Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2013)
176, *Part* 2, *pp.* 565–584

Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models

N. Tollenaar

Ministry of Security and Justice, The Hague, The Netherlands

and P. G. M. van der Heijden

University of Utrecht, The Netherlands

[Received January 2011. Revised March 2012]

Summary. Using criminal population conviction histories of recent offenders, prediction models are developed that predict three types of criminal recidivism: general recidivism, violent recidivism and sexual recidivism. The research question is whether prediction techniques from modern statistics, data mining and machine learning provide an improvement in predictive performance over classical statistical methods, namely logistic regression and linear discriminant analysis. These models are compared on a large selection of performance measures. Results indicate that classical methods do equally well as or better than their modern counterparts. The predictive performance of the various techniques differs only slightly for general and violent recidivism, whereas differences are larger for sexual recidivism. For the general and violent recidivism data we present the results of logistic regression and for sexual recidivism of linear discriminant analysis.

Keywords: Data mining; Linear discriminant analysis; Logistic regression; Machine learning; Prediction; Predictive performance; Recidivism

1. Introduction

Risk assessment is a widespread phenomenon in forensic contexts. A plethora of violence, sexual and general recidivism risk instruments have been developed from scratch as well as continually evaluated and compared, on various specific and general criminal target populations.

An important distinction in these instruments is whether they use static or dynamic information. *Static* offender characteristics are attributes that cannot change (e.g. age at first conviction) or change in only one direction (like age and number of previous convictions). *Dynamic* characteristics are subject to change, like having a job or other type of education. If the goal of an instrument is actuarial risk assessment, static factors will usually suffice and be the best choice. If one is interested in change after an offender has been provided with an intervention and change in risk is assessed, dynamic factors are required.

One widely applied actuarial scale for the general population of prison and probation, the offender group reconviction score (OGRS), was developed in the UK. Probation staff and corrections researchers have been using it since the 1990s. The scale was required to be quick and

Address for correspondence: N. Tollenaar, Research and Documentation Centre, Ministry of Security and Justice, Schedeldoekshaven 131, 2311 EM, Den Haag, Zuid-holland, The Netherlands.
E-mail: n.tollenaar@minvenj.nl

© 2012 Royal Statistical Society 0964–1998/13/176565

easy to score, to be predictively accurate and to give a view of the criminogenic needs of the offender. It was shown that with the limited number of five static predictors a reasonably accurate estimate of the (2-year) group reconviction probabilities could be constructed (Copas and Marshall, 1998). Since then, the OGRS has undergone two major revisions expanding it with more risk factors (Maden *et al.*, 2005; Howard *et al.*, 2009).

In the Netherlands, a similar instrument was developed to assess the risk of recidivism and influenceability of adult (e.g. 18 years or older) offenders quickly, the 'Quickscan' (de Ruiter and de Jong, 2006), which includes two subscales: one for the static risk (Wartna *et al.*, 2009) and a scale for dynamic risk of recidivism. The results of the quick scan are intended for the Public Prosecutor to establish the possibilities of influencing the offender's behaviour. The requirements of this instrument were twofold: it needed to be administered quickly and it should include only the limited information available for probation officers. Quickscan needed to consist of a dynamic part to supply a brief view of the criminogenic needs and a static part to have an accurate estimate of the recidivism. The latter led to the development of scale 1 of the quick scan that consisted of the assessment of the static risk of recidivism, called StatRec. This scale is similar to the original OGRS but includes an additional risk factor, namely country of birth, and targets the 4-year instead of the 2-year reconviction rate. It was constructed on the basis of the total population of Dutch adult offenders in 1999 with a valid settlement by the court or Public Prosecutor (Wartna *et al.*, 2009). It turned out that the predictive performance of the scale was very similar to the OGRS and is concurrent with the performance of the majority of risk assessment instruments that are used in criminological research in diverse offender populations.

In making pre-sentence reports for sexual and violent offenders, Dutch probation also had a need for a quick assessment of sexual and violent recidivism, based on the limited available information. This need is in line with developments in actuarial risk assessment in the UK. One of the aims of this paper is to take a first step towards such instruments for the Netherlands. Taylor (1999) developed a revised OGRS that also predicts violent and sexual recidivism by using static information. Two other important sexual recidivism scales, the popular static 99 (Hanson and Thornton, 1999) and its successor static 2002, also contain mainly static information. The risk matrix 2000 (Thornton *et al.*, 2003) is another actuarial tool used by the England and Wales probation service. This instrument has separate scales for violent and sexual recidivism risk. Besides static factors it also includes victim information.

Traditionally, the most suitable method for estimating and predicting the probability of an event at a single point in time is standard linear logistic regression (Hosmer and Lemeshow, 2000). This model has been used for both the OGRS and StatRec. It is a statistical model that is estimated via maximum likelihood, which models the logit of the probability of recidivism linearly. Since the 1960s, however, many different approaches have been developed in the field of machine learning and data mining for predicting a binary choice, that can generate individual probabilities. From a theoretical point of view these methods have several advantages and disadvantages when compared with each other and classical statistical methods. For instance, instead of maximizing the likelihood of a probability model, these methods typically tend to optimize the predictive performance directly. 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. The emphasis of these modern approaches lies more on prediction than on explanation and interpretability of covariate effects.

