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

Fawn T. Ngo,[1] **Ramakrishna Govindu,**[2] **& Anurag Agarwal**[3]
University of South Florida Sarasota-Manatee, United States of America

Abstract

In the U.S., prison administrators often rely on risk assessment instruments to place and supervise inmates, as well as manage, plan and allocate resources. Hence, any improvement in the accuracy performance of risk assessment instruments is likely to result in significant benefits for offender classification and rehabilitation, management systems, and public safety. To date, researchers have explored the relative predictive performance between regression and non-regression methods and the overall evidence is inconclusive. In this study, we seek to advance the debate regarding the efficacy of traditional regression methods versus the utility of machine learning techniques in forecasting inmate misconduct by exploring the prospect that each technique may be more suitable for a specific performance measure. We examined the relative performance of a traditional regression method, logistic regression, and two machine learning techniques, random forest and neural networks, in classifying the proportion of inmates who engaged in serious misconduct (sensitivity), the proportion of inmates who did not engage in serious misconduct (specificity), and the proportion of inmates who did and did not engage in serious misconduct (overall accuracy). We found that to maximize sensitivity, the ensemble method should be employed, to maximize overall accuracy, the neural networks technique should be utilized, and to maximize specificity, either the random forest or neural networks approach will suffice.

Keywords: Serious Inmate Misconduct, Machine Learning Techniques, Comparative Statistical Techniques, Importation Model, Deprivation Model.

[1]Associate Professor of Criminology, College of Liberal Arts & Social Sciences, University of South Florida Sarasota-Manatee, Sarasota, Florida, USA. E-mail: fawnngo@sar.usf.edu.
[2]Instructor of Information Systems/Decision Sciences, College of Business, University of South Florida Sarasota-Manatee, Sarasota, Florida, USA. E-mail: rgovindu@sar.usf.edu
[3] Professor of Information Systems/Decision Sciences, College of Business, University of South Florida Sarasota-Manatee, Sarasota, Florida, USA. E-mail: agarwala@sar.usf.edu

Introduction

Since the 1920s, risk assessment instruments for offenders have played a role in the U.S. criminal justice system's decision-making process, including decisions on inmate placement and supervision (Gottfredson & Moriarty, 2006). These risk assessment instruments have evolved from being very subjective in nature, where decisions were largely based on "expert" opinions by criminal justice professionals and clinical psychologists, to being more objective in nature involving data-driven empirical methods grounded in theory and research. The former type of risk assessment is known as the clinical approach or first generation of risk assessments and the latter type is known as the actuarial approach or second generation of risk assessments.

Actuarial risk assessment techniques are typically based on generalized linear models (i.e., linear or logistic regression). Recently, as a result from the criticisms charged against conventional actuarial techniques (Gendreau et al., 2002; Gotfredson & Gotfredson, 1986; Glover et al., 2002; Steadman et al., 2000), more specialized applications of offender assessments using machine learning and data mining techniques have been proposed and evaluated (Berk & Bleich, 2013, Berk, Kreigler & Baek, 2006; Berk et al., 2009; Hamilton et al., 2014; Liu et al., 2011; Ngo, Govindu, & Agarwal, 2014). However, some scholars have questioned as well as refuted the claim that non-regression methods would lead to improved predictive validity (Hamilton et al., 2014; Tollenaar & van der Heijden, 2013).

In this study, we explore the prospect that traditional regression methods and machine learning techniques may be appropriate for *different* predictive performance measures (e.g., overall accuracy, sensitivity, etc.). Specifically, we examine the utility of a traditional regression method, logistic regression (LR), and two machine learning techniques, random forests (RF) and neural networks (NN), in predicting serious inmate misconduct using four specific performance measures - overall accuracy, sensitivity, specificity, and AUC under ROC. We also investigate the utility of combining the results of these three statistical techniques as an ensemble, in enhancing the predictive performance for the above four performance measures.

Our study extends prior research in four ways. First, we employ an outcome variable that is rarely examined in prior comparative research, serious inmate misconduct. To the best of our knowledge, to date, there are only two studies that have evaluated and compared the predictive performance of conventional regression methods with machine learning techniques in forecasting inmate misconduct (Berk et al., 2006; Ngo et al., 2014) and only one involved the outcome of serious inmate infractions (Berk et al., 2006). Second, in addition to employing the traditional regression approach and classification techniques drawn from the data mining and machine learning literature (i.e., RF and NN), our study also proposes and investigates the utility in combining the results from these "distinct" classification techniques as an ensemble. It is noteworthy that our proposed ensemble technique is distinct from the usual practice involving the ensemble of the same technique (e.g., RF is an ensemble of multiple classification and regression trees). Third, whereas prior research has either examined the performance of a single theoretical model using a single classification technique (e.g., studies using LR to examine the efficacy of the importation model in accounting for inmate misconduct) or comparing the performance of multiple theoretical models using a single classification technique (e.g., studies using LR to examine the efficacy of the importation, deprivation, and situational models in accounting for inmate misconduct), our study evaluates the relative performance of multiple theoretical models using multiple classification techniques simultaneously (i.e.,

using LR, RF, and NN to assess the predictive accuracy of the importation and deprivation models in accounting for serious inmate misconduct). Finally, given the equivocal findings regarding the relative predictive performance of traditional regression methods and machine learning and data mining techniques, our study is perhaps the first to propose and explore the relative efficacy of these statistical approaches for different predictive performance measures (i.e., overall accuracy, sensitivity, etc.).

The remainder of this paper is organized as follows. First, we provide a brief overview of the LR, RF, NN and ensemble methods employed in the current study. We also delineate the importation and deprivation perspectives on inmate adaptation to prison from which we draw our predictor variables. Next, we describe our data, analysis methods, variables, and performance measures. Finally, we report our results and discuss their implications.

