PROJECT MILESTONE 5

PACE: Plan Stage

- What am I trying to solve or accomplish?

How to predict verified status to help client understands how video characteristics relate to verified users.

- What are my initial observations when I explore the data?

There were 298 missed values in the 'claim_status', 'video_transcription_text', and in all 5 count variables.
After removing the missed values, approximately 93.7% of the dataset represents videos posted by unverified accounts and 6.3% represents videos posted by verified accounts. So, the outcome variable is highly unbalanced.

PACE: Analyze Stage

- What are some purposes of EDA before constructing a multiple logistic regression model?

Some of the purposes of the EDA are analyzing and discovering data from the dataset, looking for correlations, missing data, potential outliers, and/or duplicates.

PACE: Construct Stage

- Do I notice anything odd?

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."

- Can I improve it? Is there anything I would change about the model?

The model could be improved by engineering new features to try to generate better predictive signal.
It could also be helpful to reconstruct the model with different combinations of predictor variables to reduce noise from unpredictive features.

PACE: Execute Stage

- What key insights emerged from the model?

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.

- What business/organizational recommendations would I propose based on the model built?

I would recommend 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.