There have been many attempts, across a multitude of disciplines and situations, to establish which model performs best (King *et al.*, 2006; Lim *et al.*, 1998, 2000; Caruana and Niculescu-Mizil, 2006). Jamain and Hand (2008) performed an extensive meta-analysis on the comparative

performance of classifiers. Although they noted that the comparability of the different studies is questionable, they found that linear discriminant analysis (LDA) was a good overall classifier as well as C4.5, a tree classifier. LDA and logistic regression are known to give very similar results, even if LDA is used inappropriately, for instance when modelling non-normal data, or when the predictors are of a categorical measurement level (Hastie *et al.* (2009), page 128). Generally, the results of the different studies indicate that there is no 'best model' for all data sets and situations, and that with a specific data set it is necessary to establish which model is best for the relevant data (Jamain and Hand, 2008).

There have been attempts to improve the predictive performance of models in the domain of criminology by using machine learning and/or data mining techniques, such as neural networks and single-classification trees (Caulkins *et al.*, 1996; Yang *et al.*, 2010). The more modern, popular and promising techniques, however, like adaptive boosting (Friedman *et al.*, 2000), random forests (Breiman, 2001) and support vector machines (SVMs) (Cortes and Vapnik, 1995) have not yet been considered. The most recent study that compared techniques in the recidivism domain was from Yang *et al.* (2011). They used the items of HCR-20 (Webster *et al.*, 1997) to predict reconvictions for violence. They compared just two alternative approaches for a binary outcome to logistic regression (classification and regression trees and neural networks), whereas survival analysis variants of these techniques should have been used because follow-up periods in their data varied from several days to 4 years.

In machine learning and data mining, performance is mainly measured by the classification error or its complement accuracy, whereas the performance of prediction scales is routinely evaluated with receiver operating characteristic (ROC) analysis (see Mossman (1994)). The latter is a general non-parametric method of establishing the ability to discriminate between two classes by using any sort of scale by comparing the ranks of each pair of class 1–class 2 individuals with respect to the scale or probability value. It can result in a graph which plots the amount of false positive results against the amount of false negative results, called the ROC curve. The higher the area under the curve, AUC, the 'better' the model. During the past 10 years, the importance of this measure of discrimination has been discovered by the machine learning community as a means of assessing predictive validity (Provost and Fawcett, 2001; Ling *et al.*, 2003). This led to the development of algorithms optimizing AUC directly (Cortes and Mohri, 2003; Clemençon and Vayatis, 2010).

In this paper we want to evaluate a broad set of prediction models on a broad set of performance metrics. We use multiple metrics so we shall just choose the best ranking or the highest accuracy model. In this paper we want to find out whether using data mining, machine learning and modern statistical models leads to a significant improvement in predictive performance of reconviction over classical statistical methods. The standard classical methods are logistic regression and discriminant analysis. We have attempted to construct and select models that optimally use the information at hand and thus generate the optimal prediction for three different types of recidivism. These types of recidivism are general recidivism, violent recidivism and sexual recidivism. General recidivism is defined as a reconviction of any type following an index case of any type, violent recidivism is a reconviction for a violent crime following an index case consisting of minimally one violent offence and sexual recidivism is defined as a reconviction for a sexual offence following a sexual index offence. This study will compare the performance of a range of classical and modern methods of predictive modelling, using only static offender information.

The data are available from the first author on request. The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Method

2.1. Source of data

The StatRec scale uses only static information that is available in the Dutch offender's index. This index is an automated, encrypted and anonymized version of the judicial information that probation officers and others can request for an offender, the 'Judicial documentation system'. It provides a chronological overview of all criminal cases in which a physical or legal entity has been suspected of a criminal offence. Criminal cases are registered for people from ages 12 years and up, as the minimal age of criminal responsibility is 12 years. For each criminal case, it is recorded when the case was registered and at what court, along with details of the crimes to which it related and how the case was dealt with. This information is also at the disposal of probation workers so it can be used to calculate a risk score. Our study concerns people aged 18 years or older. This means that we restrict our scales to people falling under adult law, with a custodial sentence, a non-custodial sentence or a disposal from the Public Prosecutor's office. In the Netherlands the Public Prosecutor may dispose of cases by offering fines, community service orders and training programmes. If the perpetrator does not accept this, the case will appear in court. The Public Prosecutor disposals are usually for less serious crimes and are supposedly similar to police cautions in England and Wales.

2.2. Definition of recidivism

Recidivism is defined as proven reoffending, and the unit of recidivism is the criminal case. A criminal case can consist of one or more indictable offences. The time to recidivism is measured as the time from registering the index criminal case to the minimal offence date of the crimes in the follow-up case. If the index case is sanctioned with a prison term, the estimated time to release was subtracted from the duration of recidivism. General recidivism includes motoring offences like leaving the scene of an accident and driving under the influence of alcohol or drugs. Cantonal court cases (i.e. misdemeanours) are not counted. Violent recidivism includes threatening, extortion, property crime by using violence and/or threatening, assault and battery, murder, homicide and culpable homicide but excludes destruction of property. Sexual recidivism includes rape, violation, indecent assault (with minors or subordinates), sexual contact below age 16 years, age 12 years and sexual contact with someone unconscious or legally incapable. It excludes indecent exposure and bigamy.