Classification Techniques and Ensemble Method

Logistic Regression

Logistic regression (LR) is a type of probabilistic statistical technique used to model a binary outcome variable. Similar to the linear regression analysis method, the goal of the LR technique is to find the best fitting model that describes the relationship between the dependent or outcome variable and a set of independent or predictor variables. However, whereas the outcome variable in a linear regression model is continuous, the outcome variable in a LR model is dichotomous and is therefore considered an appropriate technique for binary classification (Grimm & Yarnold, 1995; See also, Liu et al., 2011; Ngo et al., 2014).

Findings from prior comparative studies have provided support for LR as one of the better methods for binary classification (Hosmer & Lemeshow, 1989; Tollenar & van der Heijden, 2013). However, because the LR model is based on generalized linear models, important nonlinearities and interaction effects must be identified by the researcher and included in the model. Accordingly, critics of the LR method have argued that unless the decision boundary for a particular forecast is simple or the researcher possesses the knowledge and required training that enable him or her to identify complex decision boundaries, derive a suitable algebraic form, and have access to data to construct an appropriate prediction model, predictions based on conventional regression-based methods such as LR could have adverse consequences (Berk & Bleich, 2013).

Random Forest

A random forest (RF) is essentially an ensemble of classification and regression trees (CART; Breiman, 2001). The CART method produces a regression tree when all of the independent variables are continuous, a classification tree when all of the independent variables are discrete, and a classification and regression tree when the independent variables consist of both discrete and random variables. Further, this approach uses predictors to split data into homogenous groups or "branches" through a series of conditional answers (Breiman et al., 1984; Ripley, 1996). More specifically, through a set of logical *if-then* conditions, the CART method divides a sample into "branches" and within each of these branches, the best predictor is determined until no more variance can be explained with the remaining variables or some other criterion (e.g., a minimum group

size) is satisfied. The resulting category groups represent subgroups of the original sample that differ in terms of the probability of the outcome variable (Liu et al., 2011; Ngo et al., 2014).

Whereas in standard classification and regression tree analysis each branch is split using the best predictor among all predictors, in RF, each branch is split using the best predictor among a subset of predictors randomly chosen at that branch. Additionally, each tree is independently constructed using a bootstrap sample of the original data set and each tree "votes" for one category group or class. In the end, the forest selects the group or class with the majority of votes (Liaw & Wiener, 2002). There is evidence that relative to CART, RF appears to possess superior predictive performance (Berk et al., 2006). RF is also known to be robust to over-fitting or shrinkage (an occurrence when a statistical model demonstrates poor predictive performance or when the predictive accuracy of a model decreases from the training sample to the test sample), suitable for identifying interactions and can be tuned to address the relative costs of false positives (a false positive is defined as a positive result on a diagnostic test for a condition in an individual who actually does *not* have that condition)and false negatives (a false negative is defined as a negative result on a diagnostic test for a condition in an individual who actually *does* have that condition; Berk et al., 2009; Liaw & Wiener, 2002; See also, Hamilton et al., 2014; Neuilly et al., 2011).

Neural Networks

Neural networks (NN), also known as artificial neural networks, are mathematical models inspired by the biological model of the brain, which is essentially a network of neurons. Just like a brain can learn to recognize patterns in the real world, an artificial NNs can learn patterns in data. A NN mimics the learning process of a brain to learn patterns. A neuron of a brain is modeled as a processing element (PE) in an artificial neural network. Many PEs are connected to each other in a certain fashion to create a network of neurons or a neural network. The PEs are connected through connections characterized by connection weights. Using a learning algorithm and some learning parameters, *learning* is accomplished through the modification of the connection weights between PEs (Kartalopoulos, 1995; See also, Liu et al., 2011; Ngo et al., 2014).

Compared to conventional statistical methods such as linear regression or logistic regression, NN models are considered more suitable for data suffering from missing values or involving large measurement errors. NNs are also suitable for identifying complex patterns and relationships (linear and non-linear) between multiple inputs that are not recognizable by the human brain. NNs can also handle noisy data and data involving a large number of predictor variables (Grann & Långström, 2007; Tollenaar & van der Heijden, 2013;).

Ensemble Method

In this study, we also examine the combined predictive performance of the above three classification techniques – LR, NN, and RF - using the maximum predicted probability values generated by each of the techniques (i.e., the ensemble maximum model or EM). For example, suppose for a given prediction, the predicted probability value generated by LR is 0.69, by RF is 0.72, and by NN is 0.70; we then use 0.72 (the maximum of the three values) as our predicted probability for the EM method for that particular case. The

same procedure is then repeated for all of the cases to generate the results for the EM method.

For the EM results, we elected to focus on the ensemble *maximum* value because we want to increase the accuracy of predicting misconduct correctly rather than the accuracy of predicting compliant behavior correctly. Since we are denoting serious misconduct as 1 and no misconduct as 0, taking the highest of the three values will improve the accuracy of correctly predicting misconduct. Our decision of using the maximum of the three values is also premised on the rationale that the cost of a false negative (i.e., when an inmate is classified as not engaging in serious misconduct but he actually does) is considered higher than the cost of a false positive (i.e., when an inmate is classified as being engaged in misconduct but he actually does not). According to extant evidence, the cost of one false negative incidence of serious inmate infraction is equal to the combined costs of ten false positive incidences (Berk et al., 2006).

