Children and Youth Services Review 91 (2018) 156–167

Contents lists available at ScienceDirect

Children and Youth Services Review

journal homepage: www.elsevier.com/locate/childyouth

Utilizing predictive modeling to enhance policy and practice through improved identification of at-risk clients: Predicting permanency for foster children

Dallas J. Elgin*

2M Research, United States

ARTICLE INFO

Keywords:
Predictive modeling
Permanency
High-risk clients
Predictive analytics
Big data
Best practices

ABSTRACT

Child welfare agencies are increasingly required to leverage their limited resources to meet nearly limitless demands. As a result, agencies are searching for new opportunities to efficiently improve policy and practice, and advances in data availability and technology have brought increased attention to the utility of predictive modeling. While the literature has often highlighted the considerable potential of predictive models leveraging "big data", discussions of the methodology and the associated best practices remain critically absent. To address this gap, this paper provides an illustrative case involving the development and testing of models used to predict the probability of whether U.S. foster children would achieve legal permanency. The models were trained and tested using a national administrative dataset of 233,633 foster care children that discharged from state child welfare systems in 2013. The optimal model, a boosted tree, predicted whether children would achieve permanency with 97.66% accuracy. The paper concludes with a discussion of best practices detailing how agencies can utilize predictive modeling to enhance policy and practice.

1. Introduction

Child welfare agencies operate in an environment that increasingly requires using limited resources to meet nearly limitless demands. Within this environment, agencies are increasingly searching for efficient management tools that will allow them to improve the effectiveness of their policies and practice (Clarke & Margetts, 2014; Lynn, Heinrich, & Hill, 2001). Due in part to advances in computing technology as well as the increased volume at which agencies collect administrative data, predictive modeling (also commonly referred to as "predictive analytics" or "data science") allows agencies to utilize data on past events to predict the likelihood of future events (James, Witten, Hastie, & Tibshirani, 2013; Kuhn & Johnson, 2013).[1] Over the past several decades, predictive modeling has been utilized in a variety of fields and settings to predict diverse outcomes, including the likelihood of hospital readmission (Kansagara et al., 2011; Raven, Billings, Goldfrank, Manheimer, & Gourevitch, 2009), identifying credit card fraud (Bhattacharyya, Jha, Tharakunnel, & Westland, 2011), predicting bankruptcies during the Great Recession (Serrano-Cinca & Gutiérrez-Nieto, 2013), and estimating risk among child welfare clients (Gillingham, 2015; Vaithianathan, Maloney, Putnam-Hornstein, & Jiang, 2013). However, a concerning trend has emerged where many of the predictive models have been deemed "proprietary" (Jackson & Marx, 2017), thereby concealing the methodological processes associated with developing and testing the accuracy of predictive models. While this issue is indicative of the important ethical and legal implications raised by the emerging methodology (Bovens & Zouridis, 2002; Cohen, Amarasingham, Shah, Xie, & Lo, 2014; Cuccaro-Alamin, Foust, Vaithianathan, & Putnam-Hornstein, 2017), the development of predictive models in accordance with best practices for methodological transparency, implementation, and interpretation can considerably enhance the efficiency, effectiveness, and equity of child welfare policies and practice.

Predictive modeling offers a dynamic approach for assessing the risk that individual clients will experience adverse outcomes, with rigorous training and testing of predictive models providing agency staff with an efficient and effective process for accurately identifying at-risk clients. Agency staff can subsequently run validated predictive models on a regular basis (e.g., monthly, quarterly, or as agency needs and resources permit) to obtain dynamic predictions that reflect the changes in a client's case. In these regards, predictive models leveraging "big data" offer considerable promise for both policy and practice. Potential

* Corresponding author.
E-mail address: delgin@2mresearch.com.

[1] More formally, predictive modeling has been defined as the process of selecting a model that best predicts the probability of an outcome (Geisser, 1993) or generates an accurate prediction (Kuhn & Johnson, 2013).

https://doi.org/10.1016/j.childyouth.2018.05.030
Received 7 March 2018; Received in revised form 25 May 2018; Accepted 27 May 2018
Available online 28 May 2018
0190-7409/

contributions include improving the ability to predict policy outcomes of interest (Cook, 2014; Jarmin & O'Hara, 2016), supporting the development of efficient policies and management within the public sector (Decker, 2014; Margetts & Sutcliffe, 2013), driving innovation within public policy and practice research (Pirog, 2014), and increasing the skills and competencies of the next generation of policy researchers (Lane, 2016). Despite the considerable promises and benefits to the field, an articulation of the predictive modeling methodology remains noticeably absent from the literature. To address this gap, this paper provides a detailed overview of the processes associated with developing, testing, and implementing predictive models, and identifies a collection of predictive modeling best practices for agencies to consider.

To achieve this purpose, this paper provides an illustrative case that details the process for developing and testing a collection of models that predict whether foster children would fail to achieve legal permanency. Establishing a permanent legal connection (i.e., "permanency") for children placed into foster care is a critical goal for child welfare agencies, as the failure to establish permanency can result in numerous negative emotional and intellectual effects for children (Freundlich, Avery, Munson, & Gerstenzang, 2006). Due to these negative consequences, child welfare policy over the past three decades has placed an increased focus on achieving permanency for children in foster care. This paper utilizes the 2013 Adoption and Foster Care Analysis and Reporting System (AFCARS) dataset[2] which provides administrative data on the national population of children that were discharged from state child welfare systems in 2013. A collection of nine distinct model types were developed and tested using the population of 233,633 foster care children that exited from care. The optimal model predicted whether children would achieve permanency with 97.66% accuracy.

This paper begins with reviews of the literatures on permanency and the use of predictive modeling within child welfare agencies. The paper then provides an overview of the methodologies used to develop the predictive models, and the associated results from predicting permanency. The concluding discussion identifies a collection of predictive modeling best practices.

2. Legal permanency and the use of predictive modeling by child welfare agencies

Achieving permanency continues to be an enduring challenge for state child welfare systems. Previous studies have found that nearly 1 in 10 foster care children lack legal ties to a permanent family (Craig & Herbert, 1997; Sheldon, 1997), and that over 20,000 children and youth annually exit foster care without permanency (U.S. Department of Health and Human Services Administration, 2017). Over the past several decades, child welfare policies and practice have focused on supporting and strengthening families to prevent the need to remove children from their homes (Pelton, 1991). However, the safety of children is paramount, and in those instances where a child's safety is threatened, child welfare agencies remove the child from the home and place him or her in a safe and stable environment. After removal, reunifying children with their families is the preferred outcome (Barth & Berry, 1987), but in some instances reunification may not be feasible, and it may be in the best interest of a child to remain in out of home care until a permanent legal connection with a parent or guardian is established. The Children's Bureau within the U.S. Department of Health & Human Services defines legal permanency as consisting of reunification with the child's parent or primary caretaker, living with other relatives, adoption, or guardianship. In contrast, reasons that children fail to achieve permanency include emancipating from state child welfare systems at the age of 18, or running away (Orsi, Lee, Winokur, & Pearson, 2017).

Establishing permanency is a critical task for child welfare agencies, as the failure to achieve permanency can have considerable adverse effects on children, including enduring difficulties in interacting with others, challenges in achieving independence, diminished academic, social and emotional development (Avery, 2010; Harden, 2004), and a decreased ability to effectively cope with stress (Freundlich et al., 2006). In turn, these effects can result in adverse outcomes that include failing to graduate from high school (Burley & Halpern, 2001), unemployment or underemployement (Courtney, Piliavin, Grogan-Kaylor, & Nesmith, 2001), and homelessness or incarceration (Keller, Cusick, & Courtney, 2007).

Given the critical importance of permanency, child welfare policy over the past three decades has sought to support permanency for children in foster care, though with mixed levels of success (Kemp & Bodonyi, 2002). The Adoption Assistance and Child Welfare Act of 1980 (P.L. 96-272) requires state child welfare agencies to engage in permanency planning and case plan reviews to ensure that foster care children are provided with a detailed plan for achieving permanency. The Multiethnic Placement Act (P.L. 103-382) in 1994 sought to remove obstacles to transracial adoptions with the goal of increasing permanency for minority children. Finally, the Adoption and Safe Families Act of 1997 (ASFA; P.L. 105-89) provided enhanced support for achieving permanency, by introducing a collection of reforms designed to increase the establishment of permanency in a timely manner.

