Machine Learning Augmented Prediction for Labor Exploitation Detection

Stanford Data Science for Social Good 2024

Enkhjin Munkhbayar, Leon Reilly, Kyler Shu

August 30, 2024

Abstract

To enhance the detection of illegal charcoal production sites in Brazil, often linked to labor exploitation, we developed a machine learning solution to address the high false positive rate of an existing Convolutional Neural Network (CNN) model that identifies potential sites from satellite imagery. Our approach integrates relevant geospatial and survey data to train a Gradient Boost classifier, which distinguishes between false positive and true positive charcoal site identifications made by the upstream CNN. By reducing the false positive rate, this process improves the efficiency of human post-processing and subsequent field inspections, potentially leading to more effective interventions against labor exploitation in Brazil's charcoal industry.

Technical Mentors: Dr. Benjamin Seiler, Dr. Kimberly Babiarz, Jonas Junnior
Faculty Mentor: Dr. Michael Baiocchi

Stanford University

Stanford Data Science

1 Introduction

Over the past decade, Brazil has been one of the world's leading producers of crude steel.[1] In 2023 it was the ninth largest producer of crude steel in the world. In 2022 it was the largest producer of charcoal in the world—fuel to feed the steel industry. Yet, the process of producing charcoal is typically intertwined with violence and exploitative labor partly as a consequence of the dangerous and illegal nature of the work. Since 1995, about 2,830 people have been rescued from what Brazilian authorities define as slavery-like conditions. It's worth pointing out that illegal charcoal production has been difficult to combat since the work is typically done in remote locations and the sites operate for a finite amount of time. The Stanford Human Trafficking Data Lab and Brazilian Federal Labor Prosecution Office are collaborating to use a remote detection tool called CHAR to locate charcoal sites quickly and accurately. The Human Trafficking Data Lab developed a computer vision model trained on low-resolution, high-frequency satellite imagery to accurately identify and pinpoint zones where forced labor is likely used to produce charcoal. Human post-processing is then used to select sites for task force site inspection. The technology allows for a "more proactive anti-trafficking intervention, moving beyond merely reactive, tip-driven initiatives."

The major wrinkle with the existing model is its high rate of false positives, confusing roads, rivers and other visually similar features to the two parallel lines that are distinctive of charcoal kilns. As a consequence human post-processing often involves looking at images that cannot be charcoal sites due to their location, e.g. being in the ocean, on a highway, etc. Successful improvements over the current manual labeling would speed up task force prioritization and therefore the potential to improve high-level anti-trafficking outcomes, that is, workers rescued. To this end, our project develops a downstream Gradient Boosting Classifier model trained on appropriately selected geospatial and survey data to augment the upstreams model's accuracy.

2 Methodology

2.1 Feature Engineering

The bedrock of our work is built on a dataset produced by the upstream CHAR model's outputs. At the time of writing, CHAR has been deployed exclusively in the state of

[1] World Steel Association: Annual Production Data 2023

Maranhão, operating with a threshold value of 0.9, that is, only sites with a probability of 90% or higher of being a charcoal site are included in the dataset. Our dataset encompasses 5,278 flagged sites each manually reviewed to classify the site as a true positive or false positive. 478 flagged sites were labeled as true positives. The remaining variables, presented in Table 1, make up the rest of the skeletal framework we use to appropriately group sites.

Variable	**Notes**
Model Score	Probability output from CHAR model
Month and Year	Month and Year of the satellite image. Ranges from July 2023 to March 2024
Geometry	Precise location of flagged site
Tiling	Satellite imagery's unique tile ID from Planet Labs

Table 1: Variables in CHAR flagged site dataset.

For feature construction, we only need to consider the geometry of a flagged site. We began with thinking carefully about what key environmental or infrastructural elements might constitute intuitively significant covariates that would generate signal for the model. Using GeoPandas, features were calculated using two key metrics: calculating the shortest distance to features and counting how many specific features lie within a defined radius of a site. In total, we constructed a suite of 17 variables using this approach.

