KJPS
Korean Journal of Policy Studies

Na, C., Oh, G., Song, J., & Park, H. (2021). Do Machine Learning Methods Outperform Traditional Statistical Models in Crime Prediction? A Comparison Between Logistic Regression and Neural Networks. *Korean Journal of Policy Studies*, *36*(1), 1–13.

Articles

Do Machine Learning Methods Outperform Traditional Statistical Models in Crime Prediction? A Comparison Between Logistic Regression and Neural Networks

Chongmin Na [a], Gyeongseok Oh [b], Juyoung Song [c], Hyoungah Park [d]

Keywords: machine learning, prediction, neural networks, logistic regression

Vol. 36, Issue 1, 2021

Although machine learning (ML) methods have recently gained popularity in both academia and industry as alternative risk assessment tools for efficient decision-making, inconsistent patterns are observed in the existing literature regarding their competitiveness and utility in predicting various outcomes. Drawing on a sample of the general youth population in the U.S., we compared the predictive accuracy of logistic regression (LR) and neural networks (NNs), which are the most widely applied approaches in conventional statistics and contemporary ML methods, respectively, by adopting many theoretically relevant predictors of the future arrest outcome. Even after fully implementing rigorous ML protocols for model tuning and up-sampling and down-sampling procedures recommended in recent literature to optimize learning algorithms, NNs did not yield substantially improved performance over LR if we still rely on a conventional dataset with relatively small sample sizes and a limited number of predictors. Nonetheless, we encourage more rigorous, comprehensive, and diverse evaluation research for a complete understanding of the ML potential in predictive capacity and the contingencies in which modern ML methods can perform better than conventional parametric statistical models.

Introduction

Predicting which individuals pose higher risks of an initial offense or recidivism within either the general youth population or ex-offender groups has long been challenging, although it is of primary interest for many researchers and practitioners (Farrington, 1987; Gottfredson & Moriarty, 2006). Determining who deserves closer monitoring and assistance is critical after release from institutions, during earlier stages of sanctioning (e.g., pretrial detention, incarceration/probation, early release, and security levels for prison inmates), or before individuals are involved in the criminal justice system (e.g., early prevention/intervention for high-risk youths) to best use limited resources when designing and implementing public safety policies and programs. In combination with the efforts toward theoretical refinements to better develop a comprehensive and evidence-based set of risk assessment instruments, methodological advancements have played significant roles in classifying individuals with higher risks of offending by minimizing prediction errors.

Inspired by the actuarial risk assessment tools in econometrics and public health, researchers have heavily relied on parametric statistical models (e.g., generalized linear models) to predict many criminological outcomes, such as drug use, arrest, recidivism, pretrial detention, incarceration, parole release, and inmate misconduct (Berk & Bleich, 2013; Casey et al., 2011; Farrington & Tarling, 1985; Gottfredson & Gottfredson, 1986; Lowenkamp et al., 2001; Ngo et al., 2015; Pew Center on the States, Public Safety Performance Project, 2011; Sears & Anthony, 2004; Skeem & Monahan, 2011). Despite the consensus that more objective and scientific applications of these statistical methods work better than clinical or professional judgments made exclusively by subjective intuitions with prior experiences or opinions of the decision-makers (Berk, 2012; Gottfredson & Moriarty, 2006; Hastie & Dawes, 2001), their level of predictive accuracy is unimpressive (Lowenkamp et al., 2001; Van Voorhis & Brown, 1997). For example, Farrington & Tarling (1985) and Farrington (1987) found that the level of predictive accuracy from widely implemented statistical methods is not high, with a proportion of false positives and false negatives greater than 0.5. Among many possible reasons, strict and sometimes unrealistic assumptions of conventional parametric models (e.g., the linearity of relationships between the predictor and outcome variables

a Corresponding author, Graduate School of Public Administration, Seoul National University, South Korea, E-mail: chongmin20@snu.ac.kr

b Police Science Institute, Korean National Police University, South Korea, E-mail: safecorea@police.go.kr

c Department of Administration of Justice, Penn State University, Schuylkill, PA, USA, E-mail: jxs6190@psu.edu

d Criminal Justice Department, Saint Perter's University, E-mail: hpark1@saintpeters.edu

and only two-way linear/additive interactions among specific predictors of interest; see also Gottfredson & Moriarty, 2006) might limit the learning process of capturing critical but previously unexpected/unrecognized patterns in the data.

Largely dissatisfied with the disappointingly low level of predictive accuracy of traditional risk assessment tools, alternative strategies have been proposed, such as the random forest (Breiman, 2001), support vector machine (SVM; Vapnik, 2010), and neural network (NN; Kartalopoulos, 1995) methods. These machine learning (ML) methods, widely applied in both academia and industry, are more effective classification algorithms, at least in principle, than traditional parametric regression models. This is because they can "automatically handle non-linearity, handle noisy data, handle a large number of candidate predictors, automatically search and estimate complex interactions, which quickly becomes both unfeasible and unstable by using classical statistics" (Tollenaar & van der Heijden, 2013, p. 566).

Nonetheless, existing comparative studies have concluded that such modern ML methods do not necessarily outperform traditional statistical models. Many comparative studies have also noted that it is unlikely that ML algorithms can make a noticeable difference unless other conditions for successful prediction are satisfied, such as the availability of good predictors, meaningful variation in the predictors, and balance between classes in the outcome variable (Kuhn & Johnson, 2013).

This study aims to assess whether a NN algorithm can outperform the traditional logistic regression (LR) when implemented to maximize its predictive capacity. Many other comparative studies have applied various ML algorithms to the same dataset to select the one that performs best with little explanation about how they work and how their optimization processes are implemented. In contrast, we focus on only the two most widely used prediction methods that are relatively comparable and compare them in a way that is accessible for both academic researchers and practitioners.