Deprivation and Importation Models of Inmate Behavior

The predictor variables for our study are derived from the deprivation and importation perspectives on inmate adaptation to prison. The deprivation model, proposed by Sykes and Messinger (1960), posits that the "pains" associated with imprisonment or the deprivations suffered by prisoners are the main determinants of an offender's conduct while incarcerated. In particular, proponents of the deprivation model argue that the existence of an inmate subculture that is in conflict with the prison administration and staff is a byproduct of the deprivations of liberty, goods and services, sexual relationships, autonomy, and security experienced by the prisoners. The existence of the conflicting subculture also leads prisoners to be aggressive, resist authority, violate prison rules or attack other inmates (Cao et al., 1997; Goodstein & Wright, 1989; Harer & Steffensmeier, 1996; Wright, 1991). On the other hand, the importation model, developed by Irwin and Cressey (1962), maintains that inmates' behavior in confinement is determined by their distinctive traits and social background prior to incarceration. That is, inmates import their prior behavioral characteristics from outside the prison into the prison culture and thus, if an inmate had proclivities towards violence prior to incarceration, he is also very likely to behave violently while incarcerated (Lahm, 2008; Mears et al., 2013).

Prior research assessing the efficacy of the importation and deprivation models have provided support for both perspectives (Ayton & Lowenstein, 2007; Cao et al., 1997; Dhami et al., 2007; Harer & Steffensmeier, 1996; Jiang and Fisher-Giorlando, 2002; Steiner & Wooldredge, 2008; Sorensen, Wrinkle & Gutierrez, 1998; Paterline & Petersen, 1999; Woolredge, 1991). In a recent project that involves data from 98 different studies published in top criminology and sociology journals, Steiner and colleagues (2014) performed a systematic review on the causes and correlates of prison inmate misconduct. The outcome variable included in Steiner and colleagues' study encompassed all types of misconduct (i.e., staff assault, inmate assault, drug/alcohol, property, etc.) and the predictor variables were derived from the importation, deprivation, and situational/administrative control perspectives. However, for their meta analysis, the researchers categorized their independent variables into three groups of measures:1) background characteristics, 2) institutional routines/experiences, and 3) prison characteristics. Steiner and colleagues found all three groups of predictor measures were significantly related to inmate misconduct albeit there was evidence of between model

variability in the effects of nearly every single predictor. Further, to date, none of the prior research comparing the efficacy of the deprivation and importation models on inmate behavior has employed multiple statistical techniques or techniques drawn from the machine learning and data mining literature.

Data and Methods

Data

Data for the current study came from the 2004 Survey of Inmates in State and Federal Correctional Facilities (SISFCF) conducted for the United States Bureau of Justice Statistics (BJS) by the Bureau of the Census (ICPSR #4572). Data collection for SISFCF involved a two-stage stratified sample design with correctional facilities chosen in the first stage and inmates within facilities chosen in the second stage. The SISFCF data provides nationally representative data on U.S. offenders held in state and federal prisons with personal interviews with the inmates occurring between October 2003 and May 2004. Inmates participating in the SISFCF provided information about their current offense and sentence, criminal history, family background and characteristics, prior drug and alcohol use, medical and mental health conditions, participation in treatment programs, gun possession and use, and prison activities, programs, and services.

A total of 14,499 inmates participated in the 2004 SISFCF and after accounting for missing data and non responses, the sample size was reduced to 10,328. From the reduced sample, approximately 1,283 inmates reported that they had been written up or found guilty of violating serious prison infractions and approximately 3,600 inmates indicated that they had not been written up or found guilty of violating any prison infractions. The remaining cases consist of inmates who were involved in only minor infractions and since these cases were not the focus of this study, they were excluded from the study.

Cross-Validation Procedure

Cross-validation is an empirical procedure to obtain an unbiased estimate of predictive accuracy (Gottfredson & Moriarty, 2006). The cross-validation procedure requires a "training" dataset to build a classification model, which can then be used to classify cases in the "testing" dataset. In this study, we employ the k-fold cross-validation procedure, a method where the entire sample is randomly partitioned into k equally-sized subsamples and one of the k subsamples is retained as the testing data to test the model, and the remaining k – 1 subsamples are used as training data. The cross-validation process is then repeated with each of the k subsamples used exactly once as the testing data (Breiman et al., 1984).It is noteworthy that the k-fold cross-validation approach, where $k > 1$, yields more reliable classification accuracy than a single-sample validation as the latter approach may result in over-fitting (Grann & Långström, 2007).

In addition to the k-fold cross-validation procedure, we were also interested in obtaining a baseline ratio of inmates who committed serious misconduct versus those who did not at 50%:50%. Accordingly, we randomly selected 1,250 cases from the pool of 1,283 inmates who violated serious prison infractions and 1,250 cases from the pool of 3,600 inmates who did not violate any prison rules and partitioned randomly the original dataset(N=2,500) into five sub-datasets of 500 cases each labeled A, B, C, D, and E. These five sub-datasets (with each set consists of 250 cases of inmates who committed serious

misconduct and 250 cases of inmates who did not engage in any misconduct) were used for model development and testing in this study.

Variables

Our outcome variable is a binary variable and denotes whether the inmate was cited for or found guilty of a serious prison violation (misconduct) or not. For our study, serious or major prison infractions include the following categories: 1) possession of a weapon; 2) physical assault on a correctional officer or other staff member; 3) physical assault on another inmate; 4) escape or attempted escape; and 5) other major violations including food strikes, setting fire, rioting, etc.

Twenty-six importation measures and eleven deprivation measures were also included in the study as predictor variables. Our importation variables include gender, age, race, marital status, education level, employment, homelessness, service in the United States Armed Forces, substance use, mental health condition, prior arrests, age at first arrest, and current type of offenses. Our deprivation variables include contact with family and friends (phone and visits), time spent in physical exercises, and participation in various prison programming (e.g., inmate assistance groups, parenting classes, life skills classes, etc.).The descriptive statistics for the outcome and predictor variables are shown in Table 1.

