Executive Summary

Machine Learning Model Evaluation

ISSUE / PROBLEM

In this part of the project we were tasked to create a well performing random forest ML model that would classify tiktoks as either claims or opinions. We also note that our preliminary analysis revealed that video_view_count is highly correlated with claim_status.

RESPONSE

We have created two tree based models, a random forest and an XGBoost model, and since we have 19,382 data points we have enough to hold out a validation set.

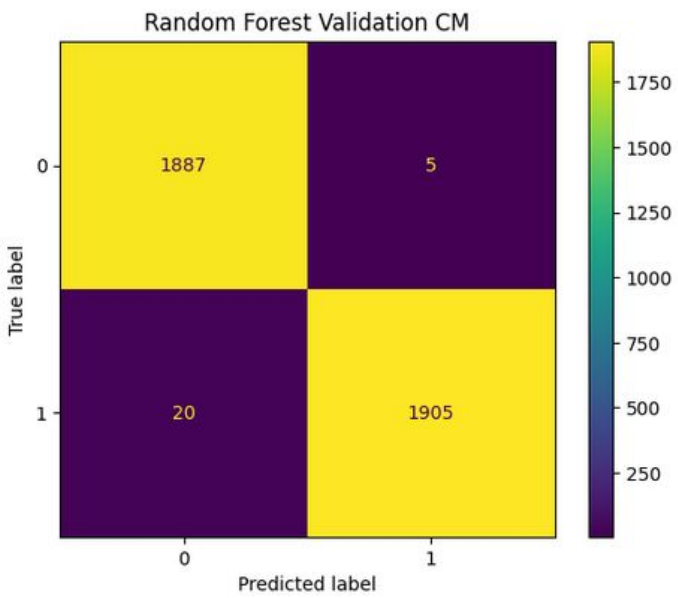
We used a 60/20/20 split of out training/validation/test data, then used the recall metric to determine which performed best. We then used out champion model on test data to evaluate its expected future performance.

IMPACT

The purpose of this model is to increase the efficiency of the process we use to handle problematic content. TikTok has an enormous number of reported tiktoks per day, and so it is essential that the review process is streamlined to minimize the number of employees it takes to review problematic content.

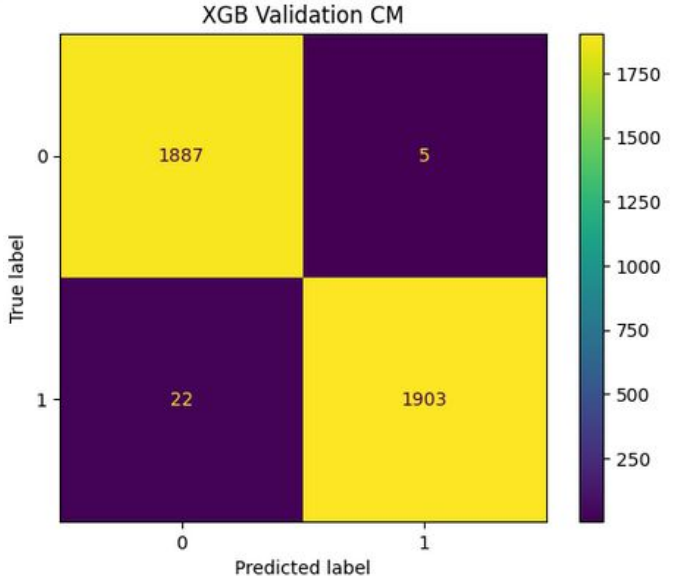
```
print(rf_results_df['mean_test_precision'][rf_cv.best_index_])
print(rf_results_df['mean_test_recall'][rf_cv.best_index_])
```

```
0.9994785483051682
0.9908534395531852
```




```
print(xgb_results_df['mean_test_precision'][xgb_cv.best_index_])
print(xgb_results_df['mean_test_recall'][xgb_cv.best_index_])
```

```
0.9989540877965151
0.9898176171763818
```



KEY INSIGHTS

Although the models both performed exceptionally, and with very similar results, we have chosen to use recall as the metric for evaluation because we deem it more important to ensure we classify all claims as such than to incorrectly classify an opinion as a claim. As we expected from our previous investigations, engagement levels, view/like count, were the best predictors of claim status.