Berk & Bleich (2013) claimed that many findings unfavorable to ML might result from improper implementations of the methods. To address their concerns, we built the final models after searching for the optimal tuning parameters that can best use the information available within the predictors and applied up-sampling and down-sampling methods that can address class imbalance in the prediction of rare events, such as arrest. Various ML methods often have inherently different learning algorithms and thus require unique optimization procedures to maximize their predictive performance. Therefore, the result might be misleading if we mechanically apply default tuning parameters preprogrammed in most statistical packages and report the summary findings for a fast and easy comparison of their relative performance. Specifically, we demonstrate and compare the logic behind these two models and the specific optimization procedures we employed to provide additional insight on whether and when ML can outperform conventional approaches, which is understudied in the existing literature.

Prior Comparative Studies on the Utility of Neural Networks Over Logistic Regression in Crime Prediction

Unlike other modern ML techniques that have garnered relatively little attention in criminal justice research, the use of NNs for the prediction of criminological outcomes is not new. Thus, several comparative studies have assessed the advantages of NNs over conventional modeling strategies, such as LR (Brodzinski et al., 1994; Caulkins et al., 1996; Kartalopoulos, 1995; Liu et al., 2011; Ngo et al., 2018; Palocsay et al., 2000; Sears & Anthony, 2004; Tollenaar & van der Heijden, 2013). Nonetheless, the overarching pattern observed in these studies with different samples and model applications demonstrates a level of predictive accuracy in crime prediction similar to that of LR when evaluations are performed on the testing sets independent of the training data. The preponderance of evidence might be better characterized as 'mixed' because some studies found that NNs outperform LR (often moderately), whereas other studies found that LR performs equally well or even better than NNs.

Palocsay et al. (2000) compared the performance outcomes of LR with those of NNs, assessing their capabilities of predicting criminal recidivism. Unlike many other studies that suggested NNs did not offer any significant improvements over conventional statistical approaches in predicting crime (e.g., Caulkins et al., 1996), they found consistently higher levels of total accuracy and sensitivity for NNs than for LR even after checking the results from different model specifications. They concluded, "NN models are competitive with, and may offer some advantages over, traditional statistical models" (271).

Similarly, Brodzinski et al. (1994) achieved 99% accuracy in predicting recidivism among 778 probationers using NNs. This study suggested that collecting and incorporating more and better predictors into the models was essential to minimize classification errors and maximize predictive accuracy. A recent study by Ngo et al. (2018) reached a similar conclusion. This study compared specific performance measures across different prediction models, such as LR, random forest, NN, and ensemble methods. The authors claimed that it is unlikely that one specific method consistently outperforms the others on different performance criteria types representing various aspects of forecasting errors. Specifically, they found that NNs were appropriate when the prediction goal was to maximize either the specificity (true negatives) or overall accuracy (true negatives and true positives).

However, even after adding more predictors than those used by Gottfredson & Gottfredson (1979, 1980), Caulkins et al. (1996) did not find any noticeable benefits of NN models predicting recidivism when they are applied traditionally. Nonetheless, they still called for more research with more predictors that can help better classify offenders because it is possible that "the failure of prediction in this case is not due to inadequate statistical methods, but rather to inadequate knowledge and theory about what kind of variables and mechanisms are linked to future criminal behavior, and to limitations in obtaining satisfactory measures of these variables" (236). Sears & Anthony (2004) also

found that both LR and NNs performed similarly. However, they also called for further comparative evaluations with different data because the benefits of NNs might be best materialized if mere linear combinations of the covariates cannot detect complex and nonlinear data patterns.

More recently, out of a similar motivation to address inconsistent findings in existing comparative studies, Liu et al. (2011) compared LR, classification and regression tree (CART), and NN models for their predictive validity in recidivism among 1,225 UK male prisoners based on a standardized risk assessment instrument. Despite using a multivalidation procedure to reduce sampling error in the estimates of predictive accuracy and controlling for the low base rate to minimize prediction error and achieve a more-balanced classification, the performance of NNs measured by the overall accuracy and area under the receiver operating characteristic curve (AUC) did not exhibit a significant improvement over those of the LR and CART models. Similarly, based on the comparison of predictive accuracy across different models in terms of various performance measures, Tollenaar & van der Heijden (2013) found that classical statistical methods, such as LR and linear discriminant analysis, perform equally well or sometimes even better than alternative ML methods, such as NNs and linear SVM in predicting recidivism.

Current Study

Given the inconsistent patterns in the existing literature, more rigorous, comprehensive, and diverse evaluation research is necessary to assess potential contributions NNs can make in crime prediction and better understand the conditions in which NNs perform better than LR. In this study, we pursue this line of research by assessing the predictive capabilities of NNs with a relatively large sample of the general US youth population. Drawing on this relatively understudied sample, the current study explores whether NNs can yield substantially improved predictive accuracy over LR based on the same set of theoretically relevant predictors by extracting additional information (e.g., nonlinearities, complex interactions, and discontinuities) that cannot be detected using a simple additive LR model. Considering that these less favorable findings for ML methods might result from improper optimizations during the training phase of model building (Berk & Bleich, 2013), we tuned the models to maximize predictive accuracy after seeking optimal tuning parameters via systematic cross-validation procedures. In doing so, we attempted to make the ML procedures as transparent and interpretable as possible for both researchers and practitioners to address the “black box” issues inherent in the application and interpretation of ML methods (Zeng et al., 2017).

Considering that too many model specifications must be considered when all possible interaction terms among predictors are added to the baseline model, we assess whether LR in its simplest additive form without any higher-order or interaction terms can still perform similarly to or even outperform NNs optimized to maximize the performance measure of primary interest. If the optimally tuned NN models perform better than the baseline LR model, we determine whether such patterns result from misspecifying the LR model.

Analytic Strategies

In summary, both LR and NN algorithms are designed to continue the iterative processes of minimizing their unique cost functions until optimal model parameters are identified, linking most properly input features (predictors) to the outcome variable. In this vein, NNs are very similar to LR because both methods search for a set of parameters θ in multidimensional settings via unique learning algorithms for optimal classification. A vector of θs (or multiple vectors of θs in NNs) identified from the training sets is evaluated on independent testing sets to assess how well these predictors forecast future outcomes presumed to be unrealized at the time of data collection. We present a summary of both methods highlighting their similarities and differences because our goals are to assess which method performs better and explicate how each procedure is implemented to maximize predictive accuracy.

