**Project Overview**

The primary goal of the project was to create a machine learning model to help with classification of videos as either claims or opinions. For this purpose, I used a dataset of around 2000 videos posted on TikTok. Out of two tree-based classification models that were built, the random forest model was chosen the final model based on its better recall score. The model had nearly perfect prediction with only five misclassified samples out of 3,817. The primary predictors were all related to video engagement levels, with video view count, like count, share count, and download count accounting for nearly all predictive signal in the data.

Business Understanding

TikTok videos receive a large number of user reports for many different reasons. Not all reported videos can undergo review by a human moderator. Videos that make claims (as opposed to opinions) are much more likely to contain content that violates the platform's terms of service.

Data Understanding

The dataset used for this project contained synthetic created by TikTok in partnership with google. It included 18,383 observations with 12 features. Important features included count variables like view, share and like counts. As shown in the visualization below, the is a balance between classes of target variable. Also, another column was created in the dataset to get the text length for each video.

Claim vs Opinion Videos
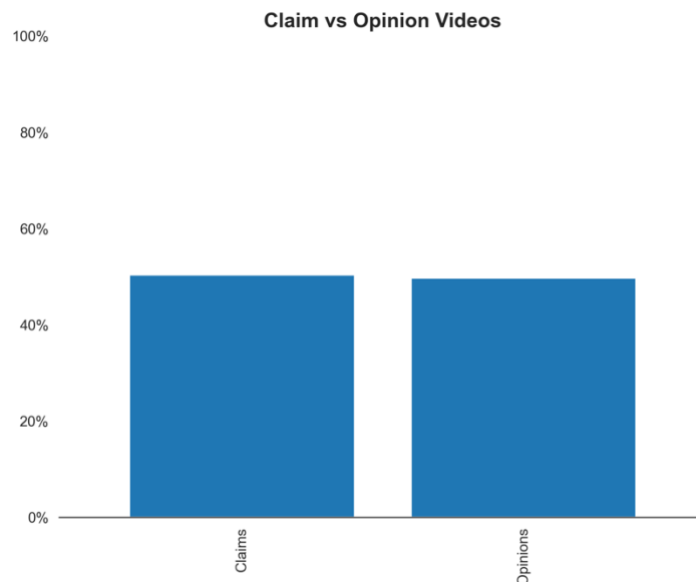
Modeling and Evaluation
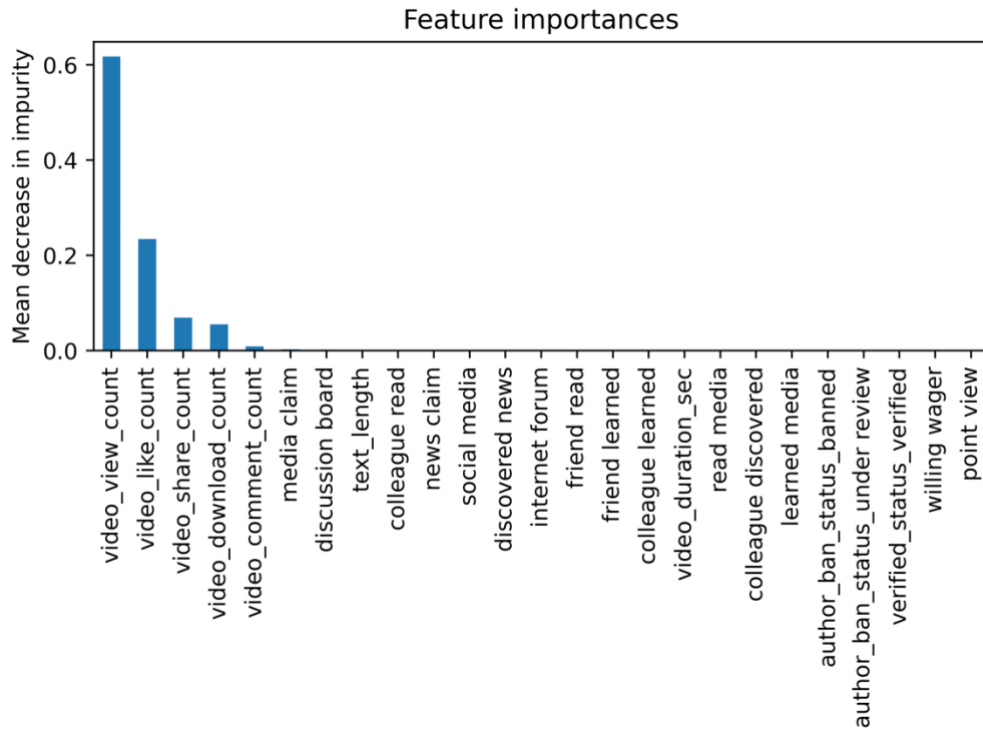
The final model, random forest, constituted 200 decision trees was used for evaluation and determining feature importance. The model's most predictive features were all related to the user engagement levels associated with each video. It was classifying videos based on how many views, likes, shares, and downloads they received. Because the model currently performs nearly perfectly, there is no need to engineer any new features.

Conclusion

With these results, we can conclude that videos with higher user engagement levels were much more likely to be claims. In fact, no opinion video had more than 10,000 views. The model would help TikTok's team and streamline their efforts to tackle user reports.