This definition of recidivism is an underestimation of the actual recidivism of the offenders as a large body of crime goes undetected: the so-called *dark number*. Thus, the recidivism defined can be seen as a lower bound of actual recidivism. Nevertheless it is the crime that the judicial system deals with. The reliability of these data is very good. If actual recidivism is preferred, a representative sample of offenders would have to be measured on self-reported crime. Besides being relatively costly to gather, these data have their own disadvantages like different sorts of biases and reliability issues. Some of them include attrition from the study, failure to report certain types of offending and overreporting other types of offending.

2.3. Data selection

StatRec predicts the 4-year reconviction rate as a risk score. Therefore, we take the most recent year of the criminal case that is available, which is 2005. The resulting data set consists of all adult perpetrators who were found guilty during criminal proceedings ending in 2005. For general recidivism and violent recidivism, memory requirements and limited computation time forced us to select two random subsets of these data sets ($N = 159\,298$ and $N = 25\,041$ respectively): 10000 for estimation of the models (the training set) and 10000 for validation of models (the

test set). For sexual recidivism, because the prevalence of sexual recidivism is relatively low, we include all sexual offenders in 2005 and only the sexual offenders in 2006 who have at least 4 years of follow-up time, to obtain a larger data set ($N = 1\,332$). 30% of the sexual recidivism data is kept as a test set. The recidivism is defined as the time from administration of the penal case to the date of first offence in a recidivism case. As detention data are not yet widely available, the follow-up time is corrected for the estimated time in prison, based on the duration of the prison term in the sentence. Cases that had less than 4 years follow-up time due to this correction were dropped. For the general recidivism data this was 1.2%, for violent recidivism 1.9% and 1.4% for the sexual recidivism data.

2.4. Variables used

The set of variables differs for the three data sets. The model of general recidivism needs to be very simple and fast to score. Therefore only a small selection of variables should be included in the model. However, we allow the models for violent and sexual recidivism to be more exhaustive with respect to the information that is available in the judicial database because predicting these types of recidivism is inherently more difficult. Table 1 contains a list of all the variables that were used in the three submodels and their sample statistics.

As noted in the previous section, the Public Prosecutor's disposals are also counted as index convictions and reconvictions. The consequence is that the general recidivism population includes less serious offenders than fall under for instance the UK definition of offenders. For example, this becomes clear in the relatively high age of first conviction.

To account for non-linearity the number of previous convictions will be included in the models transformed by taking the natural logarithm. The possible non-linear effect of age will be dealt with by including also age squared, after centring on the mean in the data. The number of previous disposals of different types may be subject to differences in criminal policy. Therefore the effect that is estimated by the models might differ over time, posing a potential threat to temporal validity. Because the models will be updated regularly we doubt whether this will have a large effect on the predictions. The three data sets yield three conditions for estimating a prediction model. The general recidivism data have a large N, a small number of parameters and a high base rate, and the violent recidivism data have a large N, a large number of parameters and a medium base rate, whereas the sexual recidivism data have a small N, a large number of parameters and a low base rate.

2.5. Criteria for predictive performance

What defines 'predictive performance'? There are many criteria to establish the performance of a prediction model, depending on what the model is used for. In prognostic modelling there are three dimensions which a good prognostic model should perform well on (see for instance Vergouwe (2003)). These are calibration of the predicted probabilities, discrimination and clinical usefulness.

Calibration of the predicted probabilities ensures that the probabilities generated correspond well to the actually observed outcomes. This means that the average probability is approximately the same as the proportion of observed positive outcomes.

Discrimination means that the model can distinguish observations with and without the outcome well, i.e. individuals with higher probabilities are more often recidivists than individuals with lower probabilities. Discrimination is usually quantified with AUC, the area under the ROC curve (Hanley and McNeil, 1982).

Clinical usefulness is important when individual decisions need to be made and requires a cut-off value for the probability. Two typical choices for this cut-off are the value 0.5 and the

Table 1. Variables used in the three prediction domains of recidivism

Variable	*Results for the following types of recidivism:*		
	General recidivism	*Violent recidivism*	*Sexual recidivism*
Total *n*	20000	20000	1332
Training *n*	10000	10000	932
Test *n*	10000	10000	400
4-year base rate (%)	40.1	23.7	5.5
Gender: female (%)	14.7	9.8	—
Age in years (mean)	35.2	34.7	38.7
Age of first conviction (mean)	27.1	25.2	30.0
Most serious type of offence (%)			
Violence	14.3	90.8	—
Sexual	0.6	0.5	—
Property crime with violence	1.4	1.1	—
Property crime without violence	23.6	3.4	—
Public order	11.2	3.2	—
Drug offence	6.6	0.4	—
Motoring offence	28.9	0.0	—
Miscellaneous offence	13.5	0.6	—
Country of birth (%)			
Netherlands	70.1	89.2	70.1
Morocco	3.1	5.2	2.7
Dutch Antilles or Aruba	3.0	4.8	5.0
Surinam	4.8	6.3	5.3
Turkey	3.1	4.7	2.6
Other western countries	8.6	6.0	5.9
Other non-western countries	7.3	9.0	8.4
Offence type present in index case (%)			
Violence component (0–1)	—	—	12.2
Sexual component	—	0.6	—
Property crime with violence	—	1.3	1.6
Property crime without violence	—	4.4	2.9
Public order	—	13.8	5.4
Drug offence	—	1.1	0.8
Motoring offence	—	1.5	0.7
Miscellaneous offence	—	6.9	8.8
Criminal history counts (mean)			
Conviction density†	0.4	0.3	0.4
Number of previous convictions‡	1.6	2.0	1.6
Previous violence offences	—	0.9	0.5
Previous sexual offences	—	0.0	0.2
Previous property crime with violence offences	—	0.20	0.13
Previous property crime offences	—	2.42	1.66
Previous public order offences	—	0.99	0.65
Previous drug offences	—	0.18	0.11
Previous motoring offences	—	0.75	0.57
Previous miscellaneous offences	—	0.37	0.25
Previous prison terms	—	0.99	0.68
Previous community service orders	—	0.42	0.28
Previous fines	—	1.17	0.88
Previous Public Prosecutor's disposals	—	0.48	0.33