Logistic Regression

Logistic regression is one of the most widely used statistical methods in many disciplines, including criminal justice. It demonstrates powerful predictive effectiveness, although it is relatively simple, fast, and easy to estimate the parameters in the hypothesis with LR compared with more-complex ML methods. Moreover, LR is a school of classification algorithms used to assign observations to a discrete set of classes. In contrast, the actual outcomes estimated by LR represent the predicted probabilities of success/failure of the discrete outcome given a set of input profiles. That is, LR is a probability-based classification algorithm in which the sigmoid function transforms the outcomes of the logit model (log of the odds) into the predicted probabilities with values between 0 and 1. In addition, LR belongs to a class of generalized linear models because the log of the odds of an outcome, which can take any value from $-\infty$ to $+\infty$, is estimated using a linear combination of predictor values (weighted by parameter estimates) unless more-complex nonlinear functional forms are explicitly modeled by adding corresponding (e.g., quadratic or cubic) terms:

$$h_\theta(x) = g(z) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}}$$

The outcome is classified as a positive event when $h_\theta(x)$ approaches 1 and a negative event when $h_\theta(x)$ approaches 0. The iterative process of maximum likelihood estimation via the gradient descent algorithm is repeated until it finds a set of parameter estimates that maximize the likelihood of observing the analyzed data. A baseline LR model can be adjusted to model a more-complex hypothesis if additional terms of nonlinearities or interactions are specified. We adopted the simplest functional form of LR for the interest of simplicity, considering that an almost infinite number of alternative functional forms exist, especially when many predictors and interaction terms must be considered.

Neural Networks

Neural networks, along with SVM and random forest methods, are widely used ML methods in various fields as

alternatives to conventional LR because they are better suited for classification problems in which complex and nonlinear relationships exist among input and output variables (Bishop, 2009). Originally designed to simulate the functioning process of biological neurons in the human brain, NNs have advantages over LR because they can relatively easily approximate any nonlinear functional form of relationship from data, unlike conventional parametric regression approaches in which functional forms must be specified a priori or after post-hoc modifications. The inductive nature of the learning process in NNs allows for detecting unknown and unforeseen patterns within the data.

Despite the complex optimization procedures developed systematically to minimize the cost function in estimating parameters, the underlying theory behind the NN algorithm is relatively simple and straightforward. Like LR models, NNs estimate parameters in the hypothesis in training sets to be used for the classification of unrealized data – approximated through independent testing sets – based on observed data features. However, unlike LR models, the outcome variable is modeled as a function of an intermediary set of unobserved variables (*hidden units*), which are modeled as a function of the original input features. These hidden units are comparable to the biological neural system processing multiple input features in the human brain.

The application of NNs can be very flexible because the number of hidden units and hidden layers can vary depending upon the research context and data structure to maximize predictive accuracy. Figure 1 reveals that each hidden unit ($a_i^{(l)}$) is modeled as a linear combination of the predictor variables (x_p) and is transformed using a nonlinear sigmoid function $g(\cdot)$, such as the logistic function (Kuhn & Johnson, 2013, p. 141). The binary outcome is nonlinearly linked to the hidden units through the logistic function.

In Figure 1, $a_i^{(l)}$ denotes the hidden units in layer l, and $\Theta^{(l)}$ is the matrix of parameters controlling the functional form mapping from layer l to layer $l+1$. Like conventional regression models, the θ coefficients in the following equations can be conceptualized as the effects of the predictors on hidden units:

$$a_1^{(2)} = g(\theta_{10}^{(1)}x_0 + \theta_{11}^{(1)}x_1 + \theta_{12}^{(1)}x_2 + \theta_{13}^{(1)}x_3 \ldots + \theta_{1p}^{(1)}x_p)$$
$$a_2^{(2)} = g(\theta_{20}^{(1)}x_0 + \theta_{21}^{(1)}x_1 + \theta_{22}^{(1)}x_2 + \theta_{23}^{(1)}x_3 \ldots + \theta_{2p}^{(1)}x_p)$$
$$a_3^{(2)} = g(\theta_{30}^{(1)}x_0 + \theta_{31}^{(1)}x_1 + \theta_{32}^{(1)}x_2 + \theta_{33}^{(1)}x_3 \ldots + \theta_{3p}^{(1)}x_p)$$
$$\ldots$$
$$a_i^{(2)} = g(\theta_{i0}^{(1)}x_0 + \theta_{i1}^{(1)}x_1 + \theta_{i2}^{(1)}x_2 + \theta_{i3}^{(1)}x_3 \ldots + \theta_{ip}^{(1)}x_p).$$

Although only one layer is incorporated in Figure 1 for the interest of parsimony, NNs can have multiple layers with varying numbers of hidden units. The outcome is modeled as a linear combination of hidden units, which is also transformed using a nonlinear sigmoid function $g(\cdot)$. For example, in a hypothetical situation in which only three hidden units are with one layer, the functional form of the NN model is expressed as follows:

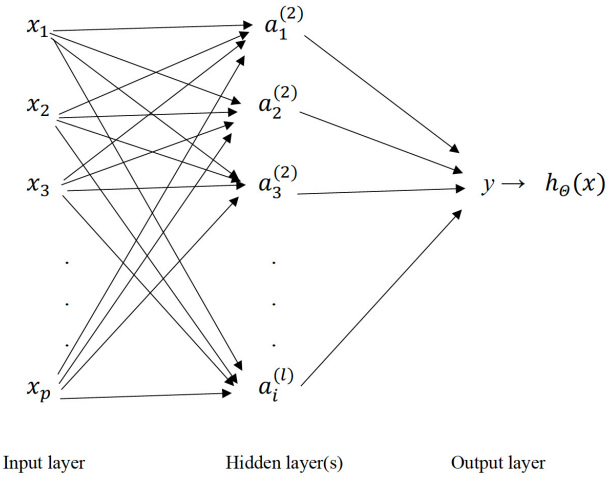

Figure 1. Conceptual Model of NNs with a Single Hidden Layer

$$h_\Theta(x) = g(\theta_{10}^{(2)}a_0^{(2)} + \theta_{11}^{(2)}a_1^{(2)} + \theta_{22}^{(2)}a_2^{(2)} + \theta_{33}^{(2)}a_3^{(2)}).$$

The model becomes extraordinarily complex as the number of hidden units and layers becomes greater. Thus, the process of a cross-validation search for the best number of hidden units and layers is often recommended to optimize the bias-variance trade-off in the prediction.

