A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions Supplementary Materials

Editors: Sorelle A. Friedler and Christo Wilson

Appendix A. Train-Test Split

In this section we describe how we partition the observed data for training and testing purposes. The basic strategy is to form a graph whose connected components correspond to children who have been observed together on at least one referral. By randomly partitioning the set of connected components in this graph we obtain a split that is guaranteed to satisfy two properties: (1) No child will appear in both the train and test test; (2) All children from a given referral will appear together in either the test or the train set.

Let c_i, $i = 1, \dots, N_c$ denote the distinct children in the data, and let r_k, $j = 1, \dots, N_r$ denote the distinct referrals. We will write $c_i \in r_k$ to mean that child i was involved in referral k. Consider the graph G formed by connecting vertices c_i and c_j whenever there exists an r_k such that $c_i, c_j \in r_k$. Let $\{G_\ell\}_{\ell=1}^{L}$ denote the (maximal) connected components of G. Randomly partition the set of connected components into two sets, $\mathcal{G}_{Train}$ and $\mathcal{G}_{Test}$. Define `Train` to be all the referral records of children c_i that are vertices of $\mathcal{G}_{Train}$. Similarly, defined `Test` to be all of the referral records of children c_i that are vertices of $\mathcal{G}_{Test}$.

Appendix B. Forecasting validation

The partitioning scheme we apply ensures non-overlapping children and referrals between the data used for training and the data used to evaluate the performance of each model. While this is a reasonable way to evaluate the predictive performance of the model, another approach is to see how a model trained on historical data predicts outcomes for future referrals. In this section we provide some performance assessments of the models in this forecasting setting. That is, we assess how the model performs on predicting outcomes for the most recent referrals in the observed data.

We evaluate four modeling alternatives. We use $46,503$ available screened-in referrals for the period April 2010 to July 2014, and split the data into training and test set based on the time of the referral. This guarantees that the same referral is never observed during both training and test phases, though the same child may be. Alternative A corresponds to a training set with referrals between April 2010 and December 2013, and a test set with referrals between January and July 2014. Alternative B uses the same period for training, but restricts the test set to non-overlapping clients with the training set. Alternative C restricts the training set to a shorter time period, using referrals only between April 2010 and December 2011, and the same test set as alternative A. Finally, alternative D uses the same period for training as C, and restricts the test set to non-overlapping clients with the training set. Table 1 shows training and test sizes for each alternative.

Figure 1 and Table 2 show the performance for models trained using Logistic Regression and Random Forests, for each alternative. From Figure 1 it can be observed that the performance is stable across the different alternatives, for both algorithms. Technically, a slight increase in the performance of alternative B as compared to alternative A might be due to the test set size, which is more restricted for alternative B. Using a smaller training set as in alternatives C and D has an effect on the performance, due to a smaller training set for learning.

The alternatives explored provide a clearer picture of the predictive performance of the model

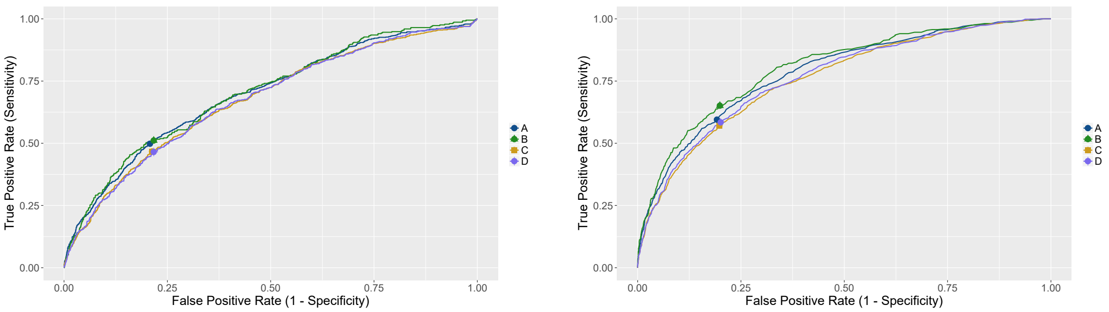

Figure 1: ROC curves for alternative training/test partitions. Left: Logistic Regression model. Right: Random Forest model.

Alternative	**Training set size**	**Test set size**
A	40,531	5,972
B	40,531	3,254
C	18,809	5,972
D	18,809	4,638

Table 1: Number of records for each alternative. Alternative A considers referrals between 2010 and 2013 for training. Alternative B uses the same time-period for training, by removing overlapping clients from the test set. Alternatives C and D follow a similar approach to A and B, respectively, using referrals between 2010 and 2011 for training.

for new cases coming into the hotline. We find that models trained to historical data do generalize well to new incoming referrals.

Method	**Alternative**	**AUC**	**TPR**	**FPR**
Random Forest	A	0.78	0.59	0.19
Random Forest	B	0.80	0.65	0.20
Random Forest	C	0.76	0.57	0.20
Random Forest	D	0.77	0.58	0.20
Logistic Regression	A	0.69	0.50	0.21
Logistic Regression	B	0.70	0.51	0.22
Logistic Regression	C	0.67	0.47	0.21
Logistic Regression	D	0.67	0.46	0.22

Table 2: Performance for each alternative. TPR and FPR correspond to the 25% highest risk cutoff (ventile scores of 16 and higher).