STANFORD BUSINESS GRADUATE SCHOOL OF

Harnessing Mixed-Scale Datasets to Illuminate Hidden Labor Abuses: The Case of Child Labor in Ghana

Antonio Torres Skillicorn[1], Prof. Dan Iancu[2], PhD, & Prof. Sarah Billington[1], PhD

Stanford School of Engineering & Doerr School of Sustainability
Civil & Environmental Engineering

Motivation: High-quality data for constructing predictive models of forced and child labor is scarce, and existing large-scale surveys often do not include interviews with children. **We focus on child labor in Ghana as a case study, combining two datasets.**

Core Research Idea: To develop a more robust predictive model at the national level, we plan to leverage 1) the increased reliability of the NORC child labor outcome variables and 2) the larger sample size and nationally representative nature of the GLSS7 data (Figure 1).

GLSS7

- N = 15,000 households
- Only in Ghana
- Nationally representative
- 665 questions
- 12 topic areas
- **Only 31% of children interviewed (n = 6,279)**

NORC at the University of Chicago

- N = 2,821 households
- Ghana and Cote D'Ivoire
- Only cocoa growing areas
- 500 questions
- Surveys with 5 stakeholders
- **All children interviewed (n = 5,534)**

Our Approach:

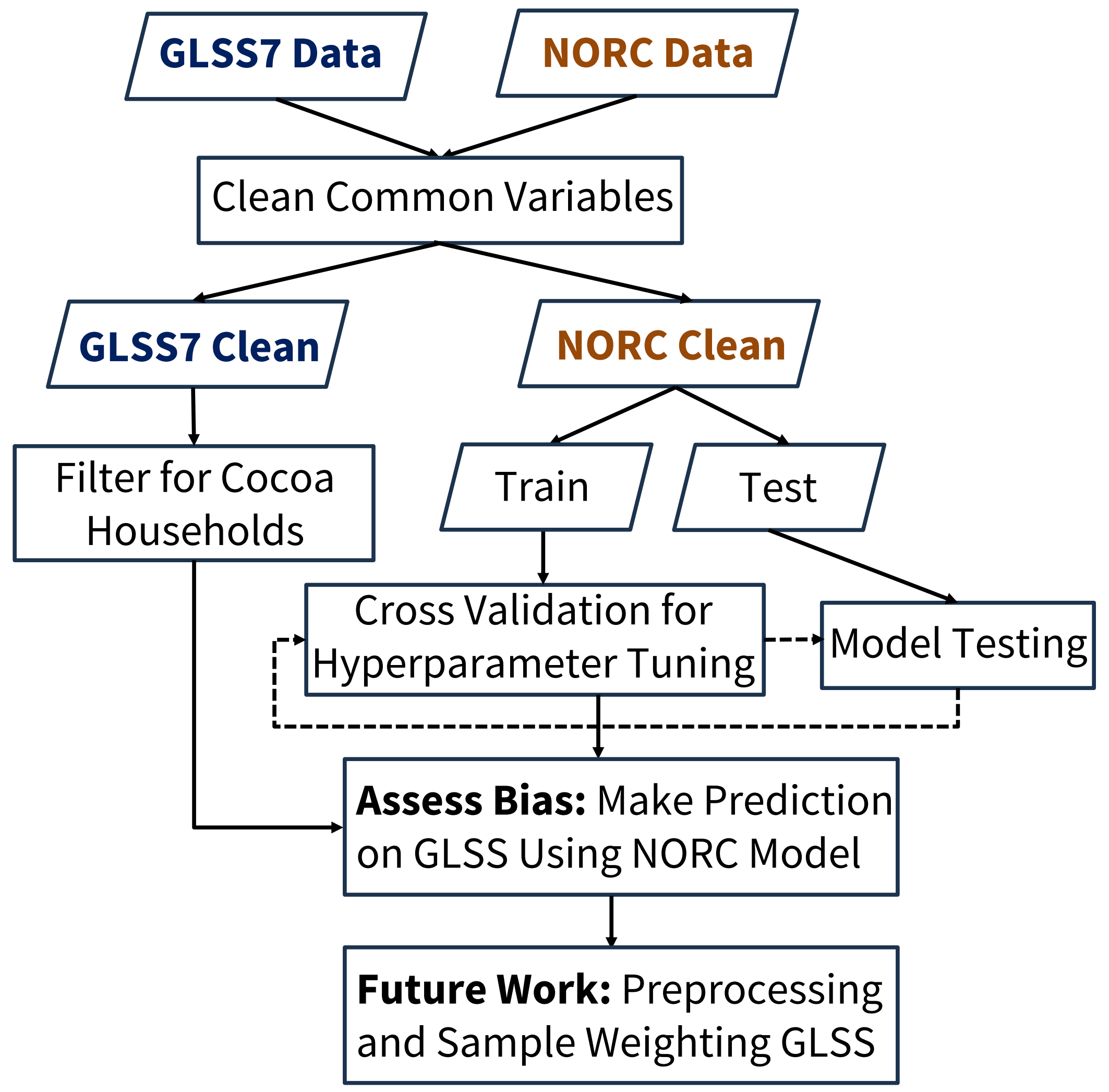
GLSS7 Data
NORC Data
Clean Common Variables
GLSS7 Clean
NORC Clean
Filter for Cocoa Households
Train
Test
Cross Validation for Hyperparameter Tuning
Model Testing
Assess Bias: Make Prediction on GLSS Using NORC Model
Future Work: Preprocessing and Sample Weighting GLSS

