Machine Learning Model Outcomes for TikTok Data

Executive Summary Report for TikTok

Overview

The data team at TikTok is nearing the end of the claims classification project. The final milestone left for the team: creating the machine learning model. The tasks for creating the model includes feature engineering, model development, and evaluation.

Objective

- Import relevant packages and TikTok data
- Exploratory data analysis
- Feature engineering
- Check model assumptions
- Model building
- Model evaluation

Results

- For the random forest model:
  - Best hyperparameters: 'max_depth': 50, 'max_features': 'sqrt', 'max_samples': 0.9, 'min_samples_leaf': 1, 'min_samples_split': 0.001, 'n_estimators': 50
  - Metric scores from validation: F1 = 0.994331, Recall = 0.98959, Precision = 0.999128, Accuracy = 0.994323
- For the XGBoost model:
  - Best hyperparameters: 'colsample_bytree': 1, 'learning_rate': 0.35, 'max_depth': 4, 'min_child_weight': 1, 'n_estimators': 35, 'subsample': 1
  - Metric scores from validation: F1 = 0.994242, Recall = 0.98907, Precision = 0.999474, Accuracy = 0.994236
- The random forest model was chosen as the champion model
- The most important features in the random forest model are video_view_count and video_like_count.

Next Steps

- Utilize the random forest model to predict the claim status of a reported video.
- Produce the production model with the best hyperparameters for random forest model using the entire dataset.