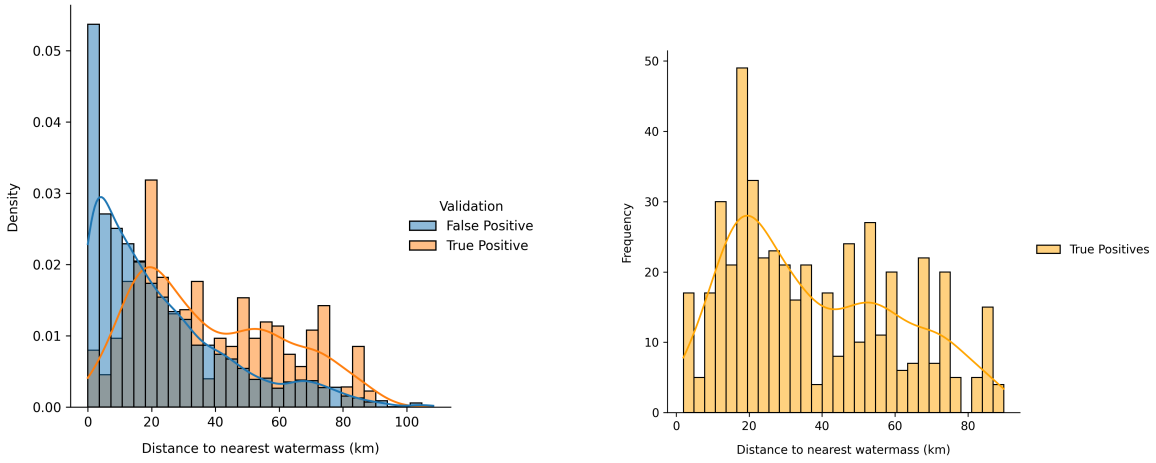


Figure 1: Distance to nearest watermass shows distinction between true and false positives

Recognizing that geography alone cannot capture all nuances associated with charcoal site locations, we expanded our feature set to include socio-economic variables leveraged from the MapBiomas and SmartLab datasets. The MapBiomas dataset—updated biweekly by the Brazilian government—provides geometries of areas that have been de-

Feature	Reasoning
Distance to nearest road	Charcoal needs to be transported to steel mills. Proximity to a road reduces logistical frictions.
Distance to nearest water mass	Charcoal sites require forests. Being too close or far may proxy for terrain or land type. Straightforwardly, charcoal sites cannot be in a river or lake.
Distance to nearest village/town	Charcoal sites want to avoid detection. We might expect sites to be further away from towns and villages.
Number of Charcoal sites	Maybe we keep this one but bc its not correct maybe not.

Table 2: Variables in CHAR flagged site dataset.

forested, independent of its legality. Since charcoal production requires wood, it stands to reason that charcoal sites are likely to be proximate to deforestation alerts.

The SmartLab dataset contains socio-economic survey data for every municipality in Brazil. This includes metrics such as literacy rate, poverty rate, and the number of workers rescued from conditions akin to slavery. Such variables serve as proxies for the socio-economic status of the region, which might influence site selection. In particular, areas with lower literacy rates and higher poverty may be more vulnerable to illegal charcoal production on account of potentially lowered detection risks and easier recruitment of labor. In total, our dataset includes 38 constructed features.

