Exploratory Data Analysis of TikTok Videos

Overview

The TikTok user can report videos they understand to violate the platform's service agreements, but this results in more videos than what a human moderator can review. A claims classification model will be made to determine what features a video will be to be a claim or an opinion.

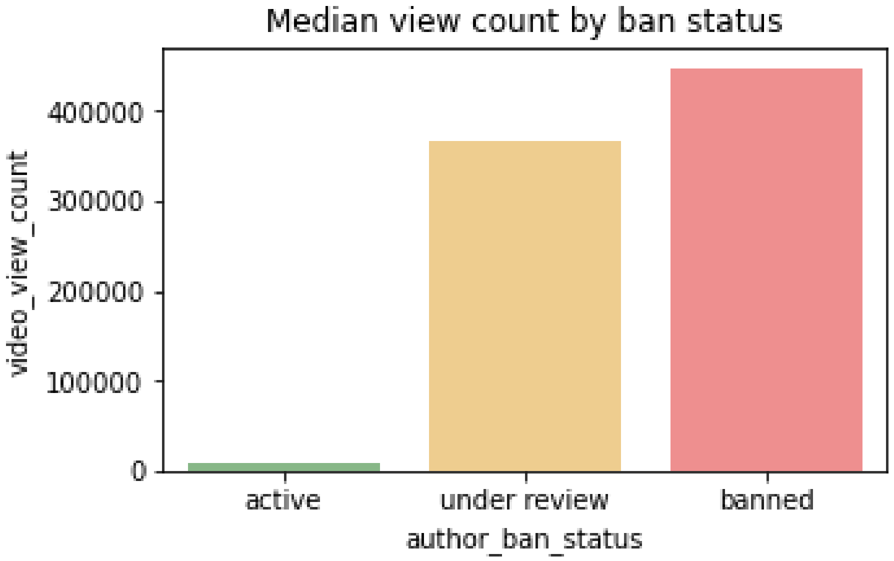
Business Understanding

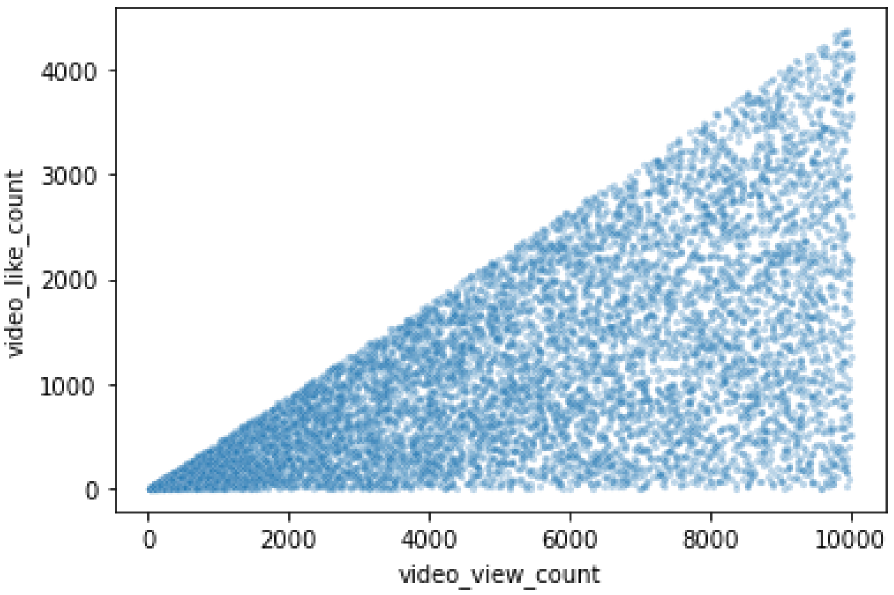
To reduce the time a human review would need to check whether a video is a claim video from an opinion video, thus saving the company time and resources.

Data Understanding

The data for this analysis shows 9,608 are claims out of a total of 19,382 (50%)

- Engagement level is strongly correlated to claim status and further examination is recommended.
- In levels of engagement from most to least: Videos from banned authors, then next is videos from authors under review, and last are videos with active authors.

