Predictive Model Building & Outcomes

Executive Summary

Overview

Overall, the TikTok data team is working to reduce the backlog of user reports and prioritize them more efficiently. To do so, they are wanting to create a model that can predict if videos are claims or opinions. In this stage of the project, the data team is to create and evaluate the model.

Problem

TikTok has a large backlog of user reports that they need help parsing through. They are hoping to identify which videos are 'claims' so that they can be sent to a human for further review.

Solution

The data has built a tree-based classification model in order to predict whether videos are claims or opinions. This model was decided as the final model after evaluating two models to find the one with the best recall score.

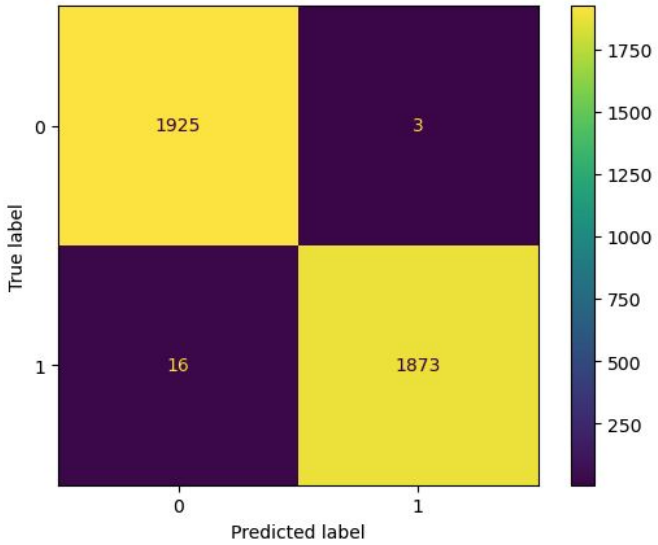
Details

Two models were evaluated - one random forest and one XGBoost. Both models performed incredibly well with 0.99+ recall scores and high precision. Ultimately the random forest model was selected as the final model due to its slightly higher scores.

As shown in the image to the right, the final model only misclassified 19 samples, for a total of .5% misclassified.

We also saw in this analysis that the most important predictors of claim status were related to engagement, as discussed in an earlier analysis. Video view, like, share, and download count all had predictive signals.

Confusion matrix for the final model on test data. Only 19 misclassifications occurred out of 3,817 total samples.



Next Steps

As discussed above, the model is an exceptional predictor of claim status. However, the data team would recommend running the model on different subsets of data to ensure its validity.