Figure 1: Project pipeline summarizing our work to date and anticipated next steps.

* Child labor is an engineered feature for the GLSS7 dataset, calculated according to the ILO common definition, which refers to hazardous conditions and thresholds of allowable working hours

Identifying Common Variables: we have identified 4 common child labor outcome variables and have cleaned 40 common predictive features between the datasets.

Child Labor Outcomes

- 7-Day (7D) Child Work
- 12-Month Child Work
- Child Labor*
- Total Hours Worked

Predictive Features

- 10 Household
- 3 Child Level
- 21 Agricultural
- 6 Economic

Outcomes: After filtering for cocoa-producing households, important variables such as child work, child labor, and total hours worked are underreported in the GLSS7 dataset compared to the NORC dataset (Figure 2).

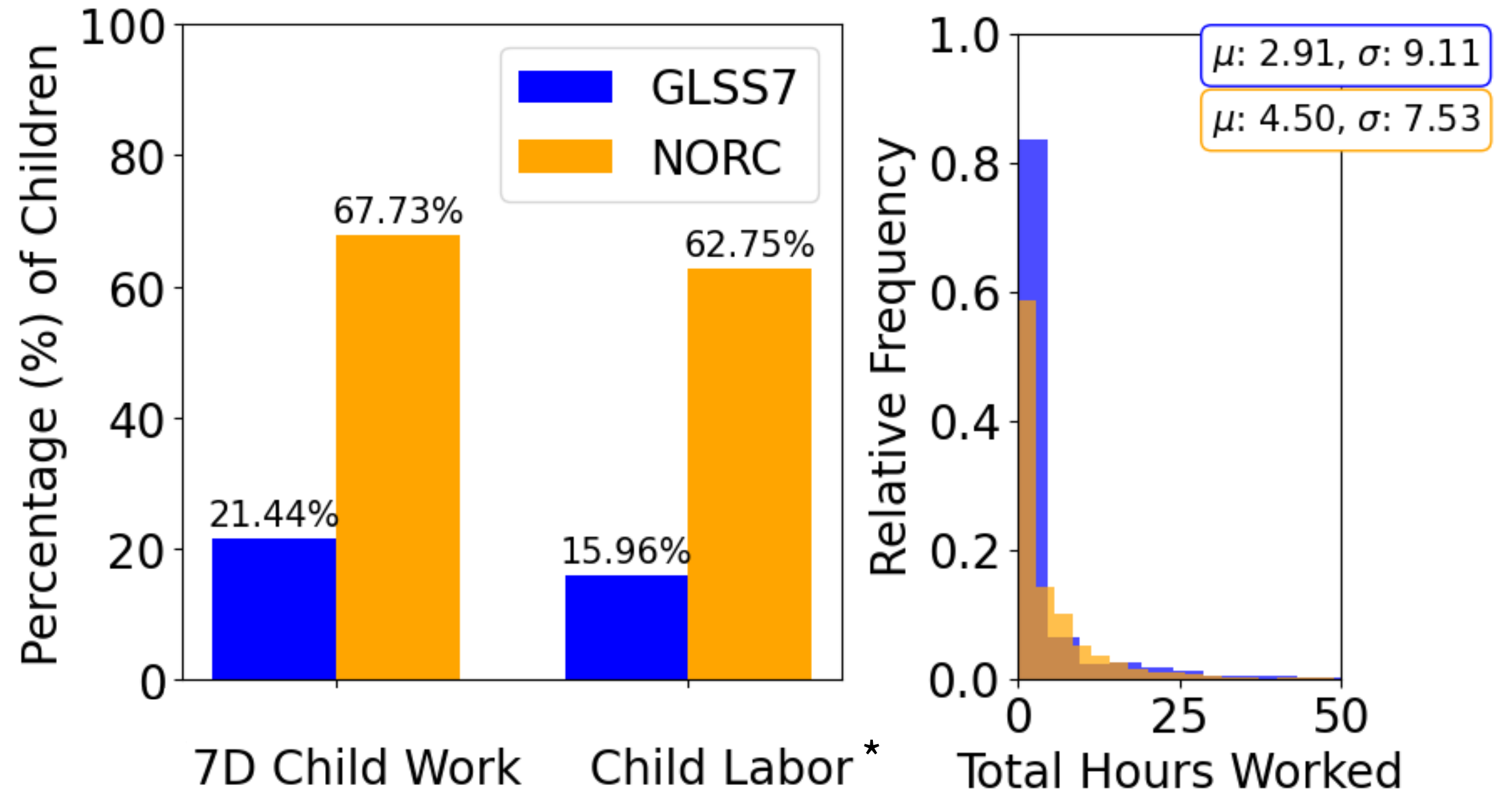

Figure 2: Comparison of child work and child labor outcome variables for the GLSS7 and NORC datasets. Outliers are removed for the total hours worked using an interquartile range (IQR) approach. Both datasets are filtered for cocoa households.

Predictors: Child and household-level demographics match reasonably well between the two datasets (figure 3)