Like LR, NNs are designed to infer the optimal parameters that minimize the cost function. Specialized learning algorithms are repeated starting with random parameter estimate values to search for optimal parameters that best relate predictors, hidden units, and the outcome variable (Rumelhart et al., 1986).[1] These are the called "feed-forward" (information sources are processed from the input variables to hidden units and then to the output variable) and "back-propagation" (parameters are modified iteratively based on the error estimates transmitted back through the process) algorithms.

Because the number of parameters increases dramatically as the number of hidden units and layers increases, NNs tend to overfit the relationship between the predictors and outcome. A regularization method, such as *weight decay*, is used to penalize large parameter estimates, as other ML methods add a penalty for large parameter values to improve the generalizability of the prediction in the future/unrealized data by reducing the variance involved in complex models (Kuhn & Johnson, 2013, p. 143-145). Along with the optimal number of hidden units and layers, the optimal value of weight decay should also be searched for via cross-validation, making it another tuning parameter in the NN procedure.

In summary, NNs have the flexibility to fit any nonlinear relationship among variables, which additionally contributes to the model's predictive capacity in the classifica-

1 Because NNs often involve numerous parameters, the model tends to find only locally (not globally) optimal parameter values. Multiple model estimation processes can be repeated with different starting values, averaging these results to produce a more stable prediction to address the instability in the model-building process (Kuhn & Johnson, 2013, p. 144).

tion. In addition, NNs allow for identifying unknown patterns among all or subsets of predictors to build a prediction model customized for the specific analyzed data.

Brief Overview of the Machine Learning Procedure

Training, Testing, and Regularization

Considering that crime prediction should be performed with data in which the outcomes are presumed to be yet unrealized, it is essential to work with two independent datasets representative of the target population for model building (training set) and model evaluation (testing set). The predictive errors result primarily from two sources: 'approximation' during the training phase and 'generalization' during the testing phase. Relying on the same dataset tends to yield overly optimistic performance for predictive models because errors resulting from too much variance in predicting future examples are masked by focusing exclusively on minimizing the estimation bias (Berk, 2012).

A systematic process of regularization is required via cross-validation to avoid such an overfitting problem of building models with too much complexity to optimize the bias-variance trade-off and maximize the overall predictive accuracy of the models. Models should be tuned to reduce a large amount of variance in the testing phase, even sacrificing accuracy in the training phase. In practice, some exploratory procedures are necessary to search for the best regularization parameters via cross-validation.

In this study, we selected a random subsample of training data (70%) to build predictive models and used the remaining sample (30%) to evaluate each model. During the model-building phase, 10-fold cross-validation was conducted to optimize the tuning parameters.

Performance Measures

In the extant ML literature, various measures have been used to assess the overall and specific aspects of model performance, such as the overall accuracy, sensitivity, specificity, precision, Kappa, and the area under the receiver operating characteristic curve (AUC)[2] (Berk, 2012). The overall accuracy is the most widely used scale representing the proportion of true positives and true negatives in the total classification outcomes. Considering that our primary goal is to maximize the predictive accuracy of those arrested in the future (true positives), sensitivity is also an important performance measure in this study, which captures the proportion of those correctly classified as positives among those with a true positive outcome. However, when comparing different models, the sensitivity is not always fixed but can be modified by adjusting the classification threshold. Thus, a trade-off exists between sensitivity and specificity, which captures the proportion of those correctly classified as negatives among those with a true negative outcome. Thus, the AUC is often preferred when such a threshold for classification is not fixed but can vary depending on the interests of decision-makers and the available resources for policy purposes. The Kappa statistic measures the concordance of the model prediction and observed classes, which is calculated by $(O - E)/(1 - E)$, where O represents the observed accuracy, and E denotes the expected accuracy based on the marginal totals of the confusion matrix. We focus primarily on the sensitivity measure, even sacrificing other measures, such as specificity, to compare the performance of LR and NN models for the above-noted reason. Accordingly, the NN models are tuned to maximize the sensitivity of classifying true positives among other performance measures.

Data

For our proposed comparative study, we used the National Longitudinal Study of Adolescent Health, a nationally representative sample of US middle and high school students enrolled during the mid-1990s. The age of the participants in the sample at the baseline interview ranged from 11 to 21 years old. The study participants were re-interviewed up to Wave 4, which was implemented in 2008. We analyzed the public-use data with a randomly selected one-half of the core sample and one-half of the oversample of African American adolescents whose parents earned a college degree (n = 6,504). Of the total sample of 6,504, only 5,067 cases had valid data for the arrest outcome and were analyzed in subsequent evaluations. These cases were randomly assigned to training (3,547; 70%) and testing (1,520; 30%) sets for model building and model evaluation, respectively. By creating an independent testing set with the current prospective longitudinal panel data, we approximated the situations in which the outcome (future arrest) was not realized at the time of data collection, although we already know the observed outcome in the testing set.

Measures

Outcome variable: Arrest

The participants were asked if they had ever been arrested (0 = "no" and 1 = "yes") to measure the arrest history of the sample during the last follow-up interview conducted in Wave 4 when the interviewees were between 24 and 34 years of age.

Predictors

Demographic Characteristics: Key demographic characteristics, such as being male, Black, Hispanic, Asian, or another race, were measured in Wave 1 and were included in the models as binary variables (0 = "no" and 1 = "yes") using female (for gender) and White (for race/ethnicity) as reference categories, respectively.

