Logistic Regression Model Results for Claim Classification Data

Executive Summary for TikTok

Overview

The next step in the claims classification project is to use the project data to create a regression model. Determine the type of regression model that is needed and develop one using TikTok's claim classification data.

Objective

- Determine the correct modeling approach
- Build a regression model
- Finish checking model assumptions
- Evaluate the model

Results

- The model produced from using 'verified_status' as the outcome variable had acceptable, but not great, predictive power. Precision, recall, and accuracy values range between 46% to 84%.
- To improve the predictive power, another version of logistic regression model was produced using the following updated conditions:
  - The dataset was split for training and testing (75%-25%) before removing rows with missing values.
    - Remove rows with missing values from the training data first before the testing data.
  - The variable 'claim_status' was used as the outcome variable.
    - Encode the values of 'claim_status' with OrdinalEncoder()
  - Only one feature variable was chosen, 'video_view_count'
    - Replace negative values with 0 before using log transform on each value added by 1.
- The resulting precision, recall, and accuracy values for the 'claim_status' values range between 98% to 100%.

Next Steps

- Consider possible alternate features to be used in the logistic regression model.
- Adjust training/testing proportions if necessary.