†This variable is computed as $\sqrt{\{\text{number of convictions}/(\text{careerlength}+1)\}}$. Copas and Marshall (1998) were the first to propose the use of the square-root-transformed conviction density as a powerful predictor.
‡To aid fast scoring on the part of the probation officer, in the models the number of previous convictions is partially categorized. In the actual models, the natural logarithms of 0–9 (+1) convictions are modelled linearly, and the categories of 11–20 and 21 and over are included as dummy variables. As the conviction density also needs the number of previous convictions, the median number of convictions in these categories is used in its computation.

base proportion of the outcome in the data, which is also known as the 'base rate' in criminological literature. Some indicators of clinical usefulness include accuracy of prediction (i.e. the percentage correctly classified), the sensitivity (the percentage of observations with the positive outcome correctly classified) and the specificity (the percentage of observations with negative outcome correctly classified). The different indicators measure different aspects of predictive performance and are not always simultaneously optimal, which has been empirically demonstrated by Caruana and Niculescu-Mizil (2004), who proposed combining three different indicators, namely AUC, accuracy and the root-mean-squared error RMSE, into one performance measure, the squared error, accuracy and ROC area, SAR, to establish optimality in different domains and to create a more 'robust' measure. They found that SAR correlated highest with a range of nine performance metrics.

For this study, we have computed a range of performance criteria to be able to provide a detailed picture of which method should be preferred in a range of situations. These criteria are as follows.

(a) AUC (the area under the ROC curve): some consider an AUC of more than 0.75 as 'large' (Shapiro, 1999; Dolan and Doyle, 2000), whereas Hosmer and Lemeshow (2000) considered AUC-values starting at 0.70 'acceptable', those from 0.80 upwards are considered 'excellent' and values of 0.90 and higher are 'outstanding'.

(b) Accuracy ACC: accuracy is simply the percentage of cases correctly classified and requires a threshold on the predicted probabilities for classifying instances as positive or negative. The accuracy is the sum of the true positive and the true negative values divided by the sample size. Absolute values of accuracy can be quite misleading in the case of a low base rate. The sheer majority of observations is classified correctly if the most prevalent outcome is chosen in all instances. This is usually known as the 'no-information rate'. We shall only compare accuracies within data sets. The accuracy is the complement of the error rate that is often used in classification studies.

(c) RMSE (root-mean-squared error): the root-mean-squared error is well known from regression analysis. It is a summary measure for the discrepancy between the observed and predicted value of the dependent variable. It can also be used for a binary dependent variable. It is described by

$$\sqrt{\left\{\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{n}\right\}}.$$

High values of RMSE show that there is a relatively large discrepancy between the observed outcome and the expected outcome.

(d) SAR: the definition of SAR (Caruana and Niculescu-Mizil, 2004) is

$$\mathrm{SAR} = (\mathrm{AUC} + \mathrm{ACC} + 1 - \mathrm{RMSE})/3.$$

(e) Overall calibration error CALerr: this is the mean calibration error over the 0–1-probability range. Calibration error is the amount of discrepancy between the expected probabilities and the proportion of the actual outcome. If the probabilities correspond well to the observed outcome, the observed proportion of outcomes would be close to the mean of the probabilities in the group. We use a window of 100 observations, which is shifted over the range and then averaged.

(f) Local calibration error (calibration error around the cut-off, CAL(0.5)): if a prediction model is used for individual decision making, the calibration should be the best around the cut-off score and deviation on the low and high probabilities are of lesser concern. We

propose to use a measure that uses only the domain around the threshold. This measure is calculated in the following way.

(i) 10% of the data are selected around the chosen cut-off value.
(ii) Two bins are created around the cut-off. The left bin is 5% on the left of the cut-off. The right bin is 5% including and larger than the cut-off.
(iii) The resulting squared residuals are summed only if the left bin has an overprediction and/or the right bin has an underprediction.

All these criteria will be used to describe the predictive performance of the models. SAR will be the decisive criterion on which we shall base the selection of the final model, because it takes into account all three dimensions of predictive validity, namely discrimination, calibration and accuracy. ACC and thus SAR will be calculated on the 0.5-cut-off, because it is used most often and has a statistical interpretation. When a probability is larger than 0.5 it is more likely to have a positive outcome than a negative outcome.

2.6. Models

The following 11 models are compared with respect to their predictive performance.

