ML Model Report - Predictive Modeling of TikTok Video Classification

Executive Summary for TikTok

Overview

The business objective is to build a machine learning model to help classify video claims vs opinions. Claims are more likely to be in violation of TikTok’s terms of service. Providing these labels can streamline the human content review process, by presenting human moderators with only the highest priority videos that are most likely to be in violation of TikTok's terms of service.

Problem

Due to the volume of content on the platform, many videos get reported, but there are too many for a human moderator to examine and consider in a timely fashion. A previous analysis of videos that were in violation showed that those in violation are more likely to be videos in which authors are making claims as opposed to sharing opinions. Additionally, high engagement levels are correlated with claim videos.

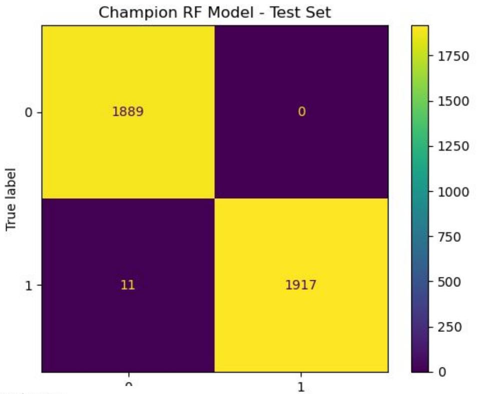
Solution

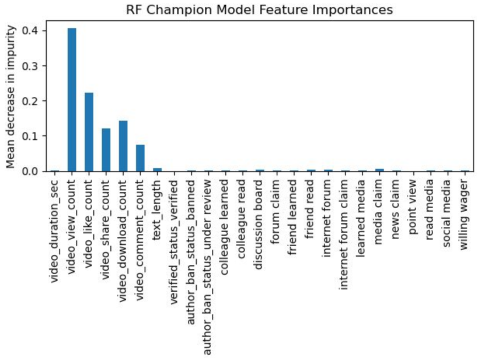
Using tree based classification models, the data team was able to deliver exceptional results with identifying claim videos, and ultimately selected the Random Forest model which had the best recall score.

Details

Both of the tuned Random Forest and XGBoost models performed well in terms of precision, accuracy, recall, and F1 scores. However, the Random Forest model did a better job of minimizing false negative predictions. There were only 11 misclassified videos out of 3,817. Given the nature of the business objective, it was imperative to make sure that likely violations were prioritized for human review. The RF model was nearly perfect for the task.

The feature importances obtained from the best model confirmed our findings during exploratory data analysis, which suggested that video engagement through likes, views, shares, and downloads were strongly related to videos that were claims. These features ended up being the most predictive signals in the classification model with the best performance.





Next Steps

Before deploying the model, additional evaluation on user data is recommended. While in the use, the distribution of the model’s most predictive signals should be monitored to make sure the model still works well considering fluctuations in levels engagement.