Positive-Unlabeled learning to identify forced labor at sea

Forced labor is a known problem in fisheries. However, the lack of data makes it hard to truly assess the extent of the problem. Here we use a machine learning approach to identify fishing vessels associated with forced labor. The model (a combination of random forests) trains on cases from reported vessels and uses a set of vessel characteristics and movement metrics obtained from worldwide tracking data to identify patterns of forced labor. Training and prediction are performed on a yearly basis, i.e. we are able to predict—for historical data—if a vessel had a high risk of forced labor at a given past year. Because we do not have enough reliable data on vessels that did not use in forced labor to train the model, we use positive-unlabeled (PU) learning; that is, a procedure to use positive cases—vessels with forced labor—alongside cases with an unknown class. Our model returns confidence scores for forced labor. We then used a density-based approach called the dedpul algorithm to find a threshold to classify between high risk and low risk of forced labor. The model proved to correctly identify vessels with and without forced labor. While this is a promising result, the data used for validation (and training) may not be representative of the diversity of forced labor cases in the world. We would like to discuss this and other caveats in our presentation. We believe that this machine learning approach is not only useful for fisheries management but also for other applications in ecology that may not have negative cases to train models.