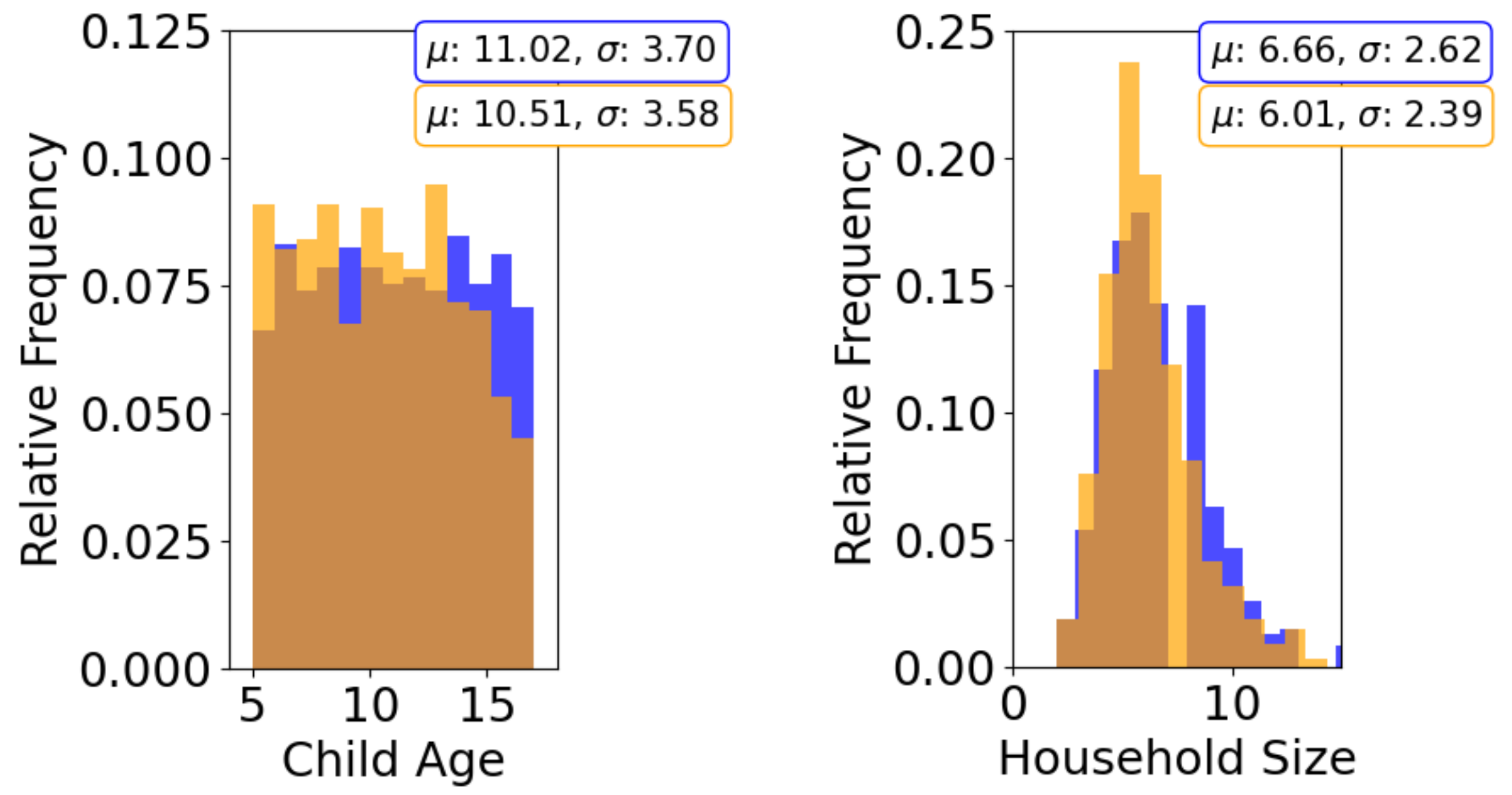

Figure 3: Comparison of child age and household size distributions for the GLSS7 and NORC datasets. Both datasets are filtered to focus on cocoa producing households.

Assessing Bias in the GLSS7 Data: We trained an XGBoost classifier on the NORC data and generated predictions of 7-day child work for both the NORC test holdout as well as the filtered GLSS7 datasets (Figure 4). As