(a) Linear logistic regression (logreg method) (see Hosmer and Lemeshow (2000)): logistic regression models the logit of a binomial probability linearly via maximum likelihood. It stems from the 1930s and is a special case of the family of generalized linear models (McCullagh and Nelder, 1989).
(b) The multivariate adaptive regression spline (MARS) method (Friedman, 1991): the MARS method is a form of non-parametric regression. This class of models automatically fits non-linear and interaction effects. The maximum number of estimated terms and the degree of interaction need to be specified in advance. They can be fitted on the entire class of generalized linear models (McCullagh and Nelder, 1989). In our case the logit link function and the binomial family are used.
(c) LDA (Fisher, 1936): this is the classical method of predicting the two-class model. It is a standard linear regression with a dummy independent variable and has more assumptions than logistic regression, regarding the underlying class distribution and equality of the class covariances. It generates a linear decision boundary. Currently, dozens of variants of discriminant analysis exist that are designed for specific conditions (for instance in the cases of highly correlated predictors or more predictors than observations).
(d) Flexible discriminant analysis (FDA) (Hastie *et al.*, 1993): this is essentially a non-parametric LDA. It uses the MARS method (see above) for estimating a non-linear decision boundary.
(e) Recursive partitioning (rpart method): a classification tree is a method to partition recursively the covariate space into maximally separated groups with respect to the dependent variable (the classification and regression tree algorithm by Breiman *et al.* (1984)). At each split the next split is searched that gives the maximum increase in the criterion that describes separation. Mostly this criterion is the Gini index. This is defined as $1 - \Sigma_j p^2$, where j is an iterator over the number of classes and p is the proportion within the class. If a threshold value for this increase is not crossed, splitting is stopped.
(f) Adaptive boosting (or adaBoost) (Freund and Schapire, 1995): this is one of the so-called 'ensemble methods'. Instead of using a single tree for predictions, an ensemble of trees is built, in this case of classification and regression trees (see above). The resulting predictions from all single trees are averaged to obtain a single prediction. The actual 'boosting' of each subsequent tree is obtained by increasing observation weights on

misclassified cases at each 'iteration' of the tree. This has been shown to be an excellent way of obtaining precise and discriminative probabilities without having to specify a model explicitly and is often called the 'best off-the-shelf classifier' (Breiman, 1998).

(g) The logitBoost (Friedman *et al.*, 2000) algorithm optimizes the logistic loss function so that it behaves like a generalized additive model. The base model that is used here is a decision stump (only the first split of a regression tree).

(h) Neural networks with one single hidden layer (nnetmethod): neural networks have been developed with the neurons in the brain as a model. They can be used to approach any function of arbitrary form. Neural networks are very sensitive to the tuning parameters and prone to overfitting the training data (see for example Hastie *et al.* (2009), page 398).

(i) Linear SVMs (Cortes and Vapnik, 1995): these classification algorithms are closely related to neural networks. The algorithm tries to find the hyperplane that separates the positive and negative examples with a certain margin. The vectors of observation points that lie close to this hyperplane within the margin are called the support vectors. The margin between these support vectors is maximized and misclassified examples are allowed or disallowed by varying the cost parameter C. The penalty of these errors is equal to the distance to the decision boundary times C. SVMs are particularly sensitive for data sets consisting of sharply unequal class sizes (also known as 'class imbalance'), because the cost parameter does not account for this. This can be overcome by using class weights to obtain balance.

(j) K-nearest-neighbours classification (K-nn method): K nearest neighbours are a rather simple method of classification (Fix and Hodges, 1951). The algorithm finds a group of k observations closest to the test observation in terms of Euclidean distance. The proportion of the votes for the positive class is then returned as a probability.

(k) Partial least squares (PLS) (see Wold (1985)): this method extracts orthogonal latent factors for the covariates and the outcome and thus indirectly models the influence of the predictors on the outcome. Component scores are estimated to optimize the covariance between the scores and the response variable. It has less strict model assumptions than for instance LDA.

Some of the aforementioned methods are not designed to generate an estimated probability as an outcome, like SVMs, which instead give the distance to the decision boundary. It is, however, possible to transform these distances by using a non-linear function fitted on the model samples using Platt's calibration (Platt, 2000). This is obtained by fitting a sigmoid function to the predicted values of the estimation data set with respect to the actual outcome on the estimation data set. The resulting function is then used to transform the predictions in the test sample.

It might be that calibration improves the estimated probabilities of the remaining methods, so every other model's probabilities are calibrated as well. These calibrated models are only shown when calibration improves the model's performance in terms of SAR. The complete tables are available as on-line supporting information.

2.7. Model selection within method

Many of the algorithms and models that are used require one or more tuning parameters to be set that have a large effect on their performance. Therefore within each type of model we must select the model that performs best in a range of tuning parameters. To select the best model over its range within the estimation set, we use tenfold cross-validation using accuracy as the main criterion. We chose accuracy because many of the models included are built to optimize accuracy. The final selected model is tested on the test set of the data, using SAR as the

final criterion. We shall prefer the classical statistical methods logistic regression or LDA if the modern methods do not perform substantially better. The following range of tuning parameters in the software was used for all three data sets:

