TikTok Machine Learning Project

Fessler 2024

Classification Model Construction

ISSUE / PROBLEM

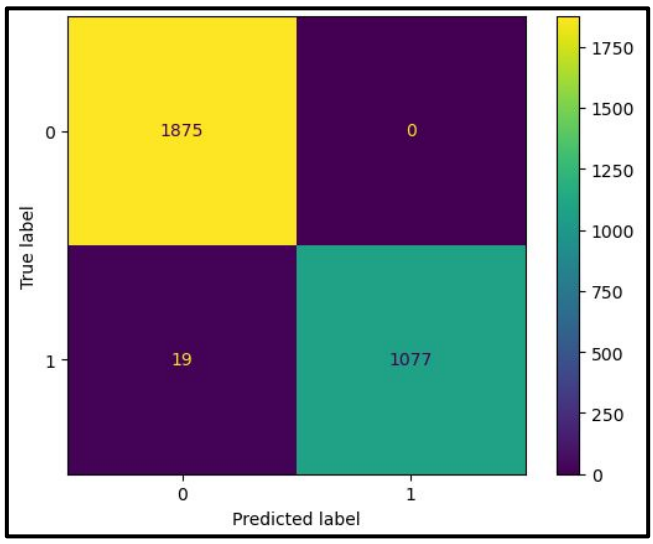
With many videos being flagged as against our terms of service every day, our moderators are in need of a way to speed the sorting of these submissions as either an 'opinion' or a 'claim'. This is currently where an immense amount of moderators' time is spent, since each video must be reviewed.

RESPONSE

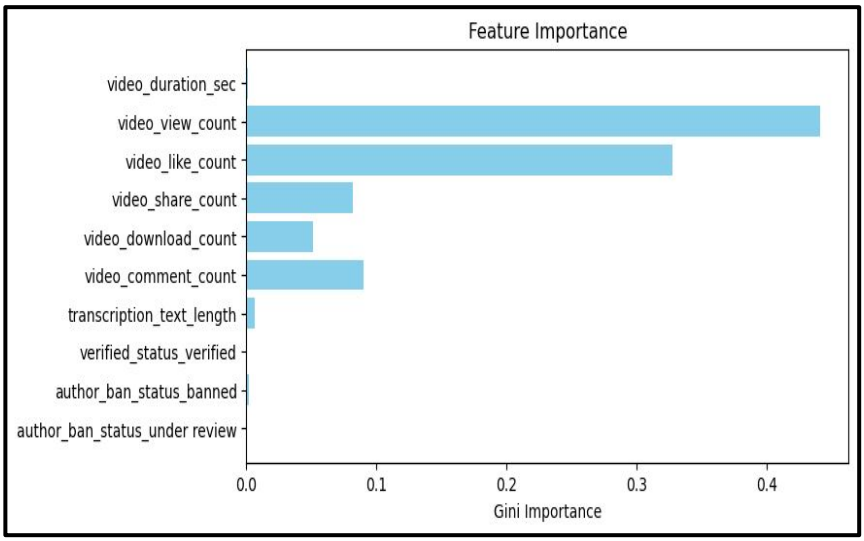
Our goal is to construct a classification machine learning model to address this issue. It will determine based off of the metadata of video submission whether the video contains an 'opinion' or a 'claim'. We will be inspecting the data and analyzing for insights along the way. The final step of the project is the construction and selection of a model. Either a Random Forest or XGBoost model appears to be most suited for the prediction we are looking for and the goals of the overall project. These two tree models will be trained and vetted on a validation set and a test hold out set.

IMPACT

Once a 'claim' has been identified then actions can be taken against the content. Therefore, grouping the 'claims' will cut out the time required for moderators to do the sorting themselves and allow them to concentrate on ensuring the content that has been flagged as a 'claim' is within our guidelines or take the necessary action.



With approximately 2900 rows of data in the testing set, we can see that the Random Forest model has consistently predicted our target variable. The issue is that the model's only mistakes are false negatives. We should look for ways to eliminate or reduce their production even more.



Video views, likes, shares, downloads, comments are the main features from which the model based its predictions. With cleaning up and further engineering the features, as well as tuning the model we can hopefully eliminate all false negatives.

KEY INSIGHTS

- With the construction of the models, we have determined that a Random Forest model is most suited for our tasks with a Precision of 100%, a Recall of 99%, and a F1-score and accuracy of nearly 100%.
- The Random Forest model was basing its predictions off of majorly 5 features: views, likes, shares, comments, and downloads. This are all indicative of community engagement.
- We could potentially improve and streamline this model by creating features such as: words/sec, likes/sec, ban status vs a categorizing of likes (low, medium, high). The worry about false negative should be addressed, since these 'claim' videos would be passed over by the moderators as they are falsely marked as 'opinion', and therefore harmful content might still by access by the community.
- We do not know of any biases in the model currently. However, we should take the time to add demographic information (age, race, location, etc), alongside the model data to see if we have any hidden biases.
- As of right now, I highly recommend, off of our current findings, that the company should move forward with starting to implement the model in our moderators' tool belt. We should do a partial roll out and take time to observe and document the increases in efficiency and whether any other biases show themselves.