expected, the model predicts a high percent of false positive cases of child work in the GLSS7 (see red box), likely due to the underreporting of actual instances of child work observed in Figure 2.

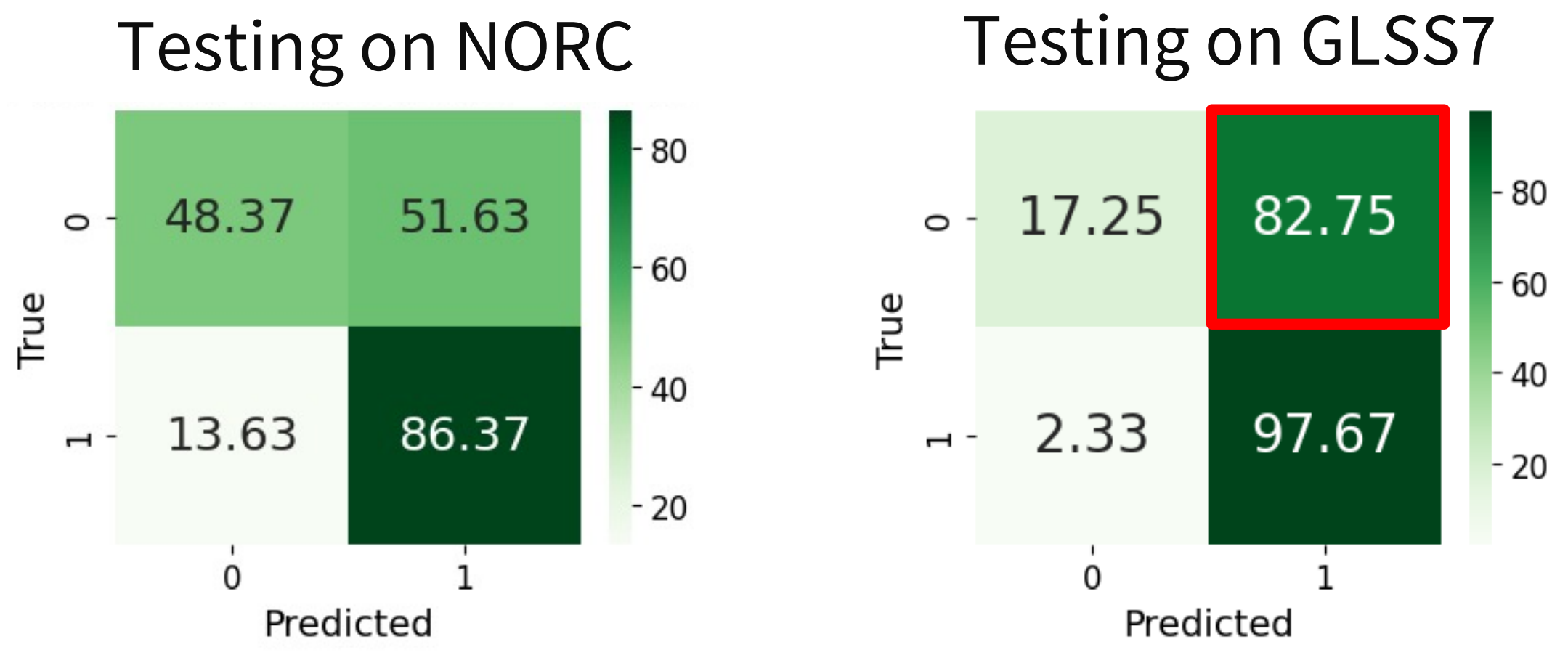

Figure 4: Confusion matrices for the XGBoost classifier's predictions on the NORC test holdout and GLSS7 datasets. Model hyperparameters were optimized using a 5-fold cross-validated grid search.