(a) recursive partitioning tree—maximum split depth of the tree in 2, 3, 4, 6, 10, 13, 15, 16 and 28, and a threshold of 0.001;
(b) single-layer feed-forward neural networks—1, 2, 4, 8 and 16 hidden nodes and decay of 0, 0.01, 0.001, 0.0001 and 0.00001; all combinations of the resulting grid are tried;
(c) MARS or FDA—maximum number of terms in the model of 15, 20, 25, 30, 35, 40 and 45, and maximum interaction degree—1, 2, 3; all combinations of the resulting grid are evaluated;
(d) adaptive boosting (adaBoost method)—50, 150, 200, 250, 300, 350, 400, 450 and 500 boosting iterations, tree maximum depth of 30 and a step size of 0.1;
(e) logistic boosting (logitBoost method)—50, 150, 200, 250, 300, 350, 400, 450 and 500 boosting iterations;
(f) linear SVM—a cost parameter of 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 2, 10 and 100, and class weights of (2,3), (1,3) and (1,20) for general, violent and sexual recidivism respectively;
(g) K-nearest-neighbours classification—$k = 5, 7, 9, 11, 13, 15, 17, 19, 21, 23$;
(h) PLS—number of latent components 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10.

Logistic regression and LDA do not need to be tuned.

2.8. Software used

All calculations were performed in R version 10.1 (R Development Core Team, 2008). A subset of the performance criteria was computed by using the `ROCR` package version 1.0-2 (Sing *et al.*, 2005). The actual training of the different models was performed by using the `caret` package (Kuhn, 2010).

3. Results

We shall successively discuss the benchmarks for the general recidivism model, the violence recidivism model and the sexual recidivism model and then make an overall judgement on the three domains of the predictive performance.

3.1. General recidivism

Table 2 shows the performance indicators of the 11 seven-variable models compared on the test data. For the SVM and PLS models the results are based on the calibrated probabilities. We shall first describe the performance with respect to the overall SAR(0.5) measure, followed by a treatment of the other metrics.

Following SAR at cut-off 0.5, the logistic regression model seems to outperform all other models. Its SAR(0.5) is, however, virtually identical to that of the LDA, FDA and MARS models. The slight differences in SAR-performance are mainly caused by minor differences in its components, AUC, RMSE and ACC. As indicated above, when the performance of models is approximately equal, we prefer the standard models over modern statistical methods, so in this case logistic regression is the preferred model. The choice for logistic regression is also supported by its performance on the other criteria where it is also doing very well and sometimes best.

Table 2. Test data performance of 11 *general* recidivism models on six indicators†

Model	*Results for the following criteria:*					
	Cut-off independent			*Cut-off = 0.5*		
	AUC	*RMSE*	*CALerr*	*ACC(0.5)*	*SAR(0.5)*	*CAL(0.5)*
logreg	0.776	0.430	0.034	0.728	0.692	0.000
MARS	0.773	0.431	0.031	0.729	0.691	0.000
LDA, calibrated	0.776	0.430	0.035	0.728	0.691	0.004
FDA	0.774	0.431	0.035	0.727	0.690	0.007
rpart	0.744	0.438	0.041	0.723	0.676	0.025
adaBoost	0.765	0.436	0.042	0.720	0.683	0.016
logitBoost, calibrated	0.710	0.473	0.036	0.680	0.645	0.000
PLS, calibrated	0.773	0.432	0.036	0.726	0.689	0.000
K-nn, calibrated	0.708	0.459	0.044	0.670	0.639	0.010
nnet	0.773	0.432	0.030	0.724	0.688	0.009
SVM, calibrated	0.768	0.436	0.054	0.725	0.686	0.011

†The no-information rate (the proportion correct if all instances are classified in the most prevalent class) of the test sample is 0.618.

The neural network model does best in terms of overall calibration error. The worst performing models are logitBoost and *K*-nn. Their AUC and accuracy values are largely behind the other models. For *K*-nn, this is not surprising given the coarse fashion in which this method operates. The calibration error around the cut-off 0.5 is zero for logistic regression, MARS and PLS. The largest error is made by the logitBoost method.

Overall, the weakest model is the logitBoost model with regression stumps. It is apparent that the single-split trees are too simple for these data. The *K*-nn model proves to be another model performing weakly in these data. Its performance criteria cannot be considered to be competitive.

3.2. Violent recidivism

Table 3 shows the performance indicators for the violent recidivism models on the test data. Following SAR at cut-off 0.5, the calibrated logistic regression model seems to outperform all other models. It is, however, almost identical to SAR(0.5) of discriminant analysis, adaBoost and the calibrated PLS models. These differences can be attributed to only very slight differences in AUC, ACC(0.5) and RMSE. Exceptionally poorly performing models, judging by SAR(0.5), are the rpart, logitBoost, *K*-nn and SVM.

The accuracy of classification is highest by LDA but it is indistinguishable from logistic regression. As opposed to the general recidivism model, overall calibration (CALerr) is best for the logistic regression model. The calibration around the cut-off scores seems to be the best for the MARS, LDA and neural network models. adaBoost has the lowest RMSE, but it is nearly identical to the RMSE of logistic regression.

3.3. Sexual recidivism

The results of the model selection on the accuracy criterion tends to favour the model that put all observations in the largest class. Therefore we use Cohen's κ (Cohen, 1960) as the model selection criterion in finding the optimal tuning parameters in the sexual recidivism models.

