IMPACT AREA FUND

Interpretability of Machine Learning in Child Welfare

By Brian Chor and Zhidi Luo

Background

Chapin Hall researchers tested an interpretation methodology on a machine learning model that predicted whether youth in a child welfare system ran away from care. This methodology might help child welfare administrators who value data-driven decisions and interpretability of more complex predictive analytic models.

Child welfare administrators are increasingly interested in using predictive analytics to inform prevention strategies. They might want to know how specific predictors, for example, age, or combinations of predictors, for example, age and gender, are associated with an increased or decreased risk of a child welfare outcome, for example, youth running away. These analyses have traditionally been conducted using regression-based approaches. An important limitation of regression-based approaches is that researchers must prespecify individual predictors and combination of predictors, making educated guesses about what characteristics might in fact be associated with various outcomes. However, it is not always possible for researchers to anticipate and, therefore, to prespecify which combinations of predictors might help explain complex phenomena such as child welfare outcomes. Machine learning (ML) models have the potential to overcome this limitation because they do not require *a priori* hypotheses about the predictors. In short, ML models mine data and "learn" underlying patterns of these predictors (Ghani & Schierholz, 2020).

Despite the predictive power of ML models, they are difficult to apply and interpret. This has stymied translation of research into practice and policy in child welfare. As a proof of concept, we sought to bridge this gap in child welfare by testing a novel interpretation methodology in ML, Shapley Additive Explanation or SHAP (Lundberg et al., 2020). First, we developed a random forest ML model to predict the risk of youth running away from care within 90 days of entering a child welfare system. Second, we applied SHAP to the random forest ML model to identify and quantify the influence of important predictors and combination of predictors on the predicted risk of runaway. Demonstrating that SHAP can be used in child welfare research might facilitate end users of ML, such as child welfare administrators and caseworkers, in making relevant policy and practice changes.

Methods

We examined 8,255 legal custody spells for youth who were 12 to 17 years old and who entered the legal custody of the Illinois Department of Children and Family Services (DCFS) between January 1, 2010 and June 30, 2018. We used DCFS administrative data to operationalize 29 predictors (see Table 1) and the outcome of youth running away from DCFS care within 90 days of entry to care. We used R 4.0.3 (The R Foundation for Statistical Computing, 2020) to construct a random forest ML model and to apply the SHAP methodology to the model to extract important predictors and combinations of predictors relevant to the outcome. This study was part of a larger study evaluating DCFS implementation of a practice model and other innovations under a Title IV-E Waiver demonstration project. The University of Chicago School of Social Service Administration-Chapin Hall Institutional Review Board and the DCFS Institutional Review Board reviewed and approved the larger study.

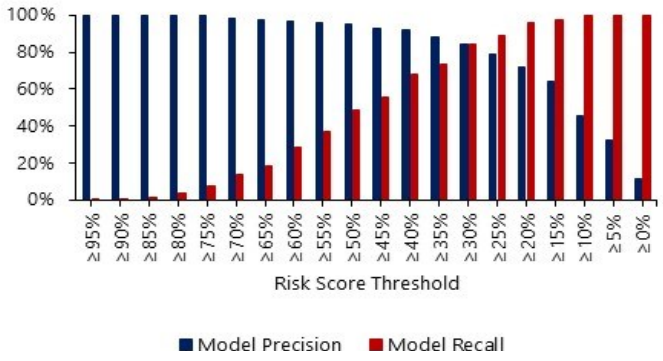
Table 1. 29 Predictors: Four Demographic Characteristics, 20 Child Welfare Characteristics, and Five Clinical Characteristics

Predictor	Description
Demographic characteristics	Gender: Female
	Race
	White (reference group)
	Black
	Other
	Age at beginning of each placement in youth's spell
Child welfare characteristics	DCFS administrative region
	Northern (reference group)
	Cook
	Central
	Southern
	Number of prior DCFS spells
	Year DCFS spell began
	Developmental disability status
	Most serious allegation prior to each placement in youth's spell: Sexual abuse
	Most serious allegation prior to each placement in youth's spell: Physical abuse
	Most serious allegation prior to each placement in youth's spell: Substance-exposed infants
	Most serious allegation prior to each placement in youth's spell: Emotional abuse
	Most serious allegation prior to each placement in youth's spell: Lack of supervision
	Most serious allegation prior to each placement in youth's spell: Environmental neglect
	Most serious allegation prior to each placement in youth's spell: Other neglect
	Most serious allegation prior to each placement in youth's spell: Substantial risk of harm
	Number of prior placements in youth's spell
	Placement with sibling
	Placement type
	Nonkinship (reference group)
	Kinship
	Residential
	Other
	Number of prior runaway events
Clinical characteristics	Most recent Child and Adolescent Needs and Strengths (CANS) assessment, within 90 days: Actionable Traumatic Stress Symptoms domain
	Most recent CANS assessment, within 90 days: Actionable Emotional/Behavioral Needs domain
	Most recent CANS assessment, within 90 days: Actionable Risk Behaviors domain
	Most recent CANS assessment, within 90 days: Actionable Social Functional Behavior domain
	Missing most recent CANS assessment, within 90 days