The enhanced policy focus on establishing permanency has informed caseworker practice in notable ways. For instance, the decision-making timeframes established under ASFA (P.L. 105-89) has underscored the need for caseworkers to utilize efficient practices for promoting permanency (Smith & Donovan, 2003). However, this focus on efficiency can have an adverse effect where caseworkers narrowly focus on routine service completion and documentation (Tilbury, 2004) as opposed to spending their time conducting high quality contacts with clients or developing the effective caseworker-client relationships that are critical for achieving case goals (but that are not explicit requirements under the policy). Further complicating matters, child welfare practice has been considerably impacted by a combination of increasing caseloads (English & Pecora, 1994), diminishing organizational resources (Malatesta & Smith, 2014), and a lack of caseworker access to pertinent data for making informed decisions about permanency (Barth & Berry, 1987). Finally, the combined challenges of administrative decision making (Jun & Weare, 2010; Lindblom, 1959; Simon, 1957) and organizational resource constraints (Heinrich, 2002) has resulted in agency administrators and caseworkers facing difficult policy and practice questions about how to most efficiently and effectively deliver services to child welfare clients.

Predictive modeling can provide agencies with access to accurate information in a timely fashion, which can positively influence decision-making processes in a manner that improves client quality of life (Walker, Damanpour, & Devece, 2010) while also improving organizational efficacy (Bretschneider, 1990). Predictive modeling has become an increasingly popular tool over the past several decades with child welfare agencies developing, testing, and implementing various predictive modeling methodologies to enhance the decision-making processes utilized by caseworkers and administrators. These include models used to predict the likelihood that a child will be maltreated (Camasso & Jagannathan, 2000), the probability of child fatalities (Florida Department of Children and Families, 2014), and the risk of being reported for maltreatment at a young age (Putnam-Hornstein & Needell, 2011; Vaithianathan et al., 2013). These and other child welfare predictive models should be commended for employing rigorous,

[2] The AFCARS data used in this publication were made available by the National Data Archive on Child Abuse and Neglect, Cornell University, Ithaca, NY, and have been used with permission. Data from the Adoption and Foster Care Analysis and Reporting System (AFCARS) were originally collected by the Children's Bureau. Funding for the project was provided by the Children's Bureau, Administration on Children, Youth and Families, Administration for Children and Families, U.S. Department of Health and Human Services. The collector of the original data, the funder, the Archive, Cornell University and their agents or employees bear no responsibility for the analyses or interpretations presented here.

empirical approaches to modeling adverse child welfare outcomes. At the same time, predictive modeling methodologies have experienced notable advances in recent years, and coinciding advances in computing technology and the collection and use of administrative data have furthered the appeal and utility of the methodology. However, an articulation of the predictive modeling methodology remains noticeably absent from the literature. This paper addresses this critical gap by providing a detailed overview of the methodological approaches used to train and test predictive models.

3. Methodological approach

This section details the methodological processes associated with training and testing predictive models in three parts. The first part discusses the data sources for the models, the processes used to transform variables to optimize model performance and details the outcome and predictor variables used in the model development process. The second discusses the processes for training predictive models, including a taxonomy of the models used to predict permanency, the cross-validation process, and estimating model performance on the training set. The third part discusses the processes for testing predictive models, including evaluating model performance and optimal models for predicting permanency.

3.1. Data

The data source for the analysis consisted of the Administration for Children and Families' 2013 Adoption and Foster Care Analysis and Reporting System (AFCARS, 2013) foster care file. AFCARS is a federally mandated data collection system that requires states to collect data on all foster children who are served by a child welfare agency.[3] The 2013 dataset contains administrative data on the national population of 640,721 children that resided in the care of state child welfare systems, and includes 101 variables pertaining to the child, the child's family, the child's case, and the agency's administrative policy and practice.

The training and subsequent testing of predictive models was conducted using an exit cohort consisting of the subpopulation of 233,633 children and youth that exited state foster care systems in 2013. Child welfare research has often favored the use of entry cohorts and survival or time-to-event analyses (Courtney & Wong, 1996; Elgin, Sushinsky, Johnson, Russo, & Sewell, 2015), as these models incorporate censored cases that had yet to experience an event during the period of analysis into the estimated outcome probabilities. However, predictive modeling methodologies differ from this approach in important ways. Notably, predictive modeling leverages data on past events to train models that predict the likelihood of future events (James et al., 2013) and then utilizes an independent test set to assess the predictive accuracy of the models. Similar to many econometric models, the current generation of predictive models requires that all observations used to train and test predictive models have a value for the outcome of interest and that any observations with missing outcomes are excluded. The exit cohort consisting of 233,633 children and youth that completed their child welfare involvements provides an appropriate dataset with a strong degree of generalizability for training and testing rigorous models that accurately predict the likelihood of whether child welfare clients will achieve permanency. Critically, the dataset consists of children and youth from all 50 states and the District of Columbia, who spent anywhere from a single day to 20 years in the care of a state child welfare system, had diverse racial and sociodemographic characteristics, experienced an array of types of child abuse and neglect, and received services under diverse child welfare policies and practice. Collectively, this dataset offers a robust collection of observations and variables that can be used to train and test predictive models that accurately predict permanency for the broader population of children and youth that remained in the care of state child welfare systems.

In accordance with the Children's Bureau's permanency definition, a dichotomous variable was constructed to measure failures in establishing permanency. Children that were reunified with their parent(s) or a primary caretaker, living with another relative, adopted, or assigned a legal guardian, were coded as establishing permanency. Conversely, cases that ended due to emancipation, running away, or a fatality were coded as failing to establish permanency. Among the children exiting care in 2013, 89.71% achieved permanency while 10.29% (or, 24,040 children) failed to establish permanency. A failure to establish permanency was designated as the “event of interest” for the predictive models during the training and testing processes. This designation requires the models to focus on generating accurate predictions for children and youth with high-levels of risk for failing to establish permanency, and also has important methodological and practical implications for false positive and false negative predictions (as discussed in later sections of the paper).

A dichotomous permanency variable incorporates a broad operationalization of permanency. A review of the permanency research by Akin (2011) highlighted the limitations of utilizing a broad definition of permanency. More specifically, studies that grouped unique types of exits into a singular outcome have been acknowledged as informative but criticized for their limited ability to describe how predictive factors might influence dissimilar exits from child welfare systems (Courtney & Wong, 1996). Admittedly, there is a recognized lack of nuance in this paper's approach to combining the different reasons that children discharge from child welfare systems (e.g., emancipation, running away, or a death). Child welfare agencies may find it more beneficial to develop multiple models that predict specific adverse outcomes such as emancipation or death, though it is critical to note that resources (e.g., time, costs, and technology) are a critical concern when developing predictive models. In light of these resource considerations, this paper's use of a dichotomous permanency variable is appropriate as it provides an illustrative case of how agencies could utilize the federal definition of permanency to develop predictive models that accurately identify children and youth that are at-risk for adverse permanency outcomes. Under this type of approach, caseworkers and administrators could subsequently review the case files of the highest-risk clients to further assess if these clients are more likely to experience particular adverse outcomes.

3.1.1. Data preparation process

Prior to running the models, a 4-step data-preparation process associated with improved model performance was applied to the 101 predictor variables in the dataset. First, categorical variables were partitioned into dummy variables, with one dummy variable omitted to mitigate perfect collinearity within the models.[4] Under this step, observations with missing values for a predictor were coded as a separate “missing” category[5]. Next, continuous variables were centered and

[3] AFCARS data is comprised of case-level data that states and tribal agencies are required to submit to the Children's Bureau under federal regulations. On a biannual basis, states and tribal agencies are required to provide case-specific information on all children in foster care or who were adopted with title IV-E agency involvement.