Educational Risk Factors: As essential risk factors for fu-

2 The receiver operating characteristic (ROC) assesses the model's capacity to discriminate between two classes by comparing the ranks of each pair of cases in both classes with respect to the probability value (Mossman, 1994). The ROC is often represented by a graph that plots the number of false positive results against the number of false negative results.

ture arrest among the general youth population, the status of high school dropouts and the level of educational difficulty were measured in Wave 1 and used in the prediction. The high school dropout status was a single binary variable (0 = "no" and 1 = "yes"), and educational difficulty was a composite scale of three individual scores measuring respondents' experiences of repeated grades, school suspension, and expulsion from school. The three dichotomously measured items (0 = "no" and 1 = "yes") were summed to the educational difficulty scale (range: 0 to 3).

Emotional Risk Factors: Emotional risk factors, such as hopelessness and depression, were measured in Wave 1 and were included as predictors for future arrest. A hopeless scale was constructed from three indicators: "felt hopeful about the future" (reverse coded), "thought your life had been a failure," and "felt life was not worth living" (within the past week). The depression scale comprised four indicators: "bothered by things," "felt depressed," "had the blues," and "felt sad" during the past week. Response categories for each indicator ranged from 0 (never/rarely) to 3 (most/all of the time), which were summed to calculate the corresponding scale scores. As another emotional risk factor, suicidal attempts were measured in Wave 1 by asking respondents how many times they had attempted to commit suicide in the past 12 months. The response categories included 0 (0 times), 1 (1 time), 2 (2-3 times), 3 (3-4 times), and 4 (6 or more times), which were recoded as a binary variable to address the issue of extreme skewness in the distribution (0 = 0 times and 1 = at least one time).

Parental Risk Factors: In Wave 3, the experience of physical abuse by primary caregivers was measured using a single item: "How often have your parents or other adult caregivers slapped, hit, or kicked you?" Response categories were 1 (one time), 2 (two times), 3 (three to five times), 4 (six to ten times), 5 (more than ten times), and 6 (this has never happened, which was recoded to 0). Neglect by parents was also measured in Wave 3 using the following two items: "By the time you started 6th grade, how often had your parents or other adult caregivers left you home alone when an adult should have been with you?" and "How often had your parents or other adult caregivers not taken care of your basic needs, such as keeping you clean or providing food or clothing?" Response categories were the same as the physical abuse item, and the scale was created by aggregating two items with scores ranging from 0 to 10.

Behavioral/Situational Risk Factors: A wide range of behavioral and situational risk factors was measured in Wave 1, such as violent behavior, victimization, minor delinquency, drug and alcohol use, smoking cigarettes, immature sexual intercourse, romantic relationships in the past 12 months, and unstructured socializing. Violent behavior comprises seven items: physical fighting, hurting someone, shooting or stabbing someone, threatening with a weapon, group fighting, using a weapon in a fight, and using a knife or gun in a fight. Each item was measured using the binary response category (0 = "no" and 1 = "yes"), and a composite scale was created by aggregating the individual scores.

A violent victimization experience scale was created by aggregating three individual items (having had a knife pulled on the individual, having been shot, or having been stabbed), also measured using two response categories (0 = "no" and 1 = "yes"). Minor delinquency was measured using 10 questions on the respondents' engagement in the following minor delinquent behaviors: graffiti, damaging property, lying to parents, taking merchandise without paying, running away, driving without permission, stealing something worth more than $50 and less than $50, trespassing, and behaving in a loud, rowdy, or unruly way in a public space. The minor delinquency scale was created by aggregating 10 items answered using binary response categories (0 = "no" and 1 = "yes").

A use-of-drugs scale was created similarly by aggregating four dichotomous response items (pot, cocaine, inhalants, and other illegal drugs). The experiences of drinking alcohol, smoking cigarettes, immature sexual intercourse, and romantic relationships were also binary variables (0 = "no" and 1 = "yes") and were included in the models as distinct predictors. Unstructured socializing was measured by asking respondents how often they spent time with their friends in the past week. Responses included 0 (not at all), 1 (1-2 times), 2 (3-4 times), and 3 (5 or more times).

Individual Protective Factor: The importance of religion was measured in Wave 1 by asking respondents whether religion was important to them. The question was initially answered using four response categories of 1 (very important), 2 (fairly important), 3 (fairly unimportant), and 4 (not important at all), but was recoded dichotomously (0 = "fairly unimportant/not important at all" and 1 = "very important/fairly important").

Neighborhood Protective Factors: First, the scale of neighborhood collective efficacy was created by summing the scores of five individual items measured by binary responses (0 = "no" and 1 = "yes") in Wave 1: "You know most people in the neighborhood." "In the past month, you have stopped on the street to talk with someone who lives in your neighborhood." "People in this neighborhood look out for each other." "Do you use a physical fitness or recreation center in your neighborhood?" "Do you usually feel safe in your neighborhood?" Second, the respondents' evaluation of their neighborhood was measured in Wave 1 by the two following items: "On the whole, how happy are you with living in your neighborhood?" "If, for any reason, you had to move from here to some other neighborhood, how happy or unhappy would you be?" These items were assessed using a 5-point Likert scale (1 = "not at all/very unhappy," 2 = "very little/a little unhappy," 3 = "somewhat/would not make any difference," 4 = "quite a bit/a little happy," and 5 = "very much/very happy"). The items were summed to create the scale after reverse coding the second item.