KEY FINDINGS

Accuracy of ML Model in Predicting Youth's Runaway

Figure 1. Relationship between Model Risk Score Threshold, Model Precision, and Model Recall.

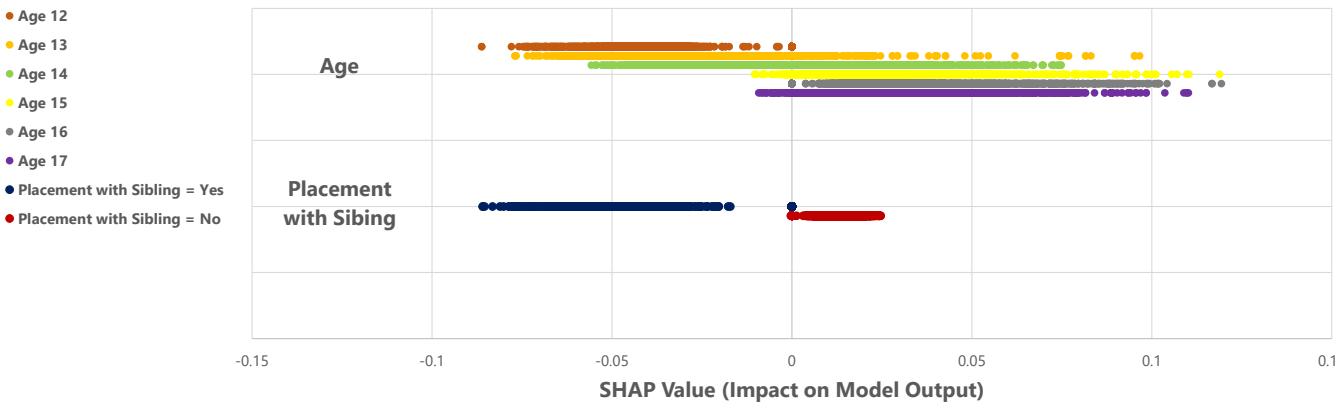


The random forest ML model achieved excellent overall prediction accuracy with an area under the sensitivity vs. 1-specificity curve of 0.88, when "perfect" prediction would yield an area under the curve of 1.00. We also examined model precision and model recall at specific risk score thresholds, which range from 0% to 100% (see Figure 1). Each risk score threshold means, "If the model-predicted likelihood of a youth running away is at or above this threshold, the model considers this youth predicted to run away." Model precision was defined as among youth's spells predicted to run away within 90 days of entry to care at each risk score threshold, the percentage of youth's spells with an actual runaway event. Model recall was defined as among all youth's spells with an actual runaway event within 90 days of entry to care, the percentage of youth's spells with an actual runaway event at each risk score threshold. As youth's predicted risk score threshold increased, so did model precision, without sacrificing model recall. Among youth with a ≥ 40% predicted risk of running away within 90 days of entry to care, 91.92% of them actually ran away. This accounted for 67.82% of all youth in the sample who ran away.

SHAP Values: Individual Predictor's Contributions to Youth's Predicted Risk of Runaway Compared to an Average Youth