[4] Within multicategorical variables, omitting a single category is required when estimating linear-based explanatory models. However, non-linear predictive models, such as neural networks, can achieve higher levels of performance when all categories are included within the model (Jensen, Qiu, & Ji, 1999). For the sake of consistency, this paper elected to utilize the same dataset, which consisted of reference categories being omitted from each multicategorical variable, when developing and testing each of the explanatory models.

[5] For instance, observations missing values for the predictor variable of whether a child received Title IV-D Child Support Funds, were coded as a separate “missing” category. While coding these observations as “missing” is an imperfect solution, it provides a more appropriate approach than replacing these observations with the mean or median value (a coarse approach to addressing missingness), conducting multiple imputation (which would significantly lengthen computing time for a dataset of this size) or dropping cases

scaled, as some predictive models have been shown to achieve higher levels of performance when variables are on a common scale (Hofmann & Gavin, 1998).[6] In the third step, 36 variables with minimal variance were removed, as these variables typically provide limited predictive power while also increasing model complexity and computational time (Guyon & Elisseeff, 2003). Finally, continuous predictor variables with correlation values above 0.75 were identified, as highly correlated predictors can increase model complexity while minimizing the overall interpretability of the model. Three predictor variables measuring the child's age at various points throughout their involvement with the child welfare system were identified as highly correlated. Of these three predictors, the child's age at the most recent removal was retained, while variables consisting of the child's age at the beginning and the end of the fiscal year were removed from the dataset. Upon completion of the data cleaning process, the dataset consisted of 189 predictor variables[7] pertaining to the child, the child's family, the child's case, and the agency's administrative policies and practices. A detailed list of the final set of predictor variables used in the model development process is included in Table 1.

The predictive models were developed and tested in R using the caret package (Kuhn, 2008). The process of developing, or "training" predictive models entails using a subset of the data (i.e., the "training set") to train or teach the models to estimate a function for accurately predicting the probability that an outcome will occur (James et al., 2013). A rigorous training process is characterized by the combination of resampling methods and the use of multiple models with differing methodological approaches. Resampling methods, such as k-fold cross-validation (Kohavi, 1995) are used to draw multiple subsamples from within the data and refit models to each sample in order to identify an optimal version of each predictive model. Upon completion of the training process, the optimal models are tested on a subset of the data excluded from the training process (i.e., the "test set" or the "validation set") to evaluate the accuracy of the predictive models. The methodological approach of utilizing an independent test set to assess predictive accuracy is a key distinguishing factor between the predictive modeling methodology and the long-standing use of traditional statistical models. For instance, while logistic regression and other linear-based models have long been used to generate predictions, the predictive modeling methodology employs an independent test set to more rigorously assess model accuracy, with models re-trained and re-tested until predictive accuracy has been maximized. Detailed discussions of the processes used to train and test the permanency predictive models follow below.

3.2. Training predictive models

After completing the data cleaning process, the dataset was subsequently partitioned into training and test sets. The training set was used to train the models to predict permanency, while a separate testing dataset was used to test the accuracy of the trained models. Partitioning the data into independent training and test sets reduces the likelihood of overfitting, a problem where a model is dependent on the unique patterns inherent to the dataset (Babyak, 2004; Domingos, 2012). Overfitted models generate predictions with a high degree of accuracy for the datasets they were trained on but generate considerably lower degrees of predictive accuracy when applied to other datasets. Thus, the data partitioning process produces a greater likelihood that the predictive model will be able to predict outcomes for new samples with a similar degree of accuracy.

Simple random sampling of the 233,633 observations was used to construct the training and tests sets. The training set used to estimate the performance of various model parameters consisted of 75% of the observations (175,248 children). The test set, consisting of the other 25% of the observations (58,415 children), was used to validate the models by providing an assessment of model accuracy that was independent of the data used to train the models. The training and test sets had a high degree of similarity across the predictor and outcome variables, and, importantly, the proportion of observations where permanency was not established was 10.29% within both datasets.

3.2.1. Taxonomy of models used to predict permanency

A collection of nine classification models were then independently trained to predict the permanency of children in the training set. The nine models utilized one of three different methodological approaches consisting of linear discriminant analysis, non-linear classification, and classification tree models (James et al., 2013; Kuhn & Johnson, 2013). The selection of an appropriate predictive model often depends upon whether the goal is inference, prediction, or a combination of the two (Athey, 2017; Breiman, 2001; Shmueli, 2010). In instances where inference is the goal, simple, inflexible predictive models are commonly preferred, as these models allow for greater interpretation of the association between the predictor variables and the outcome (Shmueli, 2010; Vaithianathan et al., 2013). In situations where generating accurate predictions is the end goal, more complex predictive models are often preferred (Delen, 2010; Rokach, 2016). These more complex models typically utilize a series of algorithms that generate higher levels of predictive accuracy at the expense of reduced interpretability.

The training set was used to train the nine predictive models with varying levels of interpretability. A model's degree of interpretability consists of the ability to ascertain the level of association between each predictor variable and the outcome variable. Along with the degree of interpretability, the computation time required to run a model is another critical factor for agencies to consider when deciding between predictive models. While advances in computation technology have significantly reduced computation time, training predictive models on large datasets can require significant time, costs, and technological resources. Accordingly, agencies will need to be mindful of the associated resource costs when training predictive models. Table 2 provides an overview of the models by methodological approach, level of interpretability, and the computation time associated with training each model on the entire training set using a robust resampling process.

As shown in Table 1, the level of interpretability and computation time can vary considerably across model types. Accordingly, these factors should be given important consideration throughout the processes of selecting, training, and testing predictive models. Linear discriminant analysis models, including logistic regressions, partial least squares discriminant analysis, and Elastic Net/Lasso models, utilize linear functions to categorize observations into groups based on predictor characteristics. These models commonly have a high degree of interpretability and require a low amount of computation time. Non-linear classification models, including neural networks, support vector machines, and multivariate adaptive regression splines (Friedman, 1991), utilize non-linear functions to categorize observations. These models are associated with low to moderate levels of interpretability, and typically have a higher computation time than linear discriminant models. Classification tree models, including classification trees, boosted trees, and random forests, utilize rules to partition observations into smaller homogenous groups. These models also differ in their degree of interpretability but commonly require a higher degree of computation time than the other model types.

3.2.2. Cross-validation process

Each of the nine predictive models were trained using the training

(*footnote continued*)
with missing observations (which would introduce significant bias on account of dropping a large number of observations).

[6] In some predictive models, such as partial least squares, un-centered and un-scaled variables can exert a considerable impact on the model, as variables with larger measurement scales will have a stronger influence on predicting outcomes. While centering and scaling variables is used to minimize the associated effects on model performance, there is an associated loss of interpretability due to the transformation of the variables.

[7] The increase from 101 to 189 predictor variables is attributed to the partitioning of categorical variables into multiple dummy variables.

Table 1
Variables Included in the Predictive Models.[a]