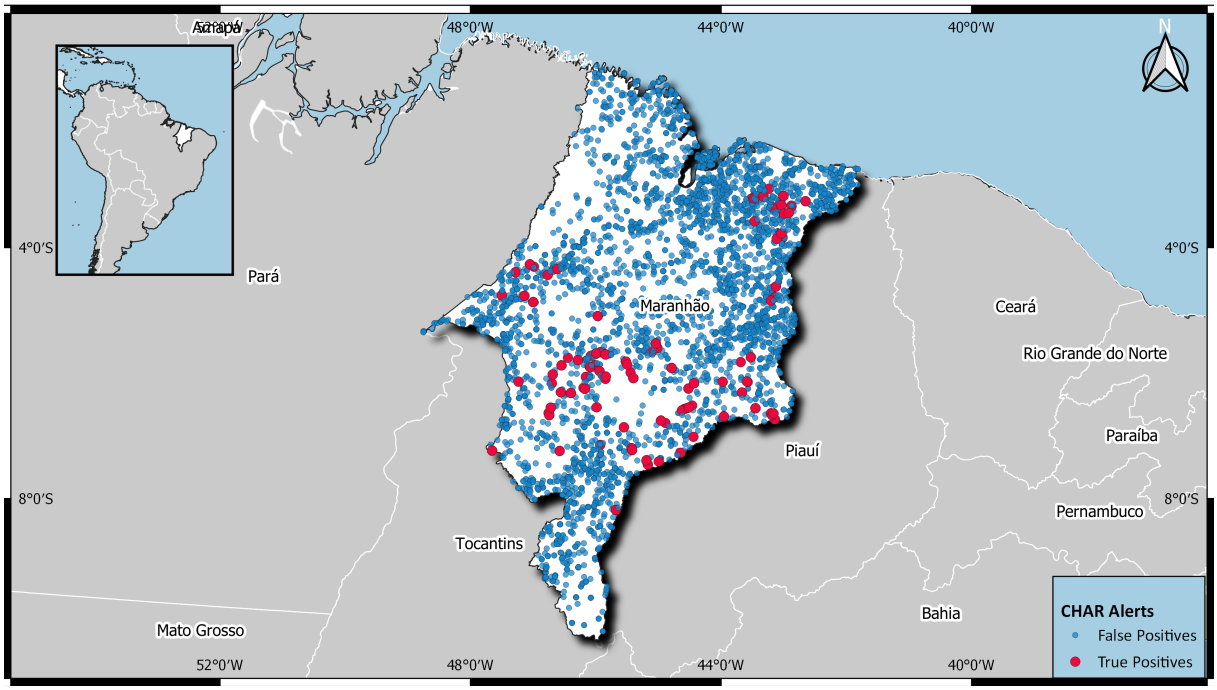


Figure 2: Map of True Positive and False Positive CHAR flagged sites.

2.2 Image Encoding Approaches

We used two approaches to encode information from the original image of the flagged site into the model, both using pretrained models. The first was a variational autoencoder with ResNet50V2 backbone, which outputted a 128 dimensional vector for each image. The second approach was using the MOSAIKS API, a pretrained global satellite imagery machine learning model for general purpose tasks. Given the latitude and longitude of each flagged site, this model outputs a 4000 feature vector encoding information about the location with 0.01 degree resolution.

2.3 Modeling

In order to train the model, it was necessary to use grouping and stratification to split the data across the training validation split and the individual cross validation folds. Because sites are often flagged multiple times over many months in the training data, it was necessary to group by unique location in order to prevent the leakage of knowledge between training and testing. Thus, all data points with the same x-tile and y-tile combination were placed in the same split. Without grouping, the model's performance would be drastically overestimated. Additionally, stratification of labels was necessary to balance the proportion of true positives across each split. This would ensure more consistent model results. Grouping and stratification was automated using sklearn's StratifiedGroupKFold and confirmed using tests run on the result splits that did indeed confirm approximately 9% true positives across all splits and no overlapping locations. We saved $\frac{1}{6}$ of the data for validation, with the remaining $\frac{5}{6}$ used for 5-fold cross validation during the training of the model.

Among the tested model architectures were logistic regression with l1, l2, and elasticnet penalties, SVMs, various tree-based methods, and a pretrained transformer-based model for tabular data known as TabPFN. Ultimately, the best performing model was a Gradient Boosting Classifier, sklearn's implementation of gradient boosted trees. Using GridSearchCV, we tuned hyperparameters such as max depth, learning rate, subsample proportion, and the minimum number of samples for a split.

2.4 Clustering

some clustering

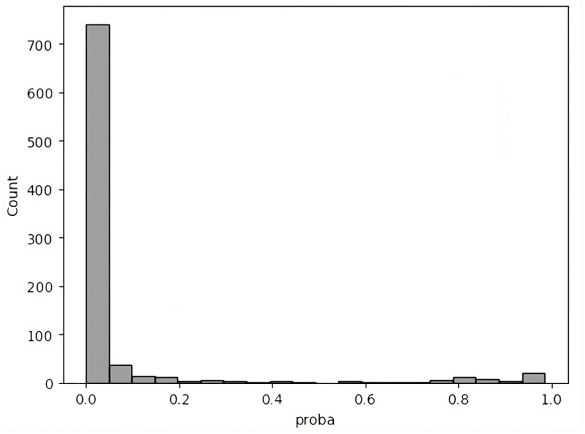
Metric	Score
F1	71.1%
Recall	65.8%
Precision	78.8%