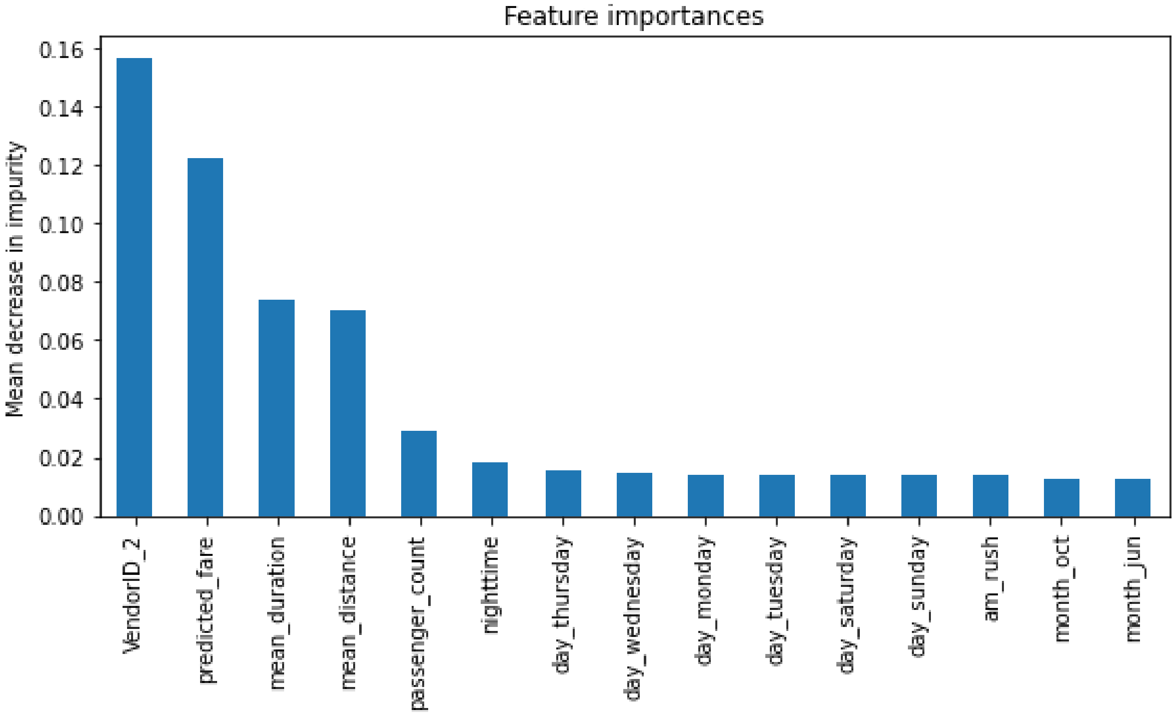


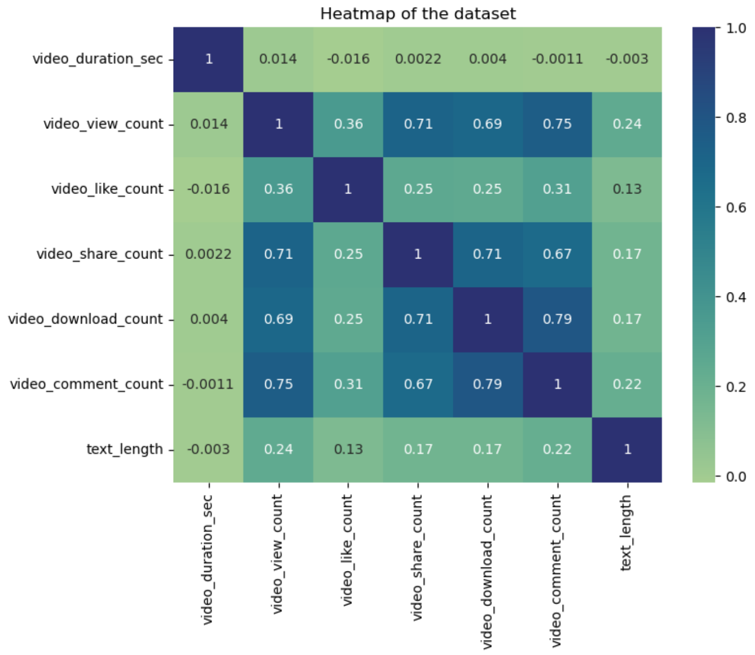


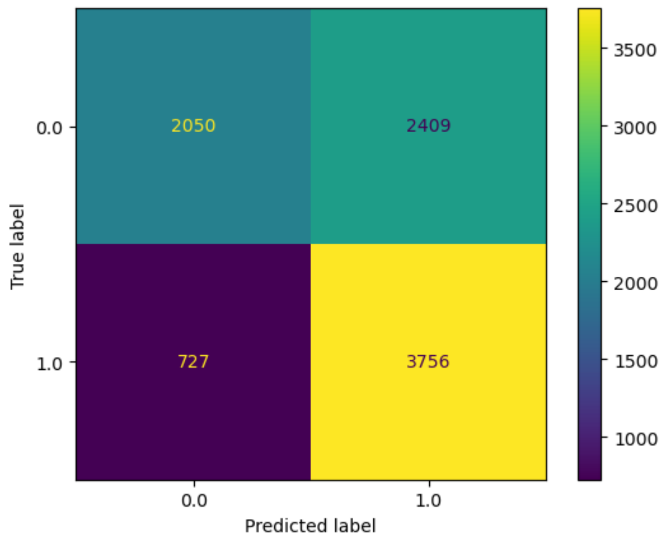
Modeling and Evaluation

The hypothesis for this data project is: The null hypothesis is that there is no statistical difference in view numbers between verified and unverified users. The alternative hypothesis is that there is a statistical difference in view numbers between verified and unverified users. There is a 5% significance level and the test was a two-sample t test. The p-value is 2.6088823687177823e-120, which is lower than the significance level of 5%. Thus, this test indicates to reject the null hypothesis, which means that there is a statistical difference between the mean video view counts between verified users and unverified users.

The test shows that there is a statistically significant difference in view counts between verified and unverified accounts. Further analysis on why this would be is recommended, such as do unverified accounts post videos that are more incendiary that compell viewers to click on them? Or do unverified accounts made by a lot of bots that superficially inflate view counts?


Feature importances
Mean decrease in impurity
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
VendorID_2
predicted_fare
mean_duration
mean_distance
passenger_count
nighttime
day_thursday
day_wednesday
day_monday
day_tuesday
day_saturday
day_sunday
am_rush
month_oct
month_jun





Conclusion

The key takeaways are:

- Since the dataset has a few very strongly correlated variables that could lead to multicollinearity in a logistic regression model, it was decided to remove video_like_count from the model.
- The logistic regression model shows that with each second added of the video is related to an 0.009 increase in the odds that the user has verified status.
- The logistic regression model was not the best but still was in the acceptable parameters of prediction. The precision was 61% (less than ideal), but the recall was 84% (overall, good). The overall accuracy is at the lower end of the acceptable range.
- The result is that the developed logistic regression model to determine the verified status based video features with decent predictive power. From the estimated model coefficients from the logistic regression, the longer a video is, the higher the likelihood that the user is verified.