Results

Table 1 lists the descriptive statistics of the variables that we analyzed in both LR and NN models. The base rate of arrest was .29, which was not low enough to negatively affect the forecasting leverage of the predictors (Berk, 2012, p. 5).[3] The demographic profiles indicate that males comprised slightly less than half of the sample (46%), and the racial composition of the participants was diverse (White: 52%; Black: 24%; Hispanic: 10%; Native American: 4%; Asian: 4%; and others: 6%). The predictors demonstrated significant variations across individuals, and our preliminary

Table 1. Descriptive statistics

Variables	Mean/%	SD	Min	Max
Outcome Variable				
Arrest* (W4)	.29	-	0	1
Predictors				
Demographic information (W1)				
Male*	.46	-	0	1
Race* (white as a reference)				
Black	.24	.42	0	1
Hispanic	.10	.30	0	1
Native American	.04	.18	0	1
Asian	.04	.18	0	1
Others	.06	.24	0	1
Educational risk factors (W1)				
Highschool dropout*	.15	-	0	1
Educational difficulty	.51	.74	0	3
Emotional risk factors (W1)				
Hopelessness	1.49	1.43	0	9
Depression	1.94	2.21	0	12
Suicidal attempt*	.04	-	0	1
Parental risk factors (W3)				
Physical abuse	.59	1.32	0	5
Neglect	.98	1.83	0	10
Behavioral/situational risk factors (W1)				
Violent behavior	.75	1.15	0	7
Violent victimization	.17	.46	0	3
Minor delinquency	1.93	1.97	0	10
Use of drug	.18	.51	0	4
Use of alcohol*	.55	-	0	1
Smoking cigarette*	.20	-	0	1
Immature sexual intercourse*	.38	-	0	1
Romantic relationships*	.56	-	0	1
Unstructured socializing	1.98	1.00	0	3
Individual protective factor (W1)				
Religion*	.78	-	0	1
Neighborhood risk factors (W1)				
Collective efficacy	3.32	1.19	0	5
Neighborhood evaluation	7.42	1.97	0	10
n = 5,067				

* denotes binary variables and their means represent the proportions

checks suggest that no redundant or noninformative predictors might negatively affect the predictive model performance (Kuhn & Johnson, 2013, p. 488).[4] Because NNs are sensitive to the nature of variation in the predictors (e.g., different scales, severely skewed distributions, and extreme outliers: Kuhn & Johnson, 2013), these predictors were centered and standardized prior to modeling to maximize the model's predictive capacity.

3 This is often called a "low base rate problem" in prediction research (Berk, 2012, p. 10) and is often observed in the study of criminological outcomes that rarely occur.

4 These results are available upon request.

Using the 'glm' function in R (R Core Team, 2018), we first ran the LR model with the simplest additive functional form, without any polynomial and interaction terms, to compare its predictive performance with that of NNs. Then, we ran a series of NN models with the model code 'nnet' in R (Venables & Ripley, 2002) to search for optimal tuning parameters via cross-validation and minimize the influence of the class imbalance by applying up-sampling and down-sampling methods. Tables 2 reports the performance of these models when they were built on the training set (n = 3,547) and evaluated on the testing set (n = 1,520) not used in the model-building procedures. To achieve our goal of maximizing the sensitivity of arrest prediction (i.e., increasing the true positives in the prediction of future arrest), we first trained and tuned the NN models with 'sensitivity' as an option during the process of selecting an optimal model (Model 2 in Table 2). However, we also used the ROC as a criterion in the optimization process to determine whether different options significantly differ in the primary findings (Model 3 in Table 2). These options were not available for the LR models and could not be applied (Model 1 in Table 2). These tuning processes of optimizing NN models adjust the model parameters to increase predictive accuracy instead of modifying the threshold for post-hoc classification of positive outcomes based on the predicted probabilities estimated using the same model parameter values. Specifically, we identified the optimal number of hidden units (neurons) and decay parameters via 10-fold cross-validation to maximize model performance.[5] Our goal was to improve the model sensitivity because our 'substantive' accuracy measure of interest was via these additional optimization processes, even sacrificing other performance outcomes, such as the specificity, overall accuracy, and Kappa value.

Table 2 demonstrates that the simple additive LR model performs equally well or slightly better than the NN models with tuning for all performance measures. In particular, the number of true arrestees correctly classified as arrested by the models (165 out of 437: sensitivity = 0.3776) was the same in Models 1 (LR) and 2 (NNs with tuning optimized using sensitivity), although Model 1 performed a bit better at predicting true negatives (not arrested) than Model 2. The results were 1,005 out of 1,083 (specificity = 0.9280) for Model 1 and 1,001 out of 1,083 (specificity = 0.9243) for Model 2. The sensitivity index of Model 3 (NNs with tuning optimized using the ROC) was slightly higher than that of Model 1 (0.3822 and 0.3776, respectively), whereas its specificity index was a bit lower than that of Model 1 (0.9224 and 0.9280, respectively). Overall, more complex and time-consuming NN models, even with optimal tuning parameters, did not outperform the conventional LR model in predicting future arrest outcomes within the current general US youth population.

Considering that sensitivity was our primary performance measure of interest, the model performance in Table 2 is not satisfactory because more than half of the true positives are still incorrectly predicted by Models 1 to 3. Such disappointing sensitivity levels could result from many known and unknown factors affecting predictive accuracy, such as the absence of good predictors, lack of variation in the predictors, and imbalance between classes. Significantly, the ML literature suggests that when a severe imbalance between negative and positive outcomes occurs, predictive models tend to be overwhelmed by the patterns observed in the majority classes (e.g., those who were not arrested in this study), and thus achieve good specificity (Kuhn & Johnson, 2013, p. 421).

Because other issues are inherent limitations that cannot be addressed with the current secondary data, we attempted to overcome the class imbalance issue by approximating a situation in which good class balance was achieved via up-sampling and downsampling the data. In the upsampling procedure, cases from the minority class are resampled with replacement until their number equals the number of majority class cases. In the downsampling procedure, majority class cases are randomly selected to match the number of minority class cases. Accordingly, the total sample sizes for the upsampled and downsampled training data were 5,054 and 2,040, respectively, although the sample sizes for the testing data remained the same for all models. As before, we also relied on the sensitivity measure to determine the optimal tuning parameter values in cross-validation for both NN models (Models 2 and 4) in Table 3.

The performance outcomes in Table 3(b) demonstrate that, if we compare the same ML methods, down-sampling performs better for LR (except for specificity) and up-sampling performs better for NNs (except for sensitivity). However, if we compare LR and NNs within the same sampling procedures, LR still performs noticeably better than NNs. The only exception is the sensitivity outcomes for the down-sampling data, which are 0.7002 and 0.7048 for LR and NNs, respectively. Considering the trade-off between the sensitivity and specificity, if we modify the threshold for classification, a much higher level of sensitivity is achieved (0.7414) when we sacrifice the specificity of the LR model to the same level as in the NN model (0.6999) by slightly lowering the cut-off value for the positive class from the conventional value of 0.5 to the value of 0.47.