Both of Child's Parents Have Relinquished their Parental Rights (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child's Age at Latest Removal from the Home (Integer Variable)
Child's Derived Race (6 Dichotomous Variables): White, Black or African American, American Indian or Alaska Native, Asian, Hawaiian or Other Pacific Islander, More than One Race ("Race Unknown" category excluded)
Child's Derived Race and Ethnicity (7 Dichotomous Variables): Non-Hispanic - White, Non-Hispanic - Black, Non-Hispanic - American Indian Alaska Native, Non-Hispanic - Asian, Non-Hispanic - Hawaiian/Other Pacific Islander, Non-Hispanic - More than One Race, Non-Hispanic - Any Race ("Race/Ethnicity Unknown" category excluded)
Child Entered Foster Care During the Fiscal Year (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child had Clinically Diagnosed Disability (3 Dichotomous Variables): Yes, No, Not Yet Determined ("Missing" category excluded)
Child Clinically Diagnosed as Emotionally Disturbed (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child Clinically Diagnosed as Requiring Other Medical Care (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child was Ever Adopted (3 Dichotomous Variables): Not Applicable, Yes- child has been legally adopted, No- Child has never been legally adopted ("Unable to Determine" category excluded)
Child was Ever Adopted - Age at Adoption (Integer Variable)
Child was in Foster Care at the Beginning of the Federal Fiscal Year (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child's Most Recent Placement Setting (8 Dichotomous Variables): Pre-Adoptive Home, Foster Home- Relative, Foster Home- Nonrelative, Group Home, Institution, Supervised Independent Living, Runaway, Trial Home Visit ("Missing" Category Excluded)
Child's Ethnicity - Hispanic Origin (3 Dichotomous Variables): Yes, No, Unable to Determine ("Not Applicable" category excluded)
Child's Race - American Indian/Alaska Native: Yes, No
Child's Race - Black/African American: Yes, No
Child's Race - Unable to Determine (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Child's Race - White: Yes, No
Child Sex: Female, Male
First Foster Caretaker's Ethnicity- Hispanic Origin (3 Dichotomous Variables): Not Applicable, Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - American Indian/Alaska Native (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - Asian (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - Black/African American (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - Hawaiian/Pacific Islander (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - Unable to Determine (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Foster Caretaker's Race - White (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
First Principal Caretaker's Year of Birth (Integer Variable)
Foster Family Structure (5 Dichotomous Variables): Not Applicable, Married Couple, Unmarried Couple, Single Female, Single Male ("Unable to Determine" category excluded)
Length (Days) in Current Placement Setting (Integer Variable)
Length (Days) Since Latest Removal (Integer Variable)
Most Recent Case Plan Goal (7 Dichotomous Variables): Reunify with Parent/Principal Caretaker, Live with Other Relative(s), Adoption, Long-Term Foster Care, Emancipation, Guardianship, Case Plan Goal Not Yet Established ("Missing" category excluded)
Number of Placement Settings During the Current FC Episode (Integer Variable)
Principal Caretaker Family Structure (5 Dichotomous Variables): Married Couple, Unmarried Couple, Single Female, Single Male, Unable to Determine ("Not applicable" category excluded)
Removal Reason- Caretaker Inability to Cope (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Child Behavioral Problem (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Inadequate Housing (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Neglect (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Parental Alcohol Abuse (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Parental Drug Abuse (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Parental Incarceration (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Physical Abuse (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Removal Reason- Sexual Abuse (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Reporting Date Month: September
Second Foster Caretaker's Ethnicity- Hispanic Origin (3 Dichotomous Variables): Not Applicable, Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - American Indian/Alaska Native (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - Asian (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - Black/African American (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - Hawaiian/Pacific Islander (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - Unable to Determine (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Second Foster Caretaker's Race - White (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Social Security Act Benefits (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
State Child Welfare System (50 Dichotomous Variables): AK, AL, AR, CA, CO, CT, DC, DE, FL, GA, HI, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NH, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WI, WV, WY
State Support Received by Child (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Title IV-A AFDC Payment (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Title IV-D Child Support Funds (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Title IV-E Foster Care Payments Made on Behalf of Child (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Title IV-E Adoption Assistance (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Title XIX Eligibility for Medical Assistance (2 Dichotomous Variables): Yes, No ("Missing" category excluded)
Total Days in Foster Care, All Episodes (Integer Variable)
Urban Rural Continuum Code (8 Dichotomous Variables): Metro: > 1 Million Population, Metro: 250 K to 1 Million Population, Metro: < 250 K Population, Non-Metro: Urban > 20 K Population - Adjacent, Non-Metro: Urban > 20 K Population - Nonadjacent, Non-Metro: 2.5 K to 20 K - Adjacent, Non-Metro: Urban 2.5 K to 20 K - Nonadjacent, Rural or < 2.5 K Population - Adjacent, ("Rural or < 2.5 K Population - Nonadjacent" category excluded)

Note: parenthetical references to "category excluded" indicate the dummy variables that were omitted during the partitioning of categorical variables to mitigate perfect collinearity within the predictive models.

[a] Metadata, including detailed definitions for each of the variables identified in the table above can be accessed via the following resource: National Data Archive on Child Abuse and Neglect. (2017). AFCARS Foster Care File Code Book. Retrieved from: https://www.ndacan.cornell.edu/datasets/pdfs_user_guides/AFCARSFosterCareCodebook.pdf.

Table 2
Overview of models used to predict permanency.

Model type	Model	Interpretability (Ability to ascertain the level of association between each predictor variable and the outcome variable)	Computation Time (in Hours)[a]
Linear discriminant analysis models	Logistic regression Brief description: A regression model that utilizes the logarithm of the odds to calculate the probability of a given outcome.	High	1.82
	Partial least squares discriminant analysis Brief description: A regression model that utilizes a dimension reduction strategy to relate a collection of predictor variables to an outcome variable.	High	0.02
	Elastic Net/Lasso Brief description: A regression model that utilizes variable selection and regularization to maximize model accuracy.	High	26.49
Non-linear classification models	Neural networks Brief description: A model resembling the physiological structure of the human brain or nervous system that utilizes multiple algorithms to process pieces of information.	Low	396.35
	Support vector machines Brief description: A model that identifies an optimal hyperplane to separate observations into distinct categories.	Low	45.35
	Multivariate adaptive regression splines Brief description: A model that fits separate piecewise linear segments (splines) to model the relationship between the predictor variables and the outcome.	Moderate	8.99
Classification tree models	Classification tree Brief description: A simple prediction model is fit by recursively partitioning data into progressively smaller, homogenous groups.	High	0.15
	Boosted trees Brief description: A model that builds upon traditional classification tree models by fitting a series of independent decision trees which are aggregated to form a single predictive model.	Low	50.93
	Random forest Brief description: A model that builds upon traditional classification tree models by utilizing bootstrapping methods to build a collection of decision trees.	Low	176.72

[a] Computation time corresponds to the time associated with running each model on a cloud computing platform with parallel processing. Parallel processing was used to reduce computing time by splitting computation tasks into smaller parts that could be executed simultaneously on multiple processors. While parallel processing significantly reduced computation time, several features of the training process increased computation time. These included the considerable size of the training set, the k-fold cross validation process, and the process of training and tuning each model across multiple parameters.

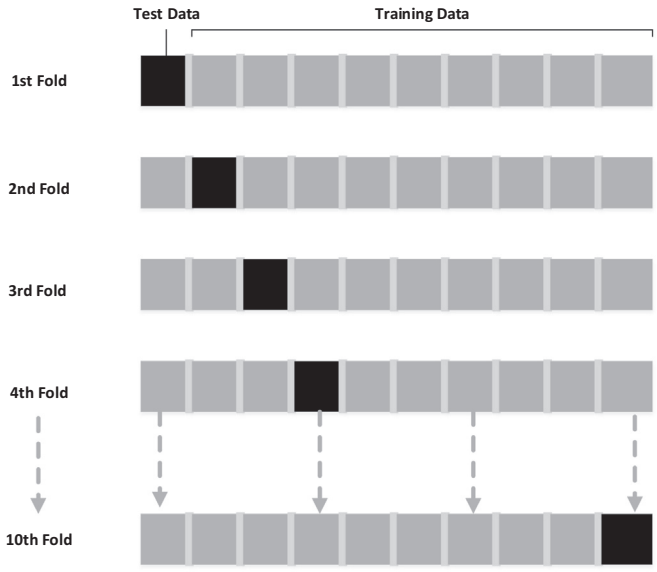

Fig. 1. Example of a single-iteration of a 10-fold cross-validation.

set of 175,248 observations. The training process utilized k-fold cross-validation methods (Arlot & Celisse, 2010; Kohavi, 1995) to resample the data to estimate multiple model parameters and identify an optimal version for each of the nine models. While k-fold cross validation and other re-sampling methods can considerably increase computation time, these methods play a critical role in fitting and evaluating predictive models. Increasing the number of folds and repetitions used in the validation process increases the overall number of subsamples used to estimate model performance while simultaneously decreasing the differences in the size of the training set and the subsamples. This in turn reduces the overall level of bias inherent within the training data, and K-fold cross validation methods have been shown to yield test errors rates with minimal bias and variance (James et al., 2013).