3 Results

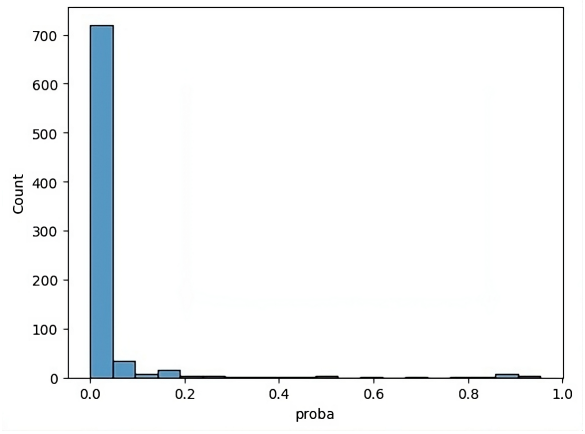
Using the above approach, our model achieved the following results at 0.25 threshold:

These results were achieved while leaving out the image encoding data because both the ResNet embeddings and MOSAIKS features did not improve and sometimes hindered the performance, despite the data containing intuitively important information. Future work will be necessary to incorporate this data in a beneficial way to the model.

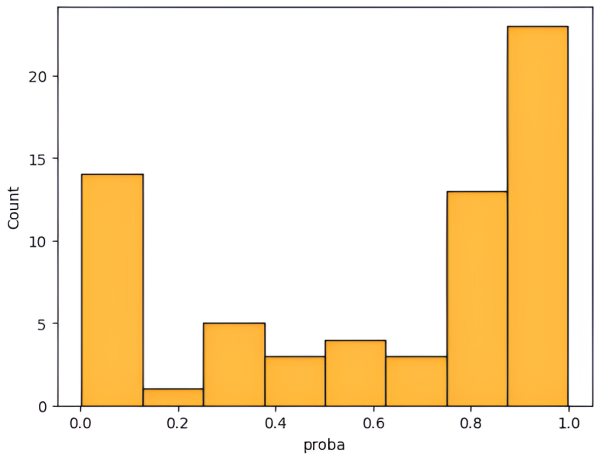
Observing the outputted predicted probabilities from the model (fig. 3a), it is evident that the model is making strong distinctions. However, it is struggling to correctly classify a few difficult data points, most notably in the false negatives (fig. 3b).



(a) Predicted probabilities for all sites.



(b) Predicted probabilities for true negatives.

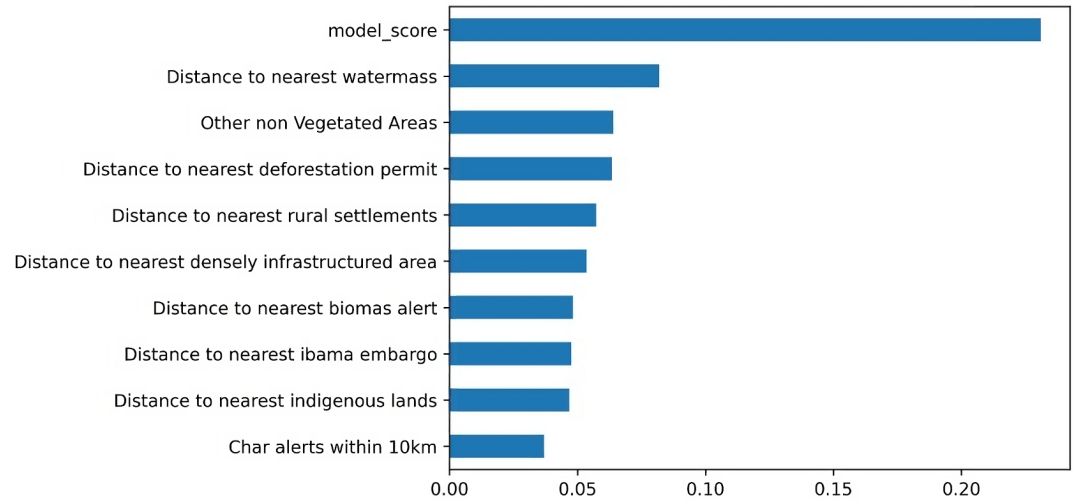


(c) Predicted probabilities for true positives.

