Unaccompanied Child Migrants Machine Learning Model

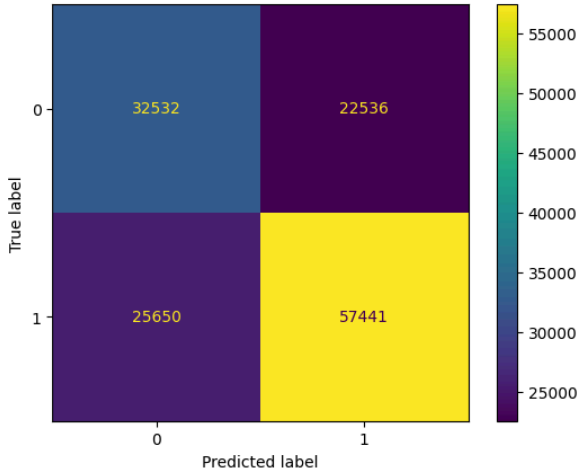
Conducted by Data Scientist Alyssa Ayala

Problem Overview

Unaccompanied child migrants to the United States who have been forced out of their native countries by war, or are lured in by better opportunities, are often detained by the **Office of Refugee Settlement** by the U.S.-Mexico border. Despite the court regulation that aims to release children to family sponsors within 20 days, **long detainment periods still remain a concern** within unaccompanied child migrants.

Project Objective

The New York Times has released data on unaccompanied child migrants from the **U.S. Department of Human Health and Services.** A machine learning model has been created to predict what makes a child more likely to stay detained for longer than 20 days based on this data.



Model Results

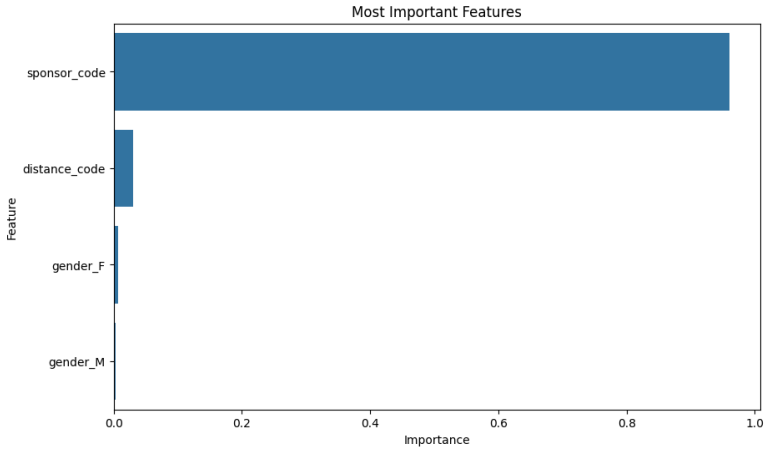
A tree-based model was fitted based on the roc_auc metric to obtain the most readable results. The outcome variable **long_detainment** boolean variable (0 = false, 1 = true) was created in advance based on the entry and release dates of each child.

The confusion matrix on the left reveals that the model's predictions are **65% accurate**. Other metrics of concern include a precision of 71%, recall of 70%, F1 score of 71%, and AUC of 68%.

Most Important Features

According to the model, the sponsor's type of relationship with the child **impacts their detainment time by 96%**. Less detainment time is likely if a parent or legal guardian sponsors their release, as opposed to a more distant family member.

It's possible that the child's native country's distance from the U.S. has more of an impact that may be demonstrated by a future model. Its current influence on detainment time is 3%.



Next Steps

For a more accurate model, we may consider trying different models for comparison, like a linear regression model, that may properly deal with any outliers and make features like the distance from the U.S. play a bigger role than what was demonstrated here. As for the model itself, **data leakage is still a concern.** Either metric may dramatically increase or decrease with future tests, which depends on the processing power of the computer being used. In this case, more project revisions may be necessary.