Under the k-Fold Cross-Validation process, the observations in the training set were randomly partitioned into ten folds, or subsets, of the same size. As depicted in Fig. 1, each model was individually fit using nine of the folds, with the first fold serving as an independent test set for predicting the permanency outcomes and estimating model performance. In the next step, the first fold was returned to the training set and the procedure was repeated with the second fold held out and used as a test set, with the process repeating until each of the ten folds were held out.

The ten-fold cross-validation was repeated five times resulting in a total of 50 random folds being used to estimate model performance. The results for all folds were averaged to obtain a k-fold cross validation estimate that was subsequently used to estimate model performance.

3.2.3. Estimating model performance on the training set

In the final step of the training process, model performance was estimated for each of the nine predictive models. This process consisted of utilizing the k-fold cross validation estimates and Receiver Operating Characteristics (ROC) curves to quantitatively assess model performance and identify an optimal version of each of the nine predictive models. ROC curves compare the sensitivity and specificity rates of a

Table 3
Average Sensitivity, Specificity, and ROC Values for the Nine Predictive Models.

Model type	Model	Sensitivity	Specificity	ROC
Classification tree model	Random forest	0.891	0.986	0.991
Non-linear classification model	Neural network	0.876	0.989	0.991
Classification tree model	Boosted tree	0.884	0.987	0.990
Non-linear classification model	Multivariate adaptive regression spline	0.778	0.987	0.984
Non-linear classification model	Support vector machine	0.774	0.990	0.982
Linear discriminant analysis model	Elastic Net/Lasso	0.723	0.991	0.980
Linear discriminant analysis model	Partial least squares discriminant analysis	0.449	0.996	0.969
Classification tree model	Classification tree	0.844	0.978	0.948
Linear discriminant analysis model	Logistic regression	0.790	0.977	0.883

model's predictions (Bradley, 1997). A model's sensitivity is the rate, measured on a scale of zero to one, that the event of interest (i.e., failure to establish permanency) is correctly predicted for all observations in which the event occurred. In contrast, a model's specificity, measured on a similar scale, is the rate that a nonevent (i.e., establishing permanency) is correctly predicted among all observations with a nonevent. A model's overall level of predictive accuracy typically involves a tradeoff between sensitivity and specificity, with efforts to increase the accuracy of predicting events (sensitivity) occurring at the expense of the prediction accuracy of non-events (specificity).

ROC curves provide one approach for evaluating the sensitivity-specificity tradeoff, by plotting the sensitivity and specificity of class probabilities across a series of thresholds. The area under the ROC curve (i.e., ROC value) is a commonly used metric for evaluating model performance (Hastie, Tibshirani, & Friedman, 2009), with values near zero indicating low levels of predictive accuracy and values approaching one having optimal accuracy. Table 3 provides the ROC value, sensitivity, and specificity values for each of the nine models.

The sensitivity and specificity results show the variance in predicting permanency across the models. Considerable variance existed in the sensitivity of models for predicting cases that failed to achieve permanency, with values ranging from 0.449 to 0.891. In contrast, the nine models exhibited considerably higher specificity rates for predicting cases that achieved permanency, with values ranging from 0.977 to 0.996. The lower sensitivity values can be primarily attributed to the characteristics of cases that failed to establish permanency, which occur less frequently and are less likely to exhibit easily discernable patterns. This variance in sensitivity and specificity further underscores the importance of using ROC values to comprehensively evaluate model performance.

The ROC values for the nine models ranged from 0.883 to 0.991. Among the models, logistic regression, classification trees, and partial least squares discriminant analysis had the lowest levels of performance, while elastic net, support vector machines, and multivariate adaptive regression splines had moderate levels. Three models, boosted trees, neural networks, and random forests had the highest levels of performance, with ROC values between 0.990 and 0.991. These models exhibited optimal performance in predicting both non-permanency and permanency outcomes, with minimum sensitivity values of 0.876 and minimum specificity values of 0.986.

4. Testing predictive models

Upon completion of the training phase, the optimal versions of each of the nine models were applied to the test set to generate permanency predictions for each of the of 58,415 children.[8] The prediction accuracy of each model was evaluated using confusion matrices, which consist of a two-by-two matrix used to show prediction accuracy for two-class problems. While ROC curves provide an appropriate metric for evaluating a model's accuracy during the training process, confusion matrices provide a greater level of detail for evaluating the model's predictive accuracy.

		Observed	
		Non-Permanency	*Permanency*
Predicted	*Non-Permanency*	**True Positive**	False Positive
	Permanency	False Negative	**True Negative**

Fig. 2. Confusion matrix for evaluating model performance.

As shown in Fig. 2, the confusion matrix provides a cross-tabulation of the predicted classes with the rows consisting of predictions for events and non-events (i.e., failure to establish permanency and establishing permanency, respectively), while the columns list the observed events and non-events. The upper left quadrant of each matrix lists the number of true positives, where the models accurately predicted that a child would not establish permanency, while the upper right quadrant lists the number of false positives. The bottom left quadrant of each matrix lists the false negative events while the lower right quadrant lists the true negative events where the models accurately predicted that a child would achieve permanency. Among the four quadrants, the emphasis is on maximizing the number of observations in the true positive (upper left) and true negative (lower right) quadrants. An overall level of accuracy is then calculated by dividing the number of observations in these two quadrants by the total number of observations in the test set.

Confusion matrices provide agencies with an important opportunity to examine the model's accuracy in predicting various types of cases, which can yield important policy and management implications. More specifically, a false negative prediction could impose greater costs for the child and the agency. A child that was falsely predicted to achieve permanency would not be accurately identified and would be subsequently less likely to receive targeted interventions and resources that could help establish permanency. In contrast, a greater number of false positive predictions could impose a considerable burden as child welfare agencies would expend critical resources on cases that would have otherwise achieved permanency.

Table 4 presents the confusion matrices for the nine models. The overall accuracy rates of these models ranged from 93.95% for the partial least squares discriminant analysis to 97.66% for the boosted tree model. Notable differences are evident regarding the comparative accuracy of the three methodological approaches. Linear-based models exhibited the lowest levels of accuracy in predicting permanency, with partial least squares discriminant analysis, logistic regression, and the elastic net/lasso model comprising three of the four lowest performing models. Non-linear models, in contrast, fared marginally better with support vector machines and multivariate adaptive regression splines demonstrating higher levels of performance. Classification tree models had the highest levels of performance, with boosted trees and random

[8] The number of children within the test set that failed to achieve permanency was proportional to the broader population, with 10.29% of children failing to achieve permanency within both groups.

Table 4
Confusion matrices for the permanency predictive models.

Model type	Model	Confusion matrices			Combined accuracy	95% Confidence interval
Classification tree	Boosted tree		Non-permanency	Permanency	97.66%	97.54%–97.79%
		Non-permanency	5333	687		
		Permanency	677	51,718		
Classification tree	Random forest		Non-permanency	Permanency		
		Non-permanency	5360	736	97.63%	97.50%–97.80%
		Permanency	650	51,669		
Non-linear classification	Neural network		Non-permanency	Permanency	97.63%	97.51%–97.76%
		Non-permanency	5259	631		
		Permanency	751	51,774		
Non-linear classification	Support vector machine		Non-permanency	Permanency	96.72%	96.57%–96.86%
		Non-permanency	4632	537		
		Permanency	1378	51,868		
Non-linear classification	Multivariate adaptive regression spline		Non-permanency	Permanency	96.49%	96.34%–96.64%
		Non-permanency	4665	706		
		Permanency	1345	51,699		
Linear discriminant analysis	Elastic Net/Lasso		Non-permanency	Permanency	96.29%	96.14%–96.44%
		Non-permanency	4337	493		
		Permanency	1673	51,912		
Classification tree	Classification tree		Non-permanency	Permanency	96.26%	96.10%–96.41%
		Non-permanency	5132	1306		
		Permanency	878	51,099		
Linear discriminant analysis	Logistic regression		Non-permanency	Permanency	95.83%	95.67%–95.99%
		Non-permanency	5428	1852		
		Permanency	582	50,553		
Linear discriminant analysis	Partial least squares discriminant analysis		Non-permanency	Permanency	93.95%	93.75%–94.14%
		Non-permanency	2691	217		
		Permanency	3319	52,188		

forests achieving the highest accuracy rates.