Performance Measures

Following prior comparative studies, we rely on multiple performance measures to evaluate the predictive accuracy of the classification techniques and ensemble method included in our study. Specifically, we report the sensitivity, specificity, overall accuracy, and the area under the curve (AUC) of the Receiver Operating Characteristic (ROC; Egan, 1975; Swets, 1988) values. The sensitivity measure represents the proportion of positives that are correctly classified (sensitivity = [the number of positives correctly identified by the model]/[the number of all positives])and the specificity measure denotes the proportion of negatives accurately classified by the classifier (specificity = [the number of negatives correctly identified by the model] / [the number of all negatives]). In our study, a misconduct is treated as "positive" and no misconduct is considered "negative." The overall accuracy measure is the combination of true positives and true negatives as a proportion of total cases (overall accuracy = [the number of positives and the number of negatives correctly identified by the model] / [the number of all positives and all negatives]).The AUC under the ROC for a binary classification problem essentially plots the true positive rate (TPR or [the number of positives correctly identified by the model]/[the number of all positives]) as a function of the false positive rate (FPR or(1 – [the number of negatives correctly identified by the model] / [the number of all negatives]) for all observed predictor values. That is, the ROC curve captures the tradeoff in the false positive rate that occurs as the true positive rate increases with lower cutoff values and vice versa (See Ngo et al., 2014). It is noteworthy that in recent years, the AUC under the ROC has been advocated as an effective and useful measure for comparing predictive accuracy because it is not affected by differential base rates (Mossman, 1994; Rice and Harris, 1995).

Table 1. Descriptive Statistics of Variables for the Sample

Importation Predictors	Mean	SD	Min	Max
Age	34.54	9.97	17	84
Age At First Arrest	18.32	10	65	65
Number of Prior Arrests	5.97	8.19	0	87
	%	Min	Max	
Gender (1=male; 0=female)	0.83	0	1	
Race (1=black; 0=non-black)	0.44	0	1	
Marital Status (1=married; 0=not married)	0.16	0	1	
Education Prior to Incarceration				
0=Less Than High School	12.60	0	3	
1=High School	74.80			
2=Some College	11.50			
3=College or Graduate Degree	1.00			
Employment Prior to Incarceration				
0= Unemployed	32.40	0	3	
1= Employed part-time	11.70			
2= Employed full-time	55.90			
Homeless Prior to Incarceration (1=yes; 0=no)	0.10	0	1	
Current Sentence				
Violent Offense (1=yes; 0=no)	0.47	0	1	
Property Offense (1=yes; 0=no)	0.18	0	1	
Drug Offense (1=yes; 0=no)	0.23	0	1	
Public Disorder Offense (1=yes; 0=no)	0.05	0	1	
Miscellaneous Offense[a] (1=yes; 0=no)	0.03	0	1	
Ever Served in the U.S. Armed Forces (1=yes; 0=no)	0.08	0	1	
Ever Used Heroin or Other Opiates (1=yes; 0=no)	0.25	0	1	
Ever Used Crack (1=yes; 0=no)	0.27	0	1	
Ever Used Cocaine (1=yes; 0=no)	0.45	0	1	
Ever Used Marijuana or Methamphetamine (1=yes; 0=no)	0.82	0	1	
Ever Used Other Drug (PCP, LSD, Ecstasy, Tranquillizers, Methaqualone, Other Drugs That Wasn't Mentioned; 1=yes; 0=no)	0.45	0	1	
Ever Diagnosed With a Depressive Disorder, Bipolar Disorder, Manic Depression, or Mania (1=yes; 0=no)	0.25	0	1	
Ever Diagnosed With Schizophrenia or Another Psychotic Disorder (1=yes; 0=no)	0.06	0	1	
Ever Diagnosed With a Post-Traumatic Stress Disorder (1=yes; 0=no)	0.08	0	1	
Ever Diagnosed With Another Anxiety Disorder Such As Panic Disorder (1=yes; 0=no)	0.09	0	1	

Ever Diagnosed With a Personality Disorder (1=yes; 0=no)	0.31	0	1
Ever Diagnosed With Any Other Mental or Emotional Condition (1=yes; 0=no)	0.07	0	1
Do You Consider Yourself to Have a Disability (1=yes; 0=no)	0.18	0	1
Deprivation Predictors	%	Min	Max
Spent Time in Physical Exercise in Last 24 Hours (1=yes; 0=no)	0.60	0	1
Allowed to Telephone Friends & Family (1=yes; 0=no)	0.82	0	1
Allowed to Have Visits (1=yes; 0=no)	0.94	0	1
Participated in a Religious Study Group Since Admission to Prison (1=yes; 0=no)	0.30	0	1
Participated in an Ethnic/Racial Organization Since Admission to Prison (1=yes; 0=no)	0.05	0	1
Participated in Inmate Assistance Groups Since Admission to Prison (1=yes; 0=no)	0.06	0	1
Participated in Other Inmate Self-Help Groups Since Admission to Prison (1=yes; 0=no)	0.10	0	1
Participated in Employment Counseling Since Admission to Prison (1=yes; 0=no)	0.10	0	1
Participated in Parenting or Child Rearing Skills Classes Since Admission to Prison (1=yes; 0=no)	0.09	0	1
Participated in Life Skills or Community Adjustment Classes Since Admission to Prison (1=yes; 0=no)	0.25	0	1
Participated in Other Pre-Release Programs Since Admission to Prison (1=yes; 0=no)	0.06	0	1

[a] The "Miscellaneous Offense" category includes violations of laws which could or did provide the offender with some financial gain but not a violent, property, or drug offense.