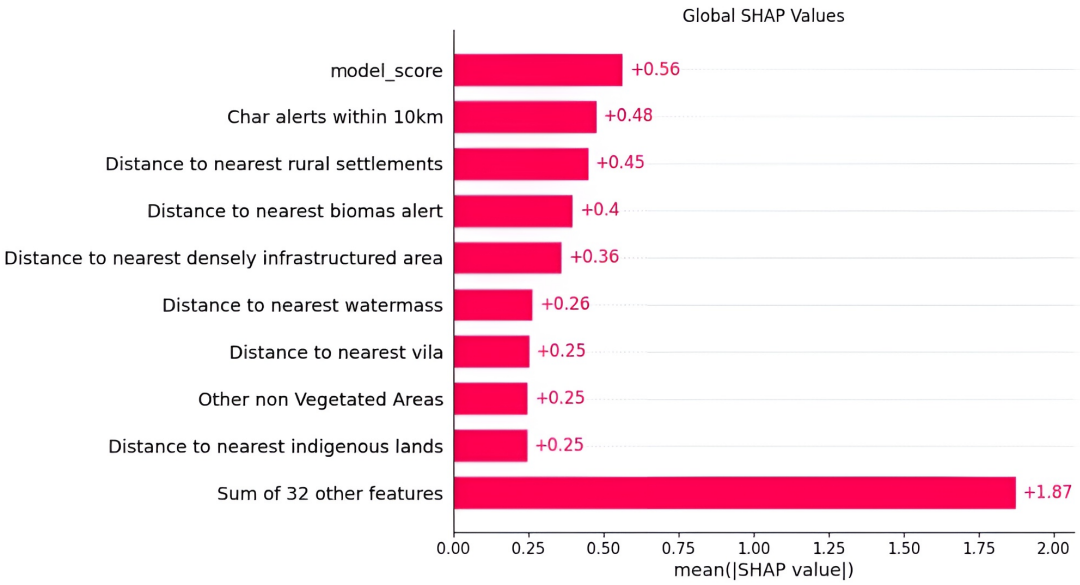
Figure 3: Predicted probabilities grouped by ground truth.

3.1 Feature Importance

We used two different methods for evaluating feature importance. The first is the built-in feature importance functionality in the gradient boosting model (fig. 4a). The second is using TreeSHAP (fig. 4b). Feature importances were calculated on the best performing data + model, excluding the image encoding approaches both because they performed worse and are unexplainable.



(a) Gradient boosting tree built-in feature importance.



(b) TreeSHAP values.

Figure 4: Feature importance results.

From these two methods, it is clear that model score, the confidence level of the first stage of the model looking only at the image, was the most important feature. Other important geospatial covariates included distance to nearest infrastructure / rural

settlements, watermasses, indigenous lands, deforestation (biomas) alerts, and the number of other flagged sites within 10km. These feature importances generally align with in-field intuition, such as the fact that charcoal sites cannot be located on indigenous lands, deforestation is necessary for charcoal production, and infrastructure and water may be necessary resources for setting up a site. A notable deviation from in-field experience is the lack of distance to nearest road as an important factor, as roads are necessary to transport charcoal away from production sites, as well as the fact that the first stage of the model often mistakes roads themselves for charcoal sites.

3.2 Clustering Results

some results

4 Conclusion

The Gradient Boosting Classifier model developed in this project aims to complement the CHAR system in detecting illegal charcoal production sites in Brazil. By addressing the issue of false positives, the model seeks to improve the efficiency of human post-processing and resource allocation for field inspections. This approach has the potential to support more targeted interventions in areas at high risk for labor exploitation. Additionally, the model's insights could contribute to the development of preventive strategies, possibly shifting some efforts from reactive measures to proactive ones. As research in this area continues, such data-driven methods may inform policy decisions and resource distribution in the fight against labor exploitation in Brazil's charcoal industry. While challenges remain, this project represents a step towards enhancing anti-trafficking efforts through improved detection techniques.

5 References