4.1. Optimal models for predicting permanency

The neural network, boosted tree, and random forest models had notably higher levels of predictive accuracy (note: detailed descriptions of the models and the associated parameters are included within the Appendix A). The combined accuracy rate of the three models was nearly identical, ranging from 97.63% for the neural network and random forest models to 97.66% for the boosted tree model. The neural network and random forest models achieved similar levels of accuracy but differed as to whether this was accomplished by maximizing model sensitivity or specificity. The neural network model maximized specificity, achieving a higher rate of accuracy in predicting instances where permanency was achieved while the random forest model maximized sensitivity, with a higher accuracy in predicting non-permanency cases.

The boosted tree model achieved a slightly higher combined accuracy rate of 97.66% (with a 95% confidence interval of 97.54% to 97.79%). Among the nine models, the boosted tree model had the highest balance between sensitivity and specificity, thereby minimizing the number of false negatives and false positives. In comparison to the other models, implementation of the boosted tree model would have the most significant implications for policy and practice. Agency leadership would be provided with increased confidence that the model had maximized predictive accuracy and would allow caseworkers to more effectively leverage the agency's resources to engage in increased efforts to establish permanency. At the same time, leadership could be further reassured that the model had sufficiently minimized the error rate and the associated costs of generating false predictions.

The high levels of predictive accuracy obtained by these three models comes with a notable tradeoff. Boosted trees, random forests, and neural networks typically achieve among the highest levels of accuracy but rely upon a complex series of algorithms that sacrifice interpretability to achieve greater accuracy. Due to this tradeoff, assessing the significance of the model's predictors becomes a complex and arduous task. Methods for assessing the significance of predictors exist, such as the caret package's Variable Importance function (Kuhn, 2008), but the application of these methods is far from universal and additional research is needed to assess the robustness of these methods. Given these limitations, linear and simple classification tree models, which provide a clearer relationship between predictors and the outcome, may be of greater utility to child welfare agencies. Using these models, agency staff could focus their attention on understanding significant predictors and then utilize performance management or continuous quality improvement methods to address the systemic impact of these predictors on adverse permanency outcomes. As discussed further in the following section, agencies should be mindful of whether their goals for developing predictive models are accurate predictions, inference, or some combination of the two.

5. Discussion

Predictive modeling is an increasingly important methodological approach that allows agencies to leverage their available data to conduct rigorous analysis to more effectively target their limited resources towards at-risk clients. This improved ability to accurately identify the highest-risk clients can allow agencies to engage in enhanced service delivery, such as conducting detailed case file reviews to identify client-specific barriers and associated services for preventing adverse outcomes. This increased ability to accurately identify and serve at-risk clients could play a critical role in supporting agencies in their pursuit of the enduring public administration goals of increased effectiveness, efficiency, and equity (Andrews & Entwistle, 2010; Frederickson, 1997; Light, 1998). At the same time, testing and implementing predictive models can present a daunting and arduous task for many agencies, due to the need for agency staff or contractors to possess substantial knowledge of predictive modeling methodologies, as well as an intimate knowledge of child welfare administrative data, and a strong familiarity with advanced statistical packages. Given these substantial requirements, this paper identifies a collection of best practices that may be helpful to agencies as they develop and implement predictive models.

5.1. Predictive models can improve upon, but not replace, traditional decision-making processes within agencies

Predictive models are rightfully lauded for their ability to process vast amounts of data to estimate the probability that an outcome will occur. The development of effective predictive models is heavily influenced by the insight and experiences of staff with expert knowledge and understanding of the problems being modeled (Kuhn & Johnson, 2013). While the resulting predictive models can estimate the probability of an outcome with a high degree of accuracy, the models cannot understand the clinical decision-making processes utilized by caseworkers and administrators, the associated policy and management implications, nor the relative costs and benefits associated with these probabilities (Russell, 2015). Accordingly, agencies should be mindful that predictive models are not a panacea nor a replacement for traditional decision-making processes. Rather, predictive models can serve as an important tool that can complement clinical decision making and support agencies in more effectively leveraging their resources to improve client outcomes. Within this context, rigorous training and testing of predictive models can provide agency staff with an efficient and effective process for accurately identifying at-risk clients. Agency staff can then run validated predictive models on a regular basis (e.g., monthly, quarterly, or as agency needs and resources permit) to obtain dynamic predictions that reflect the changes in a client's case.

5.2. Agencies should promote transparency by clearly articulating the methodological approach and the predictive accuracy of their models

Whether utilizing linear, nonlinear, or classification models, agencies should integrate methodological transparency throughout the predictive modeling process. Notably, agencies should be especially cognizant of the tradeoffs between interpretability and predictive accuracy. While models with a high degree of interpretability provide a clear relationship between how the predictors influence the outcome, less interpretable models utilize more of a “black box” approach where a complex series of algorithms are used to generate an outcome prediction. Providing internal stakeholders (such as caseworkers and administrators) and external stakeholders (such as child welfare clients, their families, and the broader community) with a detailed understanding of how an agency's predictive model utilizes the associated predictors to calculate probabilities and predict outcomes will play a critical role in the subsequent use and overall effectiveness of the models for improving policy and practice.

A model's performance on the test set is another area that requires agencies to provide clear, detailed information. The model's overall rate of accuracy, as well as its associated accuracy rates for predicting events and non-events, should be clearly articulated to provide stakeholders with a clear understanding of the areas where the models have a high degree of performance, as well as those areas in which the models under-perform. Clear articulation of model performance allows for greater evaluation of a model's internal and external validity and may help agencies to identify areas where a model can be further improved.

5.3. The predictors comprising predictive models should be interpreted cautiously

Predictive models are often developed with the distinct goals of prediction, causal inference, or a combination of the two (Athey, 2017; James et al., 2013; Shmueli, 2010). While recent research has sought to increasingly utilize predictive models to obtain causal inference (Grimmer, 2015; Wager & Athey, 2017), the majority of models are trained and tested with the goal of maximizing predictive accuracy. As demonstrated by the models used in this paper, the interpretability of predictive models can vary greatly, with linear and basic classification tree models providing a clear relationship between predictors and the outcome, while nonlinear and more sophisticated tree models rely upon a complex series of algorithms that often sacrifice interpretability for the sake of accuracy.

Furthermore, many of the methods used to bolster model accuracy make interpreting the model's predictors an exceptionally demanding task. For instance, the accuracy of predictive models can often be significantly bolstered through the inclusion of a large number of predictors, which can minimize the ability to ascertain causal relationships. In addition, methods that standardize predictors within the models, such as centering and scaling variables, can further minimize model interpretability and causal inference. Due to these factors, agencies should be clear about whether their intended goals in training predictive models are accurate predictions, inference, or a combination of the two. In instances where prediction is a prominent goal, agencies should be cognizant of the methods used to increase accuracy, any tradeoffs between interpretability and accuracy within the models, and exercise considerable caution when interpreting the relationship of the predictors within the models.

5.4. Agencies should consider opportunities for incorporating community engagement into the predictive modeling process

While predictive modeling provides a valuable approach for helping agencies to utilize their limited resources more effectively, the methodology also raises notable ethical and legal concerns (Cohen et al., 2014) related to privacy, equitable representation, access, and transparency (Bovens & Zouridis, 2002; Meijer, 2013). These ethical and legal concerns can produce justifiable apprehension among stakeholders that are less familiar with predictive models and the associated methodological strengths and weaknesses. As administrators have ethical and legal obligations to involve the public in decision-making processes, community engagement can provide a mechanism to make predictive models transparent while also enhancing agency accountability (Bingham, Sandfort, & O'Leary, 2008). Citizen access to government information has been viewed as an essential component of the democratic process (Jaeger & Bertot, 2010), and multiple opportunities exist for engaging the community to solicit stakeholder feedback throughout the process of developing and implementing predictive models. These opportunities include establishing an ethical framework and committee or providing other venues for soliciting and incorporating stakeholder feedback across key components of the predictive modeling process, including the initial design of the predictive models, the processes associated with the collection of data, interpreting findings, assessing model performance, and the processes by which the model results will be used to inform future interventions and decision-making processes. Collectively, these opportunities for incorporating stakeholder feedback can empower and educate community members, further inform public agencies and managers, and subsequently improve administrative processes (King, Feltey, & Susel, 1998).