Table 3. Test data performance of 11 *violent* recidivism models on six indicators†

Model	*Results for the following criteria:*					
	Cut-off independent			*Cut-off = 0.5*		
	AUC	*RMSE*	*CALerr*	*ACC(0.5)*	*SAR(0.5)*	*CAL(0.5)*
logreg, calibrated	0.739	0.396	0.040	0.781	0.708	0.004
MARS	0.736	0.397	0.033	0.780	0.707	0.000
LDA	0.739	0.397	0.039	0.781	0.708	0.000
FDA, calibrated	0.736	0.400	0.049	0.779	0.705	0.006
rpart	0.702	0.404	0.036	0.770	0.689	0.076
adaBoost	0.740	0.395	0.033	0.777	0.708	0.016
logitBoost, calibrated	0.671	0.408	0.035	0.772	0.678	0.025
PLS, calibrated	0.741	0.396	0.038	0.777	0.708	0.020
K-nn	0.678	0.411	0.048	0.766	0.678	0.022
nnet	0.720	0.402	0.043	0.776	0.698	0.000
SVM, linear	0.726	0.405	0.058	0.770	0.697	0.037

†The no-information rate (the proportion correct if all instances are classified in the most prevalent class) of the test sample is 0.761.

Judging by SAR(0.5), we would choose the LDA model. The second-best model for the sexual data is the calibrated PLS model, closely followed by the SVM. Judging from the individual components of SAR, we can see that the variation is almost exlusively due to the variations in AUC. The accuracy is not usable as it never surpasses the no-information rate. The difference in performance of the models from the logistic regression is now substantially larger than in the previous two data sets. AUC for the logistic regression is considerably smaller than AUC for LDA and PLS. The relatively low RMSE does not make up for this. The MARS, FDA, rpart and *K*-nn methods seem to fail completely. Surprisingly, the overall calibration error is lowest in the logistic regression model. This model also has the lowest calibration around the 0.5 cut-off. RMSE is, however, lowest in the SVM model. The overall performance of the prediction models of sexual recidivism data is disappointing when compared with the former two data sets.

3.4. Summary of the results

The assertion of many researchers that there is no best model for all data holds true with respect to these three data sets, although some models do well on all data. Following the dimensions of prognostic models from Vergouwe (2003), we shall discuss which model would be best in each situation.

3.4.1. Discrimination

The LDA and logistic regression model discriminate well when modelling general and violent recidivism. The discrimination of the logistic regression is less good with respect to sexual recidivism. As others have found, overall the LDA models are good discriminators on all three data sets. Another good overall discriminator is the PLS model.

The single-tree recursive partitioning method should not be used if discrimination is important. It is most suitable for optimizing classification accuracy.

Table 4. Test data performance of 11 *sexual* recidivism models on six indicators†

Model	*Results for the following criteria:*					
	Cut-off independent			*Cut-off = 0.5*		
	AUC	*RMSE*	*CALerr*	*ACC(0.5)*	*SAR(0.5)*	*CAL(0.5)*
logreg	0.613	0.203	0.016	0.958	0.789	0.029
MARS	0.379	0.224	0.019	0.953	0.702	0.365
LDA, calibrated	0.725	0.202	0.021	0.955	0.826	0.094
FDA, calibrated	0.681	0.206	0.026	0.953	0.809	0.179
rpart	0.500	0.202	0.021	0.958	0.752	0.030
adaBoost	0.602	0.203	0.037	0.958	0.785	0.104
logitBoost	0.524	0.248	0.031	0.900	0.725	0.074
PLS, calibrated	0.722	0.203	0.018	0.955	0.824	0.271
K-nn	0.541	0.204	0.017	0.958	0.765	0.105
nnet	0.673	0.224	0.017	0.940	0.796	0.352
SVM, calibrated	0.711	0.199	0.022	0.958	0.823	0.234

†The no-information rate (the proportion correct if all instances are classified in the most prevalent class) of the test sample is 0.958.

3.4.2. Calibration

There is no real pattern in the performance of the models on calibration. Logistic regession does, however, tend to be well calibrated over data sets, overall as well as around the cut-off. The logitBoost method has the worst-calibrated probabilities but performs worst on all criteria. The recursive partitioning has a rather poor overall calibration.

3.4.3. Clinical usefulness

Clinical usefulness was defined as the ability of the model to predict the right class. This aspect is captured by the accuracy of the model. On these data there is no model that has a substantially better classification accuracy than the remaining models. Variants of discriminant analysis (FDA and LDA) and the MARS method seem to do well across data sets on this criterion. The 'classify all in the largest class' scheme obscures the real usefulness in low base rate models. Sensitivity and specificity can bypass this scheme by focusing on the positive and negative predictive accuracy, but they are also heavily dependent on the choice of cut-off point and calibration of the probabilities. The comparability of clinical usefulness can be assessed by taking the following point of departure. If we consider false positive and false negative results to be equally costly, we shall want to make an equal percentage of errors in both types of predicted events. In this case the amount of false positive results is as large as the amount of false negative results. This yields the predictive accuracy when the sensitivity equals the specificity. For all models this measure is given in Table 5. It does not contain the calibrated models as this measure is invariant under order preserving transformations.

Table 5 shows again that different models are optimal on different data sets. For general recidivism the LDA model has the best score on the sensitivity–specificity balanced accuracy. However, the difference between the various models is again very small. The logitBoost model remains the worst performer.

For violent recidivism, the adaBoost model is best at predicting the right class. It is, however, very closely followed by the PLS and MARS models.