Analysis Methodology

Implementing the k-fold cross-validation approach, we ran the analysis five times for each of the classification techniques - LR, RF, and NN- using each of the five sub-datasets as the test sample and the remaining four sub-datasets together as the training sample. For example, in the first run of the analysis, subset A (500 cases)was used as the test sample while subsets B, C, D, and E were combined (2000 cases) and used as the training sample. Likewise, in the second run of the analysis, subset B was used as the test sample while subsets C, D, E, and A were combined as the training sample and so on (i.e., the combinations of the five sub-datasets are as followed with the letter on the left side represents the testing sample and the letters in the right side represent the training sample: Sample 1=A/BCDE; Sample 2=B/CDEA; Sample 3=C/DEAB; Sample 4=D/EABC; and Sample 5=E/ABCD). We employedSTATISTICA®11.0 software package to build LR, RF, and NN models and we evaluated the four performance measures (sensitivity, specificity, overall accuracy, and AUC under ROC) using the cut-off probability set at 0.5

(we selected 0.5 as the cut-off probability because 50% of the inmates in the five sub-samples were cited or found guilty of breaking major prison regulations) and the prediction probabilities obtained from STATISTICA®.

Finally, we also conducted analysis of variance (ANOVA) for each of the four performance measures to determine the overall significance of the main effects of the factors (i.e., model and technique) and the model-technique interaction effect. The ANOVA analyses were conducted using Minitab (Bower, 2000). In particular, we generated a 3^1x4^1 full-factorial design with five replications (representing the five-folds of cross-validation) and performed fixed-effects ANOVA on each of the four performance measures. Once the ANOVA results were obtained, the residuals (error terms) were subjected to model adequacy checks. The model adequacy checks involved: (i) normality test on the residuals using the histogram of the residuals, the normal probability plot, and the Anderson-Darling test; and (ii) homoscedasticity (equality of variance of the error terms) involving the residuals versus fits plot and residuals versus variables plots along with the modified Levene's test. If any violations of model adequacy are found in the residual analysis, we then subject the responses to an appropriate data transformation. Next, we perform ANOVA on the transformed responses and repeat the residual analysis. Once the model is determined to be valid and adequate, we conduct F-tests to check the significance of the main effects and interactions. If any of the main effects and/or interaction effects are found to be significant, we then employ multiple comparison procedures using Tukey's method to report and interpret the results (Bower, 2000; Montgomery, 2013; NIST, 2012; Rafter, Abell, & Braselton, 2002; Tukey, 1949).

Results

Table 2. Average Sensitivity, Specificity, Overall Accuracy, and AUC under ROC[a] values for the Classification Techniques

Technique	Sample	Performance Measure			
		Sensitivity	Specificity	Accuracy	AUC
LR	Training	.6986	.7212	.7099	.7764
	Testing	.6872	.6960	.6916	.7573
	Combined	.6963	.7162	.7062	.7726
NN	Training	.6990	.7218	.7104	.7789
	Testing	.6920	.7320	.7120	.7604
	Combined	.6976	.7238	.7107	.7752
RF	Training	.6654	.7634	.7144	.7768
	Testing	.6248	.7440	.6844	.7405
	Combined	.6573	.7595	.7084	.7695
EM	Training	.7538	.6710	.7124	.7824
	Testing	.7352	.6536	.6944	.7544
	Combined	.7501	.6675	.7088	.7768

[a] Entries are averages of the five training samples (N=2,000), five testing samples (N=500), and five total samples (N=2500)

Table 2 shows the average sensitivity, specificity, overall accuracy, and AUC under ROC values for the LR, NN, RF, and EM techniques. The reported values are the averages calculated from the five folds of the k-fold cross validation procedure and are presented separately for three sample datasets – training, testing, and total (training and testing combined).

Table 3. ANOVA Tables

Sensitivity:

Source	DF	SS	MS	F	P
Sample	4	0.0070592	0.0017648	19.13	0.000
Technique	3	0.0309888	0.0103296	111.95	0.000
Error	12	0.0011072	0.0000923		
Total	19	0.0391552			

Specificity:

Source	DF	SS	MS	F	P
Sample	4	0.0256208	0.0064052	16.23	0.000
Technique	3	0.0248256	0.0082752	20.97	0.000
Error	12	0.0047344	0.0003945		
Total	19	0.0551808			

Accuracy:

Source	DF	SS	MS	F	P
Sample	4	0.0062428	0.0015607	38.47	0.000
Technique	3	0.0020592	0.0006864	16.92	0.000
Error	12	0.0004868	0.0000406		
Total	19	0.0087888			

AUC under ROC:

Source	DF	SS	MS	F	P
Sample	4	0.0123682	0.0030920	52.91	0.000
Technique	3	0.0011531	0.0003844	6.58	0.007
Error	12	0.0007012	0.0000584		
Total	19	0.0142225			

Note: The source of variation term "Technique" is found to be statistically significant for all the performance measures.

According to Table 2, pertaining to the measure of sensitivity, the average predictive values generated for the five testing samples by the four techniques (LR, NN, RF, and EM) ranged from 0.62 to 0.74 with the EM method having the highest predictive value (0.74). For the measure of specificity, the average predictive values generated for the five testing samples by the four techniques (LR, NN, RF, and EM) ranged from 0.65 to 0.74 with the RF approach having the highest predictive value (0.74) and the NN method came close at second (with the predictive value of 0.73). In terms of overall accuracy, the average predictive values generated for the five testing samples by the four techniques (LR, NN, RF, and EM) ranged from 0.68 to 0.71 with the NN technique having the highest predictive value (0.71). Finally, with regard to AUC under ROC, the average predictive values generated for the five testing samples by the four techniques (LR, NN,

RF, and EM) ranged from 0.74 to 0.76 with the LR and NN methods having the highest predictive values (0.76). The results also reveal that the average predictive value generated by the EM method almost matched the values generated by the LR and NN approaches (Table 2).