SHAP Values: SHAP values provide preliminary insights into potential drivers of child work like child age and the amount of land households cultivate (Figure 4).

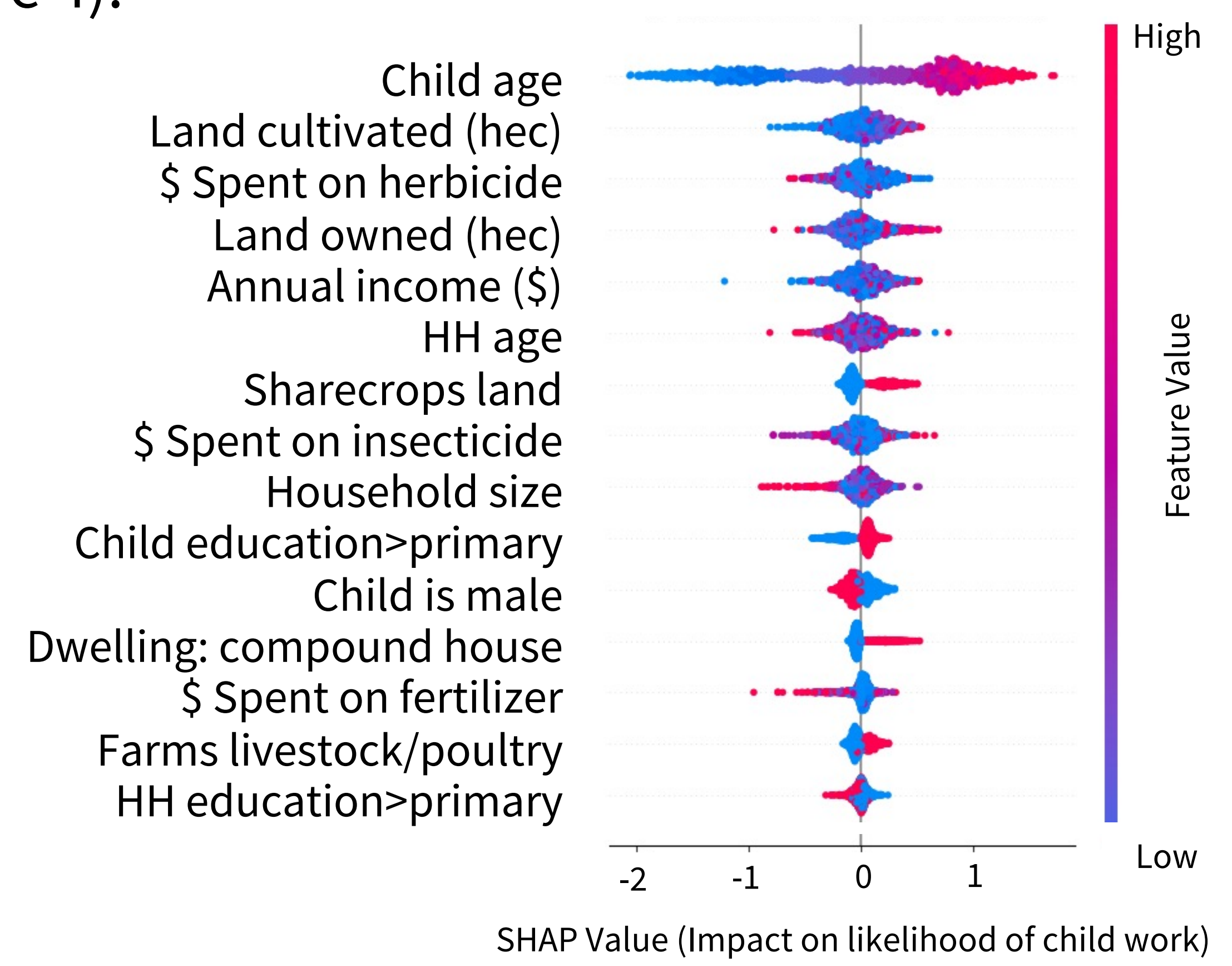

Figure 5: SHAP values for the NORC XGBoost model when making predictions on the NORC test holdout.

Next steps: In the future, we will 1) explore sample weighting and pre-processing approaches to 'de-bias' the GLSS7 dataset and 2) train a larger predictive model using the de-biased GLSS7 data that will include a broader range of predictive features with the goal of creating a more generalizable model for the occurrence of child work.

[1] Civil and Environmental Engineering

[2] Graduate School of Business