6. Limitations

This paper provides an illustrative case for training and testing predictive models, and the findings should be considered with an appreciation of the study limitations. Notably, the predictive modeling methodology differs from other methodologies commonly applied within the child welfare literature, including time-to-event models (Courtney & Wong, 1996; DePanfilis & Zuravin, 1999). While time-to-event models incorporate censored cases into the estimation of outcome probabilities, predictive models exclude censored cases from the modeling process. The extent that the findings from predictive models differ from time-to-event models is a question worthy of greater consideration, and future research will benefit from an improved understanding of the inherent differences, strengths, and limitations of these methodologies. The increasingly prevalent application of the predictive modeling methodology and the coinciding advances in computing

technology and access to administrative data, will ensure that ample opportunities exist for these future research efforts.

A second study limitation is that the permanency outcome variable has both policy and practice nuances that warrant greater attention. As described in the Methodological Approach section, the permanency outcome variable collapsed multiple outcomes into a single construct. This grouping of unique exit reasons into "permanency" or "non-permanency" has been recognized as informative but also criticized for offering a decreased ability to understand dissimilar exits and how predictive factors influence them (Courtney & Wong, 1996). This paper leveraged an illustrative case and a national administrative dataset to provide an improved understanding of the process by which a methodologically diverse collection of predictive models is designed, tested, and implemented. For the sake of simplicity, this paper elected to train and test nine predictive models to generate predictions utilizing the Children's Bureau's definition of permanency. An alternative approach would have been to repeat the same modeling processes for specific types of permanency outcomes (e.g., reunification, adoption, emancipation, fatalities, etc.) but this would have introduced considerable complexity and diminished the paper's ability to provide a concise articulation of the predictive modeling methodology that has been noticeably absent from the literature. Future research will benefit from subsequent applications of the methodology to model specific types of child welfare outcomes.

A final limitation pertains to this paper's lack of a detailed discussion centered on interpreting the highest performing predictive models and assessing the relative influence of each of the variables in predicting permanency. As discussed throughout this paper, predictive models often trade interpretability for the sake of maximizing accuracy, and vice versa. Methods for assessing the significance of predictors across models exist (Kuhn, 2008) but the application of these methods is far from widespread and further research is needed to assess the robustness of these methods, including their inherent strengths and limitations. These are critical issues to be carefully considered in future research (Athey, 2017), and subsequent predictive models will benefit from careful attention to better understanding the relative influence of the variables in predicting adverse events with a high rate of accuracy.

7. Conclusion

Predictive modeling offers an increasingly important methodological approach that allows agencies to more effectively target their finite resources towards clients that are at-risk for adverse events. As applications of predictive modeling are noticeably absent from the literature, this paper has sought to provide an improved understanding of the process by which predictive models are designed, tested, and implemented. Leveraging an illustrative case with a national administrative dataset, this paper detailed the methodological processes associated with training and testing nine types of models used to predict whether foster care children would achieve legal permanency. The resulting models predicted permanency with a high degree of accuracy, with the optimal model achieving a 97.66% accuracy rate. Given the considerable potential of predictive models, but the nascent state of the literature, this paper identified a collection of best practices detailing how agencies can more effectively utilize predictive modeling. Adopting these best practices will allow agencies to develop methodologically rigorous models that meet scientific standards and incorporate community engagement, while allowing child welfare agencies to serve at-risk clients in a more effective, efficient, and equitable manner.

Conflicts of interest

The author has no conflicts of interest to report.

Acknowledgement

The author would like to thank the three anonymous reviewers for providing highly constructive and helpful comments on the paper. In addition, the author is grateful to Steve Garasky and Rebecca Orsi for providing sage comments on earlier drafts.

Appendix A. Detailed overview of the highest-performing models

Given the higher levels of performance for the boosted tree, random forest and neural network models, brief descriptions of each model's methodological approach are provided. Table A1 provides the confusion matrices and the tuning parameters for each of the models.

Table A1
Confusion matrices and tuning parameters for highest-performing models.

Boosted tree
Tuning parameters:
- Maximum Number of Trees: 500
- Interaction Depth: 10
- Default Shrinkage Rate: 0.10
- Minimum Number of Observations in Terminal Nodes: 10

	Non-Permanency	Permanency
Non-Permanency	5333	687
Permanency	677	51,718

Combined Accuracy (95% Confidence Interval): 97.66% (97.54%–97.79%)

Random Forest
Tuning Parameters:
- Number of Predictors Sampled for Splitting at Each Node: 64

	Non-Permanency	Permanency
Non-Permanency	5360	736
Permanency	650	51,669

(*continued on next page*)

Table A1 (*continued*)

Combined Accuracy (95% Confidence Interval): 97.63% (97.50%–97.80%)

Neural Network
Tuning Parameters:
- Number of Hidden Units: 10
- Weight Decay: 2

	Non-Permanency	Permanency
Non-Permanency	5259	631
Permanency	751	51,774

Combined Accuracy (95% Confidence Interval): 97.63% (97.51%–97.76%)

Neural networks are an increasingly prominent predictive modeling approach that have often been characterized as resembling the physiological structure of the human brain or nervous system due to the model's use of multiple layers (or algorithms) for processing information (Lantz, 2013). Each layer of a neural network is responsible for processing a different piece of information, which is subsequently passed on to inform additional layers. The layers are comprised of a collection of artificial neurons consisting of input, hidden, and output units. Information is received by the input unit, passed along to the hidden unit for processing, and then passed on to the output unit, which connects to other input units.

In contrast, boosted tree models build upon traditional classification trees, which consist of a series of nested if-then statements that partition the data into smaller homogenous groups. Unlike traditional classification trees, boosted tree models fit a series of independent decision trees and then aggregate the trees to form a single predictive model. The process of building a boosted tree model involves sequentially growing a series of classification trees, with each tree using information from previously grown trees to avoid overfitting of the training data.

Finally, random forest models build upon traditional classification tree models by utilizing bootstrapping methods to build a collection of decision trees. At each point where a split in a tree is considered, a random sample of predictors are chosen as potential candidates for the split. This process of considering a smaller subset of predictors minimizes the likelihood of a high degree of correlation among multiple trees.

References

AFCARS (2013). Annual foster care and adoption reporting system report. Retrieved from: http://www.acf.hhs.gov/sites/default/files/cb/afcarsreport20.pdf.

Akin, B. A. (2011). Predictors of foster care exits to permanency: A competing risks analysis of reunification, guardianship, and adoption. *Children and Youth Services Review, 33*(6), 999–1011.

Andrews, R., & Entwistle, T. (2010). Does cross-sectoral partnership deliver? An empirical exploration of public service effectiveness, efficiency, and equity. *Journal of Public Administration Research and Theory, 20*(3), 679–701.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*(6324), 483–485.

Avery, R. J. (2010). An examination of theory and promising practice for achieving permanency for teens before they age out of foster care. *Children and Youth Services Review, 32*(3), 399–408.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421.

Barth, R. P., & Berry, M. (1987). Outcomes of child welfare services under permanency planning. *Social Service Review, 61*(1), 71–90.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613.

Bingham, L. B., Sandfort, J., & O'Leary, R. (2008). In L. B. Bingham, & R. O'Leary (Eds.). *Big ideas in collaborative public management* (pp. 270–285). Ajmonk, NY: M.E. Sharpe.

Bovens, M., & Zouridis, S. (2002). From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review, 62*(2), 174–184.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Bretschneider, S. (1990). Management information systems in public and private organizations: An empirical test. *Public Administration Review,* 536–545.

Burley, M., & Halpern, M. (2001). *It's my life: A framework for youth transitioning from foster care to successful adulthood.* Seattle, WA.

