TikTok Claims Classification Project

Logistic regression model - Executive Summary

ISSUE / PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, I build the model.

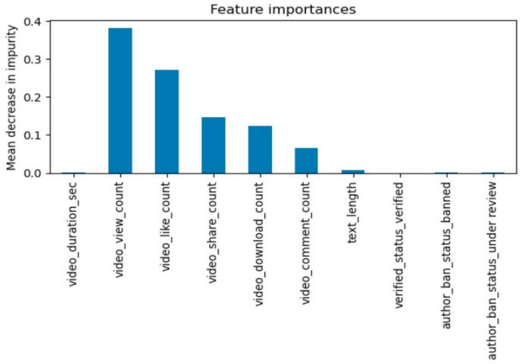
RESPONSE

The Tik Tok data team decided to build a random forest and XGboost models. With the purpose of comparing which model predicts better the claim status of the videos. In order to accomplish this, I will conduct both models, taking claim status as the target variable, then compare which model performed better, and finally pick a winner model.

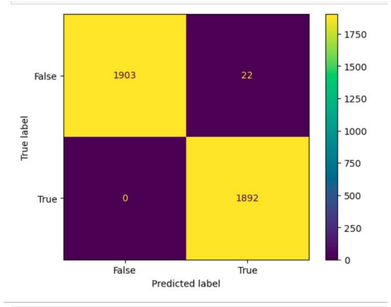
IMPACT

During the creation of the models, I created a new column about description length. We discovered that the claim video used to have slightly larger descriptions. Both models performed well, with a score over 99% because of the lower number of n_estimators we used in the XGBoost model (whose time of fitting is more demanding). I chose the random forest model as the winner.

As keys component of this project I made a histogram to compared descriptions length between claim status, we conduct 3 confusion matrices to evaluate models' performances. In this case both models performed well. Finally we plot the importances of the features for the winner model.

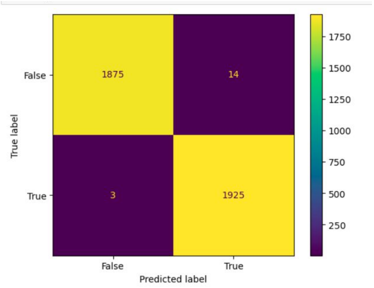


As we can see in this visual, views and likes counts are the two variables with more weight in the winner model. Ban and verified status lack of weight.



As we can see, the XGBoost model performed well. However, this model was built using a low number of n_estimators due to its demanding fitting time. That gives this model a certain risk of overfitting.

As we can see, the random forest model performed well; this model was chosen as the winner. It's got a low number of false positives and negatives, and it has precision, recall, and F1 scores over or equal to 0.99.



KEY INSIGHTS

As part of the winner model, I did a classification report of it. Here are the results:

Opinion

- The model had a presion of 100% in it prediction.
- The model had a recall of 99% in it prediction.
- The F1 score scored good as consequence. It had a score of 99%

Claim

- The model had a presion of 99% in it prediction.
- The model had a recall of 100% in it prediction.
- The F1 score scored good as consequence. It had a score of 99%