In summary, we found that our simple additive LR models perform better overall than more-complex NN models and better classify true positives correctly in particular. Coupled with the better interpretability of the parameter estimates in the LR models, LR is still a powerful and useful ML algorithm in crime prediction with the current sample of the general youth population, not to mention its mathematical efficiency compared with more-complex ML methods. Primarily, we failed to discover any noticeable advantages of NNs over LR even after implementing the additional procedures of model tuning and up-sampling and down-sampling methods recommended in the recent literature (Berk, 2012; Kuhn & Johnson, 2013) to better optimize the model's predictive capacity.

5 We did not vary the number of hidden layers during the cross-validation for parsimony to save time during the model estimation process because our preliminary analyses suggested that adding more layers did not improve the performance of the NN models.

Table 2. Performance of the Models without Up- and Down-sampling

a. Contingency Tables of Classification

Model 1: LR (without tuning)

	Observed (True)		Total
Predicted (Classified)	Not arrested	Arrested	
Not arrested	1005	272	1277
Arrested	78	165	243
Total	1083	437	1520

Model 2: NNs (with tuning optimized with 'sensitivity')

	Observed (True)		Total
Predicted (Classified)	Not arrested	Arrested	
Not arrested	1001	272	1273
Arrested	82	165	247
Total	1083	437	1520

After 10-fold cross-validation, # of hidden units = 5 and decay = 1 were applied.

Model 3: NNs (with tuning optimized with 'ROC')

	Observed (True)		Total
Predicted (Classified)	No-arrest	Yes-arrest	
Not arrested	999	270	1269
Arrested	84	167	251
Total	1083	437	1520

After 10-fold cross-validation, # of hidden units = 10 and decay = 2 were applied.

b. Performance Outcomes

	Model 1	Model 2	Model 3
	LR (without polynomial and interaction terms)	NNs (with tuning optimized with 'sensitivity')	NNs (with tuning optimized with 'ROC')
Accuracy	0.7697	0.7671	0.7671
Kappa	0.3522	0.3468	0.3489
Sensitivity	0.3776	0.3776	0.3822
Specificity	0.9280	0.9243	0.9224

For NNs with tuning optimized by "sensitivity', # of hidden units = 5 and decay = 1 were applied via 10-fold cross-validation.

For NNs with tuning optimized by "ROC", # of hidden units = 10 and decay = 2 were applied via 10-fold cross-validation.

Conclusion

Our primary goal was to fully assess whether NNs can outperform LR in predicting future arrests, especially when fine-tuned through cross-validation to achieve class balance in the outcome variable using up-sampling and down-sampling procedures to maximize model sensitivity. Contrary to our expectations, no discernible differences were observed in the predictive accuracy between LR and NN models. As often reported in earlier or even more recent comparative studies, even simple additive LR models without interaction terms performed as well as more-complex NN models, even after searching for an optimal solution in the bias-variance trade-off. Thus, these models are mathematically less efficient and less intuitive to both researchers and practitioners.

Such patterns were observed consistently across different model specifications. We followed standardized procedures in modern ML methods by creating two independent samples representing the target population to build pre-

Table 3. Performance of the Models with Up- and Down-Sampling

a. Contingency Tables of Classification

Model 1: LR (with Up Sampling)

	Observed (True)		Total
Predicted (Classified)	Not arrested	Arrested	
Not arrested	799	140	939
Arrested	284	297	581
Total	1083	437	1520

Model 2: NNs (with Up Sampling)

	Observed (True)		Total
Predicted (Classified)	Not arrested	Arrested	
Not arrested	787	145	932
Arrested	296	292	588
Total	1083	437	1520

After 10-fold cross-validation, # of hidden units = 6 and decay = 0 were applied.

Model 3: LR (with Down Sampling)

	Observed (True)		Total
Predicted (Classified)	Not arrested	Arrested	
Not arrested	794	131	925
Arrested	289	306	595
Total	1083	437	1520

Model 4: NNs (with Down Sampling)

	Observed (True)		Total
Predicted (Classified)	No-arrest	Yes-arrest	
Not arrested	758	129	887
Arrested	325	308	633
Total	1083	437	1520

After 10-fold cross-validation, # of hidden units = 10 and decay = 2 were applied.

b. Performance Outcomes

	Up-sampling		Down-sampling	
	Model 1	Model 2	Model 3	Model 4
	LR	NNs	LR	NNs
Accuracy	0.7211	0.7099	0.7237	0.7013
Kappa	0.3800	0.3580	0.3912	0.3570
Sensitivity	0.6796	0.6682	0.7002	0.7048
Specificity	0.7378	0.7267	0.7331	0.6999

For NNs with Up-sampling, # of hidden units = 6 and decay = 0 were applied via 10-fold cross-validation.

For NNs with Down-sampling, # of hidden units = 10 and decay = 2 were applied via 10-fold cross-validation.

dictive models and evaluate how they perform with new datasets (Berk, 2012, p. 10). In particular, we used a representative sample of the general US youth population. We created two independent samples through a random selection process: a training set with inferred model parameters and a testing set to assess the learned ML algorithms and inferred parameters.

We emphasize that it is critical to use an independent but representative population sample in the model evaluation to approximate future examples with unrealized outcomes. We optimized the NN models by searching for the best functional form and tuning parameters via cross-validation to address the concerns of the proponents of ML methods. They claim that less optimistic conclusions re-

garding the predictive capacity of various ML methods in the extant comparative studies result from insufficient or even improper implementations of the procedures during the training phase of model building (e.g., Berk & Bleich, 2013). Nonetheless, LR worked equally well as or even outperformed NNs for all compared performance measures, including sensitivity, which is of primary interest in future arrest prediction.