Camasso, M. J., & Jagannathan, R. (2000). Modeling the reliability and predictive validity of risk assessment in child protective services. *Children and Youth Services Review, 22*(11), 873–896.

Clarke, A., & Margetts, H. (2014). Governments and citizens getting to know each other? Open, closed, and big data in public management reform. *Policy & Internet, 6*(4), 393–417.

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs, 33*(7), 1139–1147.

Cook, T. D. (2014). "Big data" in research on social policy. *Journal of Policy Analysis and Management, 33*(2), 544–547.

Courtney, M. E., Piliavin, I., Grogan-Kaylor, A., & Nesmith, A. (2001). Foster youth transitions to adulthood: A longitudinal view of youth leaving care. *Child Welfare, 80*(6), 685.

Courtney, M. E., & Wong, Y. L. I. (1996). Comparing the timing of exits from substitute care. *Children and Youth Services Review, 18*(4–5), 307–334.

Craig, C., & Herbert, D. (1997). *The state of the children: An examination of government-run foster care (NCPA policy report no. 210).* Dallas, TX: National Center for Policy Analysis.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review, 79*, 291–298.

Decker, P. T. (2014). Presidential address: False choices, policy framing, and the promise of "Big Data". *Journal of Policy Analysis and Management, 33*(2), 252–262.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506.

DePanfilis, D., & Zuravin, S. J. (1999). Predicting child maltreatment recurrences during treatment. *Child Abuse & Neglect, 23*(8), 729–743.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87.

Elgin, D. J., Sushinsky, J., Johnson, A., Russo, G., & Sewell, T. (2015). Factors affecting permanency for legally free children & youth: A study of Colorado's legally free population across age groups, 2008–2014. *Children and Youth Services Review, 57*, 60–67.

English, D. J., & Pecora, P. J. (1994). Risk assessment as a practice method in child protective services. *Child Welfare, 73*(5), 451.

Florida Department of Children and Families (2014). *Child fatality trend analysis: January 1, 2007 through June 30, 2013. Tallahassee, FL.* Author.

Frederickson, H. G. (1997). *The spirit of public administration.* San Francisco, CA: Jossey-Bass Incorporated Pub.

Freundlich, M., Avery, R. J., Munson, S., & Gerstenzang, S. (2006). The meaning of permanency in child welfare: Multiple stakeholder perspectives. *Children and Youth Services Review, 28*(7), 741–760.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67.

Geisser, S. (1993). *Predictive inference. Vol. 55*. New York, NY: CRC Press.

Gillingham, P. (2015). Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the 'black box' of machine learning. *The British Journal of Social Work, 46*(4), 1044–1058.

Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics, 48*(01), 80–83.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182.

Harden, B. J. (2004). Safety and stability for foster children: A developmental perspective. *The Future of Children,* 31–47.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning.* New York, NY: Springer.

Heinrich, C. J. (2002). Outcomes–based performance management in the public sector: Implications for government accountability and effectiveness. *Public Administration Review, 62*(6), 712–725.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623–641.

Jackson, D., & Marx, G. (2017, December 6). *Data mining program designed to predict child abuse proves unreliable, DCFS says.* Chicago: Tribune. Retrieved from http://www.chicagotribune.com.

Jaeger, P. T., & Bertot, J. C. (2010). Transparency and technological change: Ensuring equal and sustained public access to government information. *Government Information Quarterly, 27*(4), 371–376.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning. Vol. 6*. New York: Springer.
Jarmin, R. S., & O'Hara, A. B. (2016). Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management, 35*(3), 715–721.
Jensen, J. R., Qiu, F., & Ji, M. (1999). Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. *International Journal of Remote Sensing, 20*(14), 2805–2822.
Jun, K. N., & Weare, C. (2010). Institutional motivations in the adoption of innovations: The case of e-government. *Journal of Public Administration Research and Theory, 26*(4), 495–519.
Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: A systematic review. *JAMA, 306*(15), 1688–1698.
Keller, T. E., Cusick, G. R., & Courtney, M. E. (2007). Approaching the transition to adulthood: Distinctive profiles of adolescents aging out of the child welfare system. *Social Service Review, 81*(3), 453–484.
Kemp, S. P., & Bodonyi, J. M. (2002). Beyond termination: Length of stay and predictors of permanency for legally free children. *Child Welfare, 81*(1), 58–86.
King, C. S., Feltey, K. M., & Susel, B. O. N. (1998). The question of participation: Toward authentic public participation in public administration. *Public Administration Review, 58*(4), 317–326.
Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai, 14*(2), 1137–1145.
Kuhn, M. (2008). Caret package. *Journal of Statistical Software, 28*(5), 1–26.
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* New York: Springer389–400.
Lane, J. (2016). Big data for public policy: The quadruple helix. *Journal of Policy Analysis and Management, 35*(3), 708–715.
Lantz, B. (2013). *Machine learning with R.* Birmingham, UK: Packt Publishing Ltd.
Light, P. C. (1998). *The tides of reform: Making government work, 1945–1995*. New Haven, CT: Yale University Press.
Lindblom, C. E. (1959). The science of" muddling through". *Public Administration Review,* 79–88.
Lynn, L. E., Jr, Heinrich, C. J., & Hill, C. J. (2001). *Improving governance: A new logic for empirical research.* Georgetown University Press.
Malatesta, D., & Smith, C. R. (2014). Lessons from resource dependence theory for contemporary public and nonprofit management. *Public Administration Review, 74*(1), 14–25.
Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of "big data". *Policy & Internet, 5*(2), 139–146.
Meijer, A. (2013). Understanding the complex dynamics of transparency. *Public Administration Review, 73*(3), 429–439.
Orsi, R., Lee, C., Winokur, M., & Pearson, A. (2017). Who's been served and how? Permanency outcomes for children and youth involved in child welfare and youth corrections. *Youth Violence and Juvenile Justice, 16*(1), 3–17.
Pelton, L. H. (1991). Beyond permanency planning: Restructuring the public child welfare system. *Social Work, 36*(4), 337–343.
Pirog, M. A. (2014). Data will drive innovation in public policy and management research in the next decade. *Journal of Policy Analysis and Management, 33*(2), 537–543.
Putnam-Hornstein, E., & Needell, B. (2011). Predictors of child protective service contact between birth and age five: An examination of California's 2002 birth cohort. *Children and Youth Services Review, 33*(8), 1337–1344.
Raven, M. C., Billings, J. C., Goldfrank, L. R., Manheimer, E. D., & Gourevitch, M. N. (2009). Medicaid patients at high risk for frequent hospital admission: Real-time identification and remediable risks. *Journal of Urban Health, 86*(2), 230–241.
Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion, 27*(January), 111–125.
Russell, J. (2015). Predictive analytics and child protection: Constraints and opportunities. *Child Abuse & Neglect, 46*, 182–189.
Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2013). Partial least square discriminant analysis for bankruptcy prediction. *Decision Support Systems, 54*(3), 1245–1255.
Sheldon, J. (1997). 50,000 children are waiting: Permanency planning and termination of parental rights under the adoption assistance and child welfare act of 1980. *Boston College Third World Law Journal, 17*, 73–100.
Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.
Simon, H. A. (1957). *Administrative behavior: A study of decision-making processes in administrative organization.* New York, NY: Free Press.
Smith, B. D., & Donovan, S. E. (2003). Child welfare practice in organizational and institutional context. *Social Service Review, 77*(4), 541–563.
Tilbury, C. (2004). The influence of performance measurement on child welfare policy and practice. *British Journal of Social Work, 34*(2), 225–241.
U.S. Department of Health and Human Services (2017). The AFCARS report. Retrieved from https://www.acf.hhs.gov/sites/default/files/cb/afcarsreport24.pdf.
Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine, 45*(3), 354–359.
Wager, S., & Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association,* 1–48. Advance online publication https://doi.org/10.1080/01621459.2017.1319839.
Walker, R. M., Damanpour, F., & Devece, C. A. (2010). Management innovation and organizational performance: The mediating effect of performance management. *Journal of Public Administration Research and Theory, 21*(2), 367–386.