# Claims Classification Project (Milestone 5)

Regression Analysis – Prepared by PhD Aleksandar Osmanli

## OVERVIEW

I seek to develop a machine learning model to assist in the classification of claims for user submissions. Earlier, I observed that if a user is verified, he is much more likely to post opinions. Since the end goal is to classify claims and opinions, it's important to build a model that shows how to predict the behavior of the account type (verified) that tend to post more opinions. So, in this part of the project, I built a logistic regression model that predicts "verified_status".

## PROJECT STATUS

The variable of "verified_status" was selected for this regression model because of the relationship seen between the verified account type and the video content. A logistic regression model was selected because of the categorical data type and distribution.

A LOOK AT THE MODEL RESULTS

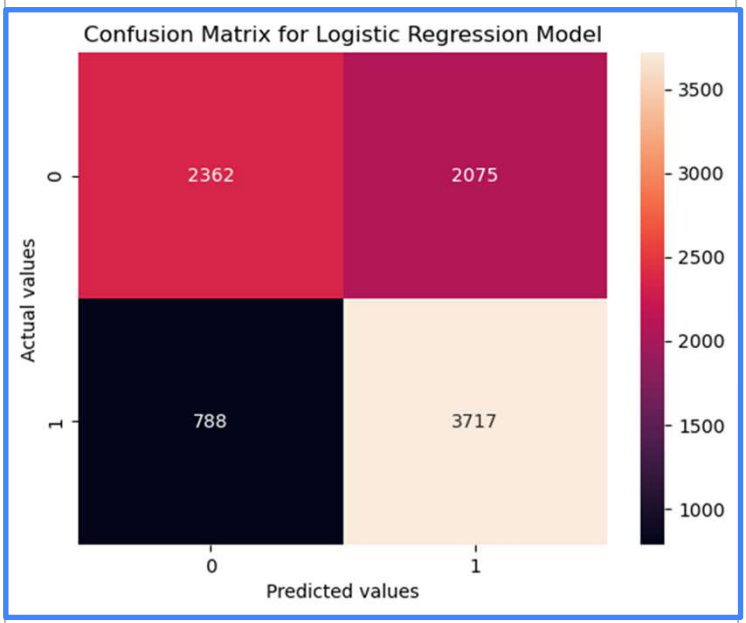
The logistic regression model achieved a precision of 70% and a recall of 68% (weighted averages). This model achieved an f1 accuracy of 67%. These model results inform key insights on video features, discussed in "key insights."

## NEXT STEPS

The next step is to construct a classification model that will predict the status of claims made by users. That is the final project and original expectation from the client. Now, there is enough information to analyze the results of that model with helpful context around user behavior.

## KEY INSIGHTS

Based on the estimated model coefficients from the logistic regression, videos with more shares and comments tend to be associated with higher odds of the user being verified. Longer videos tend to be associated with smaller odds of the user being verified.
Other video features have small estimated coefficients in the model, so their association with verified status seems to be small.




*Upper row: the number of videos posted by unverified accounts. Lower row: the number of videos posted by verified accounts.*