Considering that our primary goal was to classify true arrestees correctly, the model performance was still unsatisfactory with a sensitivity value of less than 0.4. Our final models addressed this issue by adopting sampling methods to balance the cases with negative and positive outcomes, which have often been recommended in the recent ML literature to minimize the undesired influence of class imbalance on model performance (Berk, 2012; Kuhn & Johnson, 2013). Overall, the sensitivity increased substantially after applying up-sampling and down-sampling procedures for both LR and NN models. However, LR remained more competitive than NNs for all criteria applied to assess the models' successful prediction of future arrest in the current sample.

Despite the less favorable results for NNs, we still call for more of this line of research, comparing the strengths and weaknesses of each ML algorithm. Our less optimistic results might result from the current data's inherent limitations, not from the predictive methods themselves. Indeed, the prediction of various criminological outcomes is dependent on the data, and the results can vary widely across various applications (Jamain & Hand, 2008). The potential of modern ML in maximizing predictive accuracy might be fully materialized with 'big data' with a much larger sample size and more input features than those adopted in this study (Berk & Bleich, 2013, p. 519). Our findings suggest that predictive models derived from a priori knowledge are not guaranteed to perform well with new datasets. Suppose other characteristics of individuals that are seemingly unrelated to the risk of arrest indeed contribute to the prediction in very implicit and subtle ways that are unknown to researchers and practitioners. In that case, modern ML techniques might be better suited for identifying and employing such hidden patterns in the data.

Most importantly, rapid progress has been made in ML theories and methods, and NNs are central to such recent developments (e.g., deep learning with the convolution NN or recurrent NN). Our traditional approach of applying basic multilayer perceptron NNs might have inherent limitations when competing with conventional statistical models in this vein. Future research should address these limitations and explore how ML methods should be implemented to realize their full potential to maximize predictive accuracy.

Submitted: October 19, 2020 KST, Accepted: November 24, 2020 KST

REFERENCES

Berk, R. A. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Public Policy*, *12*(3), 513–544. https://doi.org/10.1111/1745-9133.12047

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (2001). Decision tree forest. *Machine Learning*, *45*(1), 5–32.

Brodzinski, J. D., Crable, E. A., & Scherer, R. F. (1994). Using artificial intelligence to model juvenile recidivism patterns. *Computers in Human Services*, *10*, 1–18.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011). *Using offender risk and needs assessment information at sentencing: Guidance from a national working group*. National Center for State Courts. http://www.ncsconline.org/

Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting the criminal recidivism: A comparison of neural network models with statistical models. *Journal of Criminal Justice*, *24*, 227–240.

Farrington, D. P. (1987). Predicting individual crime rates. In D. M. Gottfredson & J. Tonry (Eds.), *Crime and justice: an annual review of research* (Vol. 9). University of Chicago Press. https://doi.org/10.1086/449132

Farrington, D. P., & Tarling, R. (1985). *Prediction in criminology*. State University of New York Press.

Gottfredson, S. D., & Gottfredson, D. M. (1979). *Screening for risk: A comparison of methods*. U.S. Government Priming Office.

Gottfredson, S. D., & Gottfredson, D. M. (1980). Screening for risk: A comparison of methods. *Criminal Justice and Behavior*, *7*(3), 315–330. https://doi.org/10.1177/009385488000700306

Gottfredson, S. D., & Gottfredson, D. M. (1986). Accuracy of prediction models. In A. Blumstein, J. Cohen, J. Roth, & C. A. Visher (Eds.), *Criminal careers and "career criminals"* (Vol. 2, pp. 212–290). National Academy of Sciences Press.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency*, *52*, 178–200.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications, Inc.

Jamain, A., & Hand, D. J. (2008). Mining supervised classification performance studies: A meta- analytic investigation. *Journal of Classification*, *25*, 87–112.

Kartalopoulos, S. V. (1995). *Understanding Neural Networks and Fuzzy Logic: Basic Concepts and Applications*. IEEE Press. https://doi.org/10.1109/9780470546826

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer-Verlag.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression trees, and neutral networks model in predicting violent re-offending. *Journal of Quantitative Criminology*, *27*, 547–573.

Lowenkamp, C. T., Holsinger, A. M., & Latessa, E. J. (2001). Risk/need assessment, offender classification, and the role of childhood abuse. *Criminal Justice and Behavior*, *28*(5), 543–563. https://doi.org/10.1177/009385480102800501

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *J. Consultng Clin. Psychol*, *6*, 783–792.

Ngo, F. T., Govindu, R., & Agarwal, A. (2015). Assessing the predictive utility of logistic regression, classification and regression tree, chi-squared automatic interaction detection, and neural network models in predicting inmate misconduct. *American Journal of Criminal Justice*, *40*(1), 47–74. https://doi.org/10.1007/s12103-014-9246-6

Ngo, F. T., Govindu, R., & Agarwal, A. (2018). Traditional Regression Methods versus the Utility of Machine Learning Techniques in Forecasting Inmate Misconduct in the United States: An Exploration of the Prospects of the Techniques. *International Journal of Criminal Justice Sciences*, *13*(2), 420–437.

Palocsay, S. W., Wang, P., & Brookshire, R. G. (2000). Predicting criminal recidivism using neural networks. *Socio-Economic Planning Sciences*, *34*, 271–284.

Pew Center on the States, Public Safety Performance Project. (2011). *Risk/needs assessment 101: science reveals new tools to manage offenders*. The Pew Center on the States. http://www.pewcenteronthestates.org/

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the micro-structure of cognition* (Vol. 1–2). The MIT Press. https://doi.org/10.7551/mitpress/5236.001.0001

Sears, E. S., & Anthony, J. C. (2004). Artificial Neural Networks for Adolescent Marijuana Use and Clinical Features of Marijuana Dependence. *Substance Use & Misuse*, *39*(1), 107–134. https://doi.org/10.1081/ja-120027768

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, *21*(1), 38–42. https://doi.org/10.1177/0963721410397271

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive methods. *Journal of the Royal Statistical Society, Series A*, *176*(part 2), 565–584.

Van Voorhis, P., & Brown, K. (1997). *Risk classification in the 1990s*. National Institute of Corrections.

Vapnik, V. (2010). *The Nature of Statistical Learning Theory*. Springer.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer.

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(3), 689–722. https://doi.org/10.1111/rssa.12227