The results for the ANOVA procedures are provided in Tables 3 and 4. From the residual analysis, we found all four performance measures - sensitivity, specificity, overall accuracy, and AUC under ROC - met model adequacy checks. Hence, we retained the original ANOVA results for these measures and according to Table 3, there were significant main effects for the four performance measures. With regard to the measure of Sensitivity, the results from Table 4 reveal that the EM method outperformed the other techniques - LR, NN, and RF - in predicting the proportion of inmates who engaged in serious misconduct (Sensitivity). Pertaining to the measure of Specificity, the results from Table 4 indicate that the RF and NN methods outperformed the EM and LR approaches in predicting the proportion of inmates who did not engage in serious misconduct. As for the performance measure of Overall Accuracy, the results from Table 4 show that the NN method outperformed the other three approaches – LR, RF, and EM – in predicting the proportion of inmates who did engage as well as did not engage in serious misconduct. Finally, with regard to the measure of AUC, results from Table 4 indicate that except for the RF method, any of the remaining approaches – LR, NN, and EM – would be an appropriate technique.

Table 4. Mean and 95% Confidence Interval of Performance Measures for 'Test' Samples for the Classification Techniques

Technique	Performance Measures							
	Sensitivity		Specificity		Accuracy		AUC under ROC	
	Mean	95% C.I.	Mean	95% C.I.	Mean	95% C.I.	Mean	95% C.I.
LR	.6872	(.678, .697)	.6960	(.677, .715)	.6916	(.685, .698)	**.7573**	(.750, .765)
NN	.6920	(.683, .701)	**.7320**	(.713, .751)	**.7120**	**(.706, .718)**	**.7604**	(.753, .768)
RF	.6248	(.615, .634)	**.7440**	(.725, .763)	.6844	(.678, .691)	.7405	(.733, .748)
EM	**.7352**	(.726, .745)	.6536	(.634, .673)	.6944	(.688, .701)	**.7544**	(.747, .762)

Note: The technique/s with significantly higher performance measures compared to the rest are highlighted in bold.

Discussion and Conclusion

In the U.S., prison administrators often rely on risk assessment instruments to place and supervise inmates, as well as manage, plan and allocate resources. Hence, any improvement in the accuracy performance of risk assessment instruments is likely to result in significant benefits for offender classification and rehabilitation, management systems, as well as public safety. Actuarial risk assessment instruments employed in correctional settings are typically based on conventional regression methods. In recent years, however,

critics of these techniques have noted their shortcomings including the "one size fits all" approach that essentially ignores individual differences in assessing risks(Steadman, Silver, Monahan, Applebaum, Robbins & Mulvey, 2000), a loss in predictive accuracy when these approaches are applied to offender populations that are different from the population originally employed to develop the model (Gendreau, Goggin & Smith, 2002; Glover, Nicolson, Hemmati, Benfield & Quinsey, 2002; Gottfredson & Gottfredson, 1986; Grove & Meehl, 1996), and the relatively high rates of false positive predictions resulted from these methods (Steadman et al., 2000). As a consequence, machine learning and data mining methods such as random forests (Breiman, 2001) and neural networks (Kartalopoulos, 1995) were proposed as alternatives to help improve the predictive accuracy of risk assessment instruments. In particular, critics of conventional regression models have argued that in addition to their abilities to automatically account for non-linear relationships, search and estimate complex interactions, and handle noisy data or data with a large number of predictors, machine learning and data mining approaches are also capable of forecasting risks in situations where the decision boundaries are complex and/or the requisite predictors may not be all available (Berk & Bleich, 2013). Notwithstanding the potential advantages of machine learning and data mining techniques over conventional regression approaches, some scholars have questioned as well as refuted the claim that non-regression methods would lead to improved predictive validity (Hamilton et al., 2014; Tollenaar & van der Heijden, 2013).

Additionally, given that safety and orderliness within prisons are potentially threatened by inmates engaging in misconduct, particularly violent or serious misconduct, research examining the relative predictive performance of traditional regression methods and machine learning techniques using the outcome of serious inmate misconduct are both crucial and warranted. Yet, to the best of our knowledge, to date, there are only two studies that have evaluated and compared the predictive performance of conventional regression methods with machine learning techniques in forecasting inmate misconduct (Berk et al., 2006; Ngo et al., 2014) and only one of these studies examines the outcome of serious inmate misconduct (Berk et al., 2006). Similarly, given the strong evidence that both the importation and deprivation perspectives are germane and essential to the inquiry and understanding of inmate infractions, comparative research studies on inmate misconduct should draw their predictors from these two theoretical models. Unfortunately, none of the prior comparative research studies on inmate misconduct (Berk et al., 2006 and Ngo et al., 2014) includes predictors drawn from both of these two perspectives.

In this study, we seek to advance the debate regarding the efficacy of traditional regression methods versus the utility of machine learning and data mining techniques in forecasting serious inmate misconduct by exploring the prospect that each technique may be more suitable for a *specific* performance measure. We also employ predictors drawn from both the deprivation and importation perspectives in our study. Specifically, we evaluate the relative performance of a conventional regression method, LR, and two machine learning approaches, RF and NN, in classifying the proportion of inmates who engaged in serious misconduct (the proportion of true positives or sensitivity), the proportion of inmates who did not engage in serious misconduct (the proportion of true negatives or specificity), the proportion of inmates who did and did not engage in serious misconduct (the proportion of both true positives and true negatives or overall accuracy),

and in presenting the tradeoff in the true positive rate as a function of the false positive rate (AUC under ROC).

