Date: January 16th 2025

OPTIMIZING CONTENT MODERATION WITH MACHINE LEARNING

Final Report Prepared by TikTok Data Team

Project Overview

The TikTok data team developed a machine learning model to classify flagged videos as either claims or opinions. Claims are more likely to violate platform guidelines, and prioritizing their review helps reduce moderator backlogs. Analysis revealed that engagement metrics, such as video views, likes, and shares, were the strongest predictors of claim status. The selected model meets all performance requirements.

Problem

TikTok receives a high volume of user reports, far exceeding what human moderators can review. Videos making claims, as opposed to opinions, are more likely to contain harmful or misleading content. TikTok needs an efficient way to prioritize flagged videos that require immediate attention.

Solution

The team tested two tree-based models, Random Forest and XGBoost, and selected Random Forest as the champion model due to its superior recall. High recall ensures that most claims are accurately identified, minimizing the risk of harmful or misleading content being overlooked. By prioritizing flagged claims, the model streamlines TikTok's moderation process, reducing backlogs and allowing moderators to focus on high-impact content.

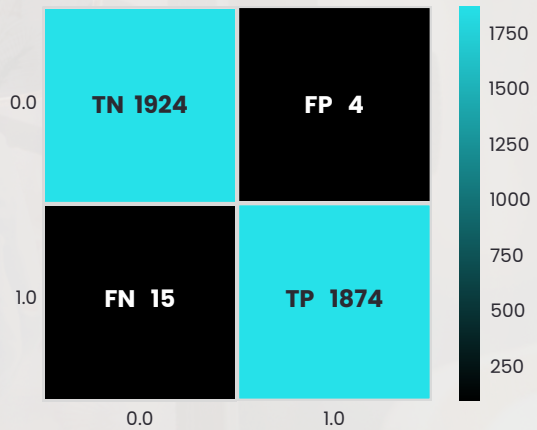
Details

Developing the champion model involved a rigorous process of testing and refinement. The team used cross-validation to optimize model settings, ensuring reliability and strong generalization to new data. Engagement metrics, such as video views, likes, and shares, emerged as the strongest predictors of claim status, guiding feature selection and refinement.

The final Random Forest model achieved:

- Accuracy: **99.50%**
- Recall (Claims): **99.21%**
- Precision: **99.79%**
- F1 Score: **99.50%**

Champion Model Confusion Matrix



Confusion matrix shows strong performance with minimal misclassifications

Key Insights

Error Analysis

The model performed exceptionally well, with **15 false negatives** and **4 false positives** out of **3,817 samples** in the test set. This demonstrates its ability to strike a balance between recall (minimizing missed claims) and precision (avoiding unnecessary flags), aligning perfectly with the goal of prioritizing critical content for moderation.

Feature Importance

The model's predictions are driven by user engagement metrics, such as **video view count, likes, shares, and downloads**. These features were the strongest indicators for distinguishing claims from opinions. In contrast, verification status and author ban status had minimal impact, challenging initial expectations.