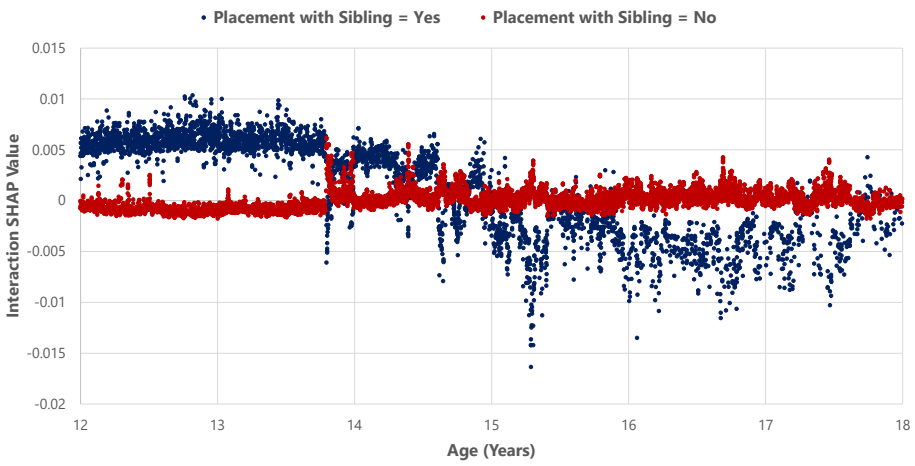
Positive SHAP values indicated an increased predicted risk of runaway relative to an average youth. Conversely, negative SHAP values indicated a decreased predicted risk of runaway relative to an average youth. Figure 2 shows the SHAP values for the two predictors with the highest average SHAP values: age (average SHAP value = 0.030) and placement with a sibling (average SHAP value = 0.021). Specifically, older youth (see yellow, gray, and purple dots) had positive SHAP values, which indicated an increased predicted risk of runaway. In contrast, youth coplaced with a sibling (see dark blue dots) had negative SHAP values, which indicated a decreased predicted risk of runaway.

Figure 2. SHAP Values for Age and Placement with a Sibling



SHAP Interaction Values: Important Combinations of Predictors of Youth's Predicted Risk of Runaway

Figure 3. SHAP Interaction Values for Age and Placement with a Sibling



Positive SHAP interaction values indicated an increased predicted risk of runaway relative to an average youth. Conversely, negative SHAP interaction values indicated a decreased predicted risk of runaway relative to an average youth. Figure 3 shows the SHAP interaction values for age and its interaction with placement with a sibling. SHAP identified this interaction as potentially important because individually the two predictors were important (per high SHAP values). In addition, age and placement with sibling appeared to interact. That is, youth younger than age 15 and coplaced with a sibling had positive SHAP interaction values (see the blue dots in the upper left corner in Figure 3), which indicated an increased predicted risk of runaway. In contrast, youth older than age 15 and coplaced with a sibling had negative SHAP interaction values (see the blue dots in the lower right corner in Figure 3), which indicated a decreased predicted risk of runaway.

Implications

When a methodologically sound ML predictive analytic model is appropriate to answer a research, policy, or practice question, child welfare administrators who value data-driven decisions and interpretability might consider using the SHAP interpretation methodology to:

- Demystify the "black box" nature of ML predictive analytic models.
- Identify relationships between predictors and outcomes not previously known, especially for complex outcomes for which there are large-scale, administrative data to leverage ML.
- Identify system-level implications, for example, tailored service pathways, by examining aggregated SHAP values of individual predictors to focus on predictors with the greatest average impact on increasing or decreasing risk.
- Identify case-level implications, for example, a specific youth's needs, by examining individual youth's SHAP values that illuminate the youth's protective and risk factors.
- Explore predictor interactions of interest, for example, "Early childhood services for coplaced children are paramount to my role as an administrator," and blind spots, for example, "I have never thought about looking at the interaction of ethnicity and age," by examining relevant SHAP interaction values that may provide similar policy and practice guidance at the system and case level.

References

Ghani, R., & Schierholz, M. (2020). Machine learning. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science: Data science methods and tools for research and practice* (2nd ed., pp. 143-191). CRC Press.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

The R Foundation for Statistical Computing. (2020). *R version 4.0.3*. The R Foundation for Statistical Computing.

Suggested Citation

Chor, K. H. B., & Luo, Z. (2021). *Interpretability of machine learning in child welfare*. Chicago, IL: Chapin Hall at the University of Chicago.

Contact

Brian Chor, Ph.D.
Senior Researcher
bchor@chapinhall.org
(773) 256-5211

Zhidi Luo, M.S.
Associate Researcher
zluo@chapinhall.org
(773) 256-5219

Statement of Independence and Integrity

Chapin Hall adheres to the values of science, meeting the highest standards of ethics, integrity, rigor, and objectivity in its research, analyses, and reporting. Learn more about the principles that drive our work in our Statement of Independence.

Chapin Hall partners with policymakers, practitioners, and philanthropists at the forefront of research and policy development by applying a unique blend of scientific research, real-world experience, and policy expertise to construct actionable information, practical tools, and, ultimately, positive change for children and families.

Established in 1985, Chapin Hall's areas of research include child welfare systems, community capacity to support children and families, and youth homelessness. For more information about Chapin Hall, visit www.chapinhall.org or @Chapin_Hall.