We also propose and examine the combined (ensemble) predictive performance of the above classification techniques (LR, RF, and NN) via the maximum predictive values generated by them (the EM results) based on the rationale that failing to identify inmates who may engage in serious misconduct poses a greater cost than misclassifying compliant inmates (Berk et al., 2006). Given the equivocal predictive performance between regression and non-regression methods, a dearth of research examining the relative performance of multiple theoretical models using multiple classification techniques simultaneously in forecasting serious inmate conduct, and the fact that risk assessment instruments are perceived as important tools for ensuring optimal public protection and a means for enhancing consistency and equity in criminal justice decision-making, we feel our study will help advance the scholarships on inmate misconduct and comparative statistical techniques.

The results from our study appear to provide support for our premise that each of the four classification methods employed in the study (LR, RF, NN, and EM) is best suited for a *specific* predictive performance measure. In particular, we uncovered that to increase the predictive accuracy in classifying inmates who are going to engage in serious misconduct (i.e., to maximize sensitivity), the EM technique should be employed (Tables 2 and 4) and to increase the predictive accuracy in classifying inmates who are *not* going to engage in serious misconduct (i.e., to maximize specificity), the RF and NN approaches should be applied (Tables 2 and 4). On the other hand, to maximize the overall predictive accuracy or to increase the predictive accuracy in classifying inmates who are going to engage in serious misconduct as well as inmates who are not going to engage in serious misconduct, we found that the NN technique was the most suitable method (Tables 2 and 4). As for the performance measure of AUC under ROC, our results reveal that to maximize this measure, the utilization of any of the following three classification techniques, LR, NN, or EM, is adequate (Tables 2 and 4).

Given the finding generated from our study that no one technique consistently outperformed the other techniques on all four performance measures, we call on future research to further explore the differential impacts among classification techniques that are based on regression and non-regression approaches. In particular, we encourage researchers to undertake comparative studies on classification techniques to determine the types of classification techniques that are appropriate for certain types of predictors (e.g., machine learning and data mining techniques may be more appropriate than conventional regression techniques in predicting future risks using dynamic predictors), the types of predictors best suited for certain outcomes (e.g., variables drawn from the importation and deprivation models of inmate behavior may be beneficial in predicting inmate-on-inmate assaults while variables derived from the administration/situational model are more suitable in predicting inmate-on-staff assaults), and the impact of specific outcome criteria on the predictive performance of classification techniques (e.g., instead of employing a cumulative measure of serious inmate misconduct, does the inclusion of specific measures, i.e., assault on staff, assault on other inmates, etc., improve the predictive accuracy of a classification technique). We also recommend that researchers recognize and consider the relative costs associated with different types of forecasting errors in future comparative studies on classification techniques. Relatedly, we encourage researchers to explore ways to translate research findings such as ours to practical applications and actions. For instance,

synthesizing findings and results generated from their research, Monahan and colleagues (2006) developed and proposed the Classification of Violence Risk (COVR) program which is an interactive software designed to estimate the risk that a person hospitalized for mental disorder will be violent to others. The COVR serves as an actuarial tool in assisting clinicians in their everyday predictive decision making.

Further, given the advocacy for wider use of machine learning and data mining techniques but the fact that these methods are not without flaws, we encourage researchers to explore and address the "black box" nature associated with these approaches. For instance, in their recent study on offender recidivism, Zeng and colleagues (2017) proposed and developed predictive models based on machine learning approaches that are accurate, transparent, and interpretable for criminal justice practitioners to use in making decisions. Specifically, the authors employed a new machine learning method known as Supersparse Linear Integer Model (SLIM; Ustun & Rudin, 2015) and produced a set of simple scoring systems to assess different decision points across the full ROC curve. The authors reported that the SLIM scoring systems were just as accurate as the other machine learning models (i.e., CART decision trees, Random forest, SVMs, SGB, etc.) in terms of predictive accuracy, but unlike the other machine learning approaches, the SLIM scoring systems were transparent and highly interpretable.

Finally, we would be remiss if we didn't recognize the limitations associated with our study. We employed self-reported data in our study and some of the shortcomings associated with self-reported data include over reporting and/or under reporting, telescoping, and memory failure and decay. The dataset employed in our study is over ten years old and more recent data may reveal new and diverse findings. We also did not include measures from the situational and administrative control model of inmate behavior (because the measures were not available in our dataset) as well as elected not to impute missing data (because data imputation has its own issues and problems). In spite of these limitations, we hope that our efforts will provide an impetus for more comparative studies involving traditional regression methods and machine learning techniques in predicting outcomes of interest in criminology such as recidivism and future risks.

References

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative Assessment. *Criminology & Public Policy, 12*(3), 513-544.

Berk, R. A., Kriegler, B., & Baek, J. (2006). Forecasting dangerous inmate misconduct: An application of ensemble statistical procedures. *Journal of Quantitative Criminology, 22*(2), 131-145.

Berk, R. A., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistics Society*, Series A 172 (part I), 191–211.

Blevins, K. R., Johnson Listwan, S., Cullen, F. T., & Lero Jonson, C. (2010). A general strain theory of prison violence and misconduct: An integrated model of inmate behavior. *Journal of Contemporary Criminal Justice*, *26*(2), 148-166.

Bower, K. M. (2000). The ANOVA procedure using MINITAB. *Scientific Computing and Instrumentation*. Retrieved from https://www.minitab.com/uploadedFiles/Content/News/Published_Articles/paired_t_test.pdf .

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Cao, L., Zhao, J., & Van Dine, S. (1997). Prison disciplinary tickets: A test of the deprivation and importation models. *Journal of Crime Justice, 25*(2), 103-113.

