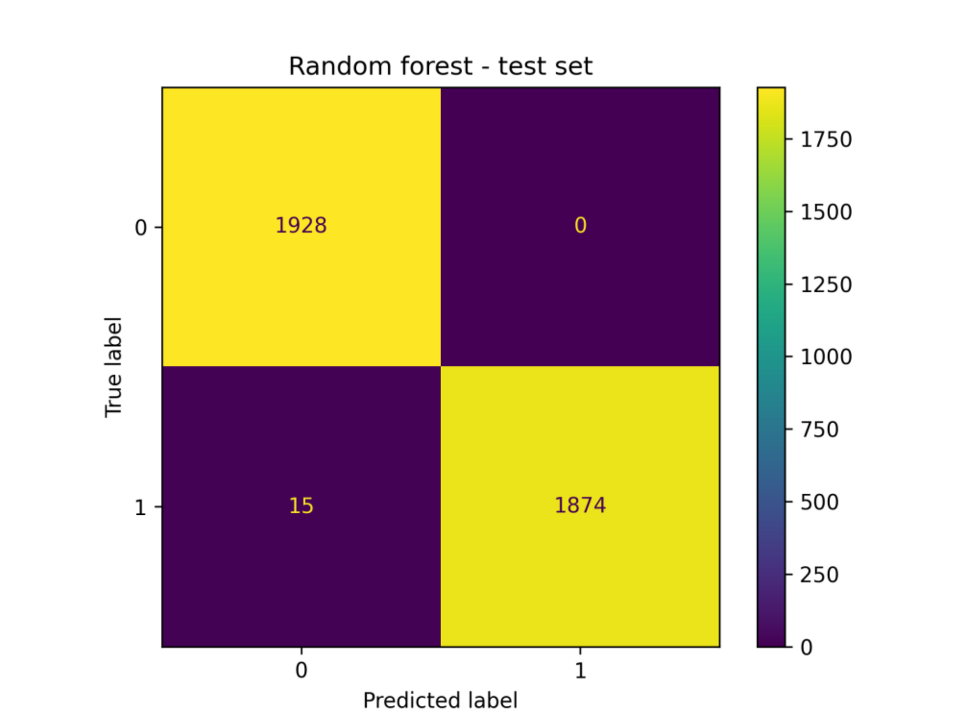
Recall: 0.9896

Recall: 0.9886

Machine Learning

Random
Forest

Test set
(20%)



Recall score: 99.2%

Accuracy score: 99.6%

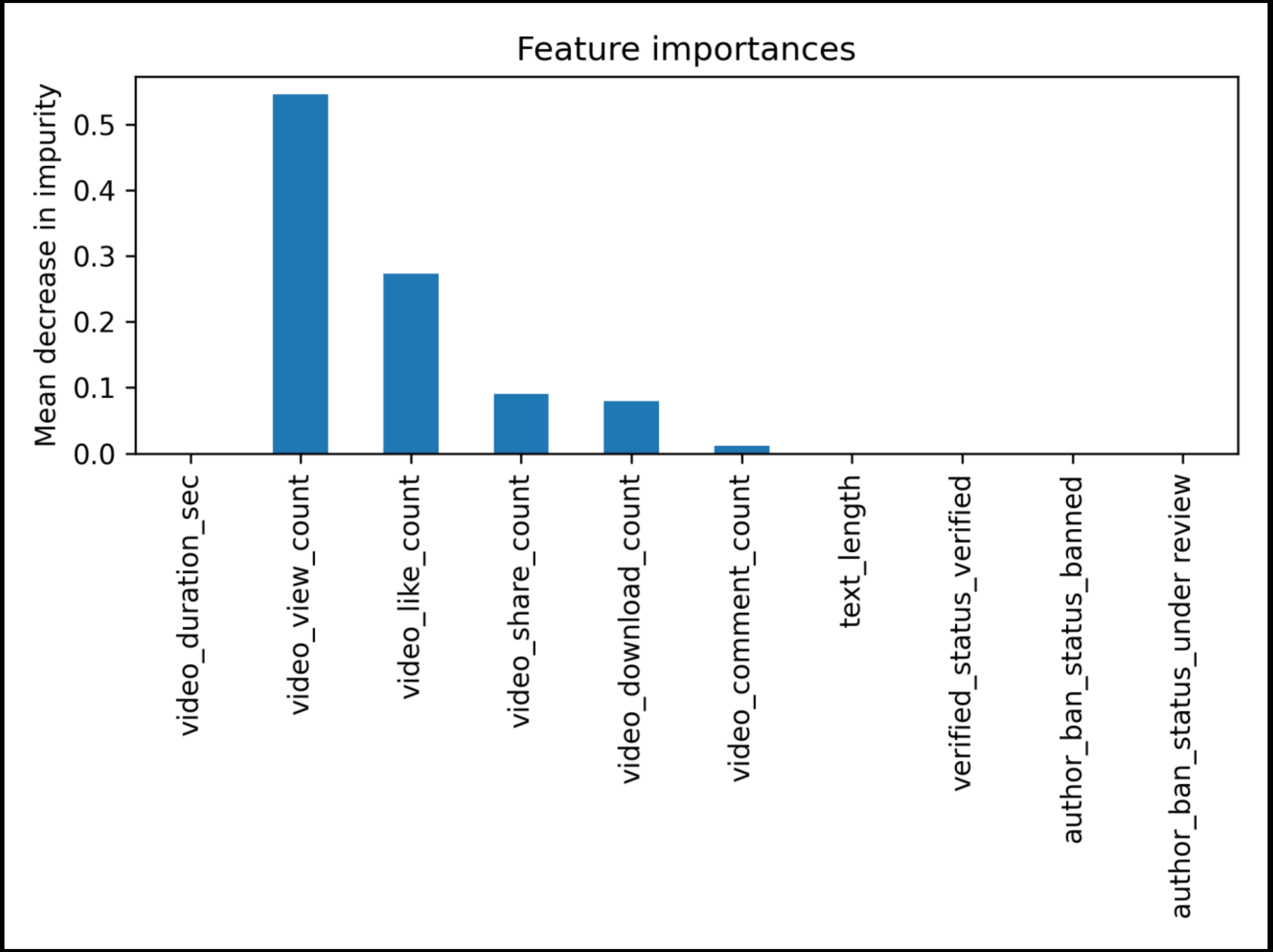
Machine Learning

Random Forest

Test set (20%)

The most predictive features is view count.

This suggests that engagement metrics are effective predictors of a video's claim status.



Conclusion

Selection Criteria

- Random Forest model chosen based on superior recall score

Test Set Performance

- Recall score: 99.2%
- Accuracy score: 99.6%

Impact

- Highly accurate predictions for video claim status
- Recommend increasing moderation efforts on high-view-count videos
- Significant improvement in content moderation workflow efficiency