Machine Learning Model Outcomes

Executive summary report for TikTok prepared by the TikTok data team

Overview

Researchers on the TikTok data team aim to build a machine learning algorithm that can classify videos based on whether they express claims or opinions. They've already discovered that video interaction metrics can be a strong indicator of whether a claim is being made or not, and are optimistic that their new model will meet expectations in terms of performance.

Problem

TikTok faces a high volume of user reports on videos for various reasons, but not all can be reviewed by a human moderator. Evidence suggests that videos making claims are significantly likelier to contain content that breaches the platform's guidelines. To improve content moderation, TikTok seeks a way to automatically identify videos making claims.

Solution

The data team developed two tree-based classification models. These models were tested on a separate validation dataset, and the model with the highest recall score was selected. This chosen model was subsequently applied to a test dataset to predict future performance.

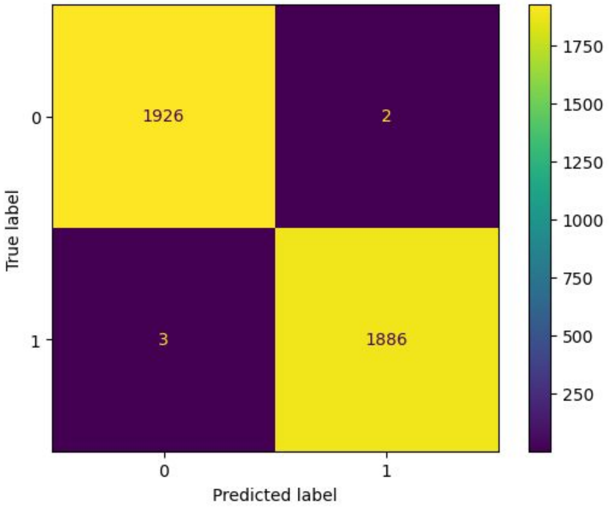
Details

The two model architectures used, random forest (RF) and XGBoost, demonstrated outstanding performance. The RF model had a higher recall score of 0.995 and was chosen as the superior model.

When evaluated on the held-out test data, the model achieved near-perfect scores, misclassifying only 5 out of 3,817 samples.

A closer look at the data confirmed that engagement metrics played a crucial role as primary predictors. Notably, view count, likes, shares, and downloads were responsible for nearly all the predictive power in the data. This suggests a strong correlation between high user engagement and claim videos, with the findings indicating that opinion videos rarely achieved view counts over 10,000.

Confusion matrix for the champion RF model on test holdout data shows only five misclassified samples out of 3,817.



Next Steps

As previously mentioned, the model showed outstanding performance on the test holdout data. However, before deploying it, the data team advises further evaluation using additional subsets of user data. Additionally, the team suggests monitoring the distributions of video engagement levels to ensure the model remains robust against variations in its most predictive features.