Dhami, M. K, Ayton, P., & Lowenstein, G. (2007). Adaption to imprisonment: Indigenous or imported? *Criminal Justice and Behavior, 34*(8), 1085-1100.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Gendreau, P., Goggin, C. E., & Smith, P. (2002). Is the PCL-R really the "unparelleled" measure of offender-risk? A lesson in knowledge accumulation. *Criminal Justice & Behavior, 29*, 397-426.

Glover, A., Nicholson, D., Hemmati, T., Benfield, G. & Quinsey, V. (2002). A comparison of predictors of general and violent recidivism among high risk federal offenders. *Criminal Justice & Behavior, 29*, 235-249.

Goodstein, L., & Wright, K. N. (1989). Inmate adjustment to prison. In L. Goodstein, & D. L. MacKenzie (Eds.), *The American Prison: Issues in Research and Policy* (pp. 229-251). NY: Plenum.

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"* (pp. 212-290). Washington, DC: National Academy of Sciences Press.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime & Delinquency, 52*(1), 178-200.

Grann, M., & Långström, N. (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34*(1), 22-36.

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and Understanding Multivariate Statistics*. Washington, D.C.: American Psychological Association.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithm) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy and Law, 2*, 293-323.

Hamilton, Z., Neuilly, M., Lee, S., & Barnoski, R. (2014). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*. doi: 10.1007/s11292-014-9221-8.

Hanson, R. K., & Thornton, D. (2003). Notes on the development of static-2002. Department of the Solicitor General of Canada, Ottawa.

Harer, M. D., & Steffensmeier, D. J. (1996). Race and prison violence. *Criminology, 34*(3), 323-355.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Howard, P., Francis, B., Soothill, K., & Humphreys, L. (2009). OGRS 3: the revised offender group reconviction scale. *Technical Report*. Ministry of Justice, London.

Irwin, J. K., & Cressey, D. (1962). Thieves, convicts, and the inmate culture. *Social Problems*. 142-155.

Jiang, S., & Fisher-Giorlando, M. (2002). Inmate misconduct: A test of the deprivation, importation, and situational models. *The Prison Journal, 82*(3), 335-358.

Kartalopoulos, S. V. (1995). *Understanding Neural Networks and Fuzzy Logic: Basic Concepts and Applications*. New York: IEEE Press.

Lahm, K. F. (2008). Inmate-on-inmate assault: A multi-level examination of prison

violence. *Criminal Justice and Behavior, 35*, 120-137.

Liaw A., & Wiener, M. (2002). Classification and regression by random forest. *R news, 2*(3), 18-22.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology, 27*(4), 547-573.

Montgomery, D. C. (2013). *Design and Analysis of Experiments* (8th Edition). Danvers, M.A.: John Wiley & Sons, Inc.

Mossman, D. (1994). Assessing prediction of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*(4), 783-792.

Neuilly, M., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 15*(2), 154-176.

Ngo, F. T., Govindu, R., & Agarwal, A. (2014). 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.

NIST (2012). NIST/SEMATECH e-Handbook of Statistical Methods. Gaithersburg, Maryland: National Institute of Standards and Technology. Retrieved from http://www.itl.nist.gov/div898/handbook.

Paterline, B. A., & Petersen, D. M. (1999). Structural and social psychological determinants of prisonization. *Journal of Criminal Justice, 27*(5), 427-441.

Rafter, J. A., Abell, M. L., & Braselton, J. P. (2002). Multiple Comparison Methods for Means. *SIAM Review, 44*(2), 259–278.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*(5), 737-748.

Ridgeway, G. (2013). Linking prediction and prevention. *Criminology & Public Policy 12*(3), 545-550.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Sorensen, J., Wrinkle, R., & Gutierrez, A. (1998). Patterns of rule-violating behaviors and adjustment to incarceration among murderers. *The Prison Journal, 78*(3), 222-231.

StatSoft Inc. (2008). Data mining, predictive analytics, statistics, StatSoft electronic textbook. http://www.statsoft.com/textbook.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., & Mulvey, E. P. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law & Human Behavior, 24*, 83-100.

Steiner, B., Butler, H. D., & Ellison, J. M. (2014). Causes and correlates of prison inmate misconduct: A systematic review of the evidence. *Journal of Criminal Justice, 42*, 462-470.

Steiner, B., & Wooldredge, J. (2008). Inmate versus environmental effects on prison rule violations. *Criminal Justice and Behavior*, *35*, 438-456.

Steinke, P. (1991). Using situational factors to predict types of prison violence. *Journal of Offender Rehabilitation, 17*(1-2), 119-132.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285-1293.

Sykes, G. M., & Messinger, S. L. (1960). The inmate social systems. In R. Cloward (Ed.), *Theoretical Studies in Social Organization of the Prison* (pp. 5-19). NY: Social Science

Research Council.

Tasca, M., Griffin, M. L., & Rodriguez, N. (2010). The effect of importation and deprivation factors on violent misconduct: An examination of black and Latino youth in prison. *Youth Violence and Juvenile Justice, 8*, 234-249.

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which methods predict recidivism best? A comparison of statistical, machine learning and data mining predictive models. *Journal of Royal Statistical Society: Series A (Statistics in Society), 176*, 565-584.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics 5*(2), 99-114.

Ustun, B., & Rudin, C. (2015) Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*, 349–391.

Wooldredge, J. D. (1991). Correlates of deviant behavior among inmates of U.S. correctional facilities. *Journal of Crime and Justice, 14*(1), 1-25.

Wright, K. N. (1991). A study of individual, environmental, and interactive effects in explaining adjustment to prison. *Justice Quarterly,* 8(2), 217-242.

Zeng, J., Ustun, B., and Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of Royal Statistical Society: Series A (Statistics in Society), 180*, 689-722.