Claims Classification Project | Milestone Six

Executive summary prepared for TikTok leadership by the TikTok data team

ISSUE / PROBLEM

TikTok leadership asked the TikTok data team to develop a ML model that will help us better understand user submissions and reduce the backlog of user reports. In the last part of the project, the data team needs to create a machine learning model that predicts whether a video is a claim or an opinion.

RESPONSE

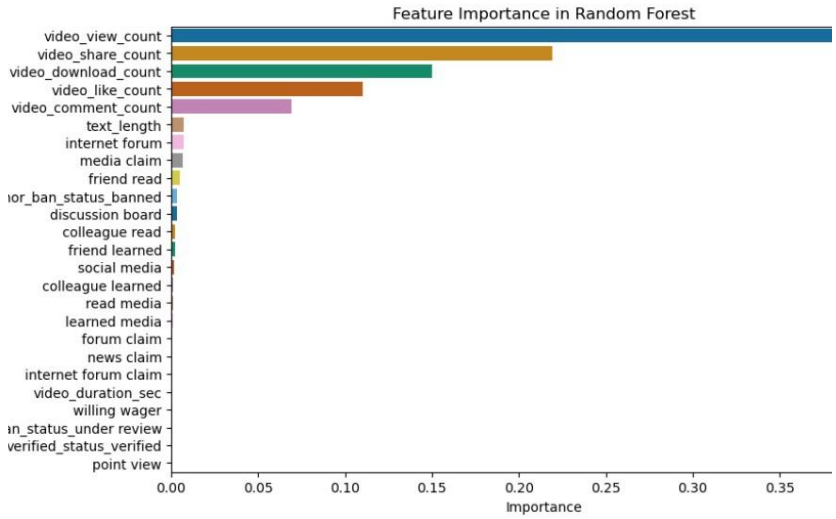
The data team chose to develop two different machine learning models (i.e., random forest and XGBoost) to cross-compare results and obtain the model with the highest predictive power (i.e., random forest).

Seeing the confusion matrix, the model correctly predicts true negatives and true positives; yielding almost all true negatives and true positives and almost no false negatives or false positives.

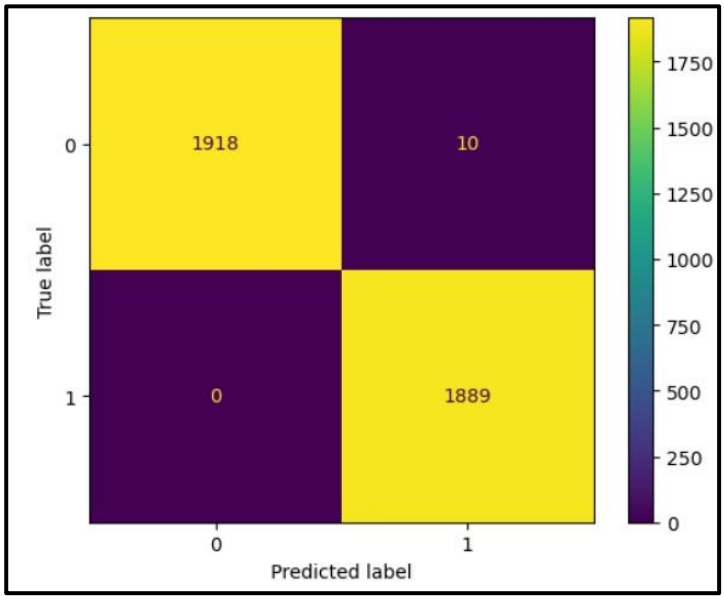
IMPACT

Developing tree-based models since Tree-based models are robust to outliers or values that deviate significantly from the norm.

The machine learning model provides an exceptional framework for predicting binary outcomes indicating whether a video is a claim or an opinion.



The bar plot above shows the relative feature importance of the predictor variables in our machine learning model.



The upper-left quadrant displays the number of opinions that the model accurately classified as so. The upper-right quadrant displays the number of opinions that the model misclassified as claims. The lower-left quadrant displays the number of claims that the model misclassified as opinions. The lower-right quadrant displays the number of claims that the model accurately classified as so.

KEY INSIGHTS

- **According to the bar plot of the relative importance of the features in the model,** the most predictive features were all related to engagement levels generated by the video. Both "video view count" and "video share count" accounted for the top two important features in the model.
- **Splitting the data, fitting models and tuning hyperparameters on the training set, and performing final model selection on the validation set,** we see from the recall , precision, and f1 scores of both the Tuned Random Forest and Tuned XGBoost that both performed exceptionally well. Yet, the Random Forest model fit the data better than XGBoost model with f1-score of 1.
- **The data team recommends sending this machine learning classification model to our operations team to predict the status of claims made by users** since the model successfully classified claims as claims and opinions as so.