Table 5. Accuracy of the models when sensitivity = specificity, test data

Model	*Results for general recidivism*	*Results for violent recidivism*	*Results for sexual recidivism*
logreg	0.704	0.672	0.587
MARS	0.705	0.676	0.464
LDA	0.705	0.673	0.660
FDA	0.704	0.676	0.681
rpart	0.690	0.653	0.500
adaBoost	0.696	0.677	0.523
logitBoost	0.671	0.645	0.486
PLS	0.705	0.677	0.705
K-nn	0.660	0.624	0.545
nnet	0.704	0.662	0.647
SVM, linear	0.699	0.671	0.602

In the sexual recidivism data, logistic regression achieves a mediocre ranking. The PLS model outperforms all models in the sexual recidivism data. If individual sexual prediction were the only goal, PLS seems to be the model of choice. The LDA model is the second best on classification accuracy. The logitBoost model is again the worst performing model, now achieving an accuracy that is worse than flipping a coin. The adaBoost model, relying on the rpart model, also suffers from poor performance in these data.

Taking all data and models together, when the sensitivity is as large as the specificity, the minimum error rate that can be achieved with these three models is about 30%.

3.4.4. Models selected

Our starting point for comparison of the models was to choose logistic regression or LDA when they did not perform worse than their modern counterparts. This turned out to be so for all data sets. In subsequently making a choice for either logistic regression or LDA, we choose logistic regression if logistic regression is doing equally well as LDA, the reason being that for logistic regression there is no need to check the assumption of homogeneous variance–covariance matrices. This leads to the choice for logistic regression for general and violent recidivism and LDA for sexual recidivism.

3.5. Description of the final models

After the model selection phase, the models that were ultimately chosen were fitted on the complete data sets. We shall describe these definitive models here.

3.5.1. General recidivism: logistic regression

The odds ratios of the seven predictor models based on 159 298 observations are provided in Table 6. Although the N is very large, some of the categories are non-significant. Table 6 shows that the conditional odds of reconviction are 32% less for women than for men. Recidivism tends to decline with age; the odds lower by 3% for each extra year. The quadratic effect of age is significant. Its estimated marginal effect is a concave quadratic function descending from 18 years and reaching a minimum at 69.6 years. The conviction density is a very powerful predictor in general recidivism. A 1-unit increase in this predictor relates to a 57% increase in the odds of reconviction. The odds ratios of offence type show that being a perpetrator of a violent

Table 6. Regression coefficients of the general recidivism logistic regression model ($N = 159\,298$)

Predictor	*Coefficient*	*Standard error*	*Odds ratio*
Constant	−0.147	0.026	0.863
Gender: male (0)/female (1)	−0.386	0.018	0.679
Age (years)	−0.060	0.001	0.942
Age (years^2)	0.0004	0.00004	1.0004
Age at first conviction	0.024	0.002	1.024
Conviction density	0.453	0.045	1.573
Most serious offence			
Violence (reference)	0	1	1.00
Sexual	−0.525	0.078	0.591
Property crime with violence	0.270	0.053	1.310
Property crime without violence	0.072	0.020	1.075
Public order	−0.005	0.023	0.995
Drug offence	−0.194	0.027	0.823
Motoring offence	−0.029	0.019	0.972
Other offence	−0.307	0.023	0.735
Country of birth			
Netherlands (reference)	0	1	1.00
Morocco	0.028	0.034	1.029
Dutch Antilles or Aruba	0.412	0.033	1.510
Surinam	0.290	0.028	1.337
Turkey	0.109	0.034	1.115
Other western countries	−0.263	0.023	0.769
Other non-western countries	−0.049	0.024	0.952
$\log_e$(number of previous convictions)	0.968	0.027	2.633
Dummy for 11–20 previous convictions	2.580	0.065	13.197
Dummy for 21 or more previous convictions	3.250	0.085	25.790

property offence (for instance, armed robbery) has the largest effect of all types of offence. Property offences and public order offences have the next highest odds ratios. The number of previous offences, parameterized as a log-term for 0–10 convictions and two dummy variables for the upper categories, show by far the largest effect on the probability of recidivism.

Age of first conviction has an effect that is opposite to what would be expected. This can be explained by the implicit effect of this variable in the calculation of the conviction density. If the career length is long, the age at first conviction is very likely to be early. Perpetrators who were born in Dutch Antilles or Aruba and Surinam have respectively 51% and 34% higher odds of recidivism than perpetrators born in the Netherlands. This might be partially explained by the fact that for immigrants we are lacking the part of the judicial history recorded in their country of birth.

3.5.2. Violent recidivism: logistic regression

Just as for general recidivism a logistic regression model was chosen. Table 7 shows the coefficients for this model. As in the general recidivism model, the largest effects can be seen in the number of previous convictions. Another large effect is discernible in the exclusive offence type. If the most severe offence in the case was property crime with violence, the odds of violent recidivism are 176% larger, keeping all other variables constant. The type of offence in the index case shows that, if the case also includes a property crime without violence, public order, a motoring or other offence, the risk of violent recidivism is heightened. The country of birth shows that being born in some countries yields a large rise in the odds of violence. Perpetrators

Table 7. Regression coefficients of the violent recidivism logistic regression model ($N = 25041$)

Predictor	*Coefficient*	*Standard error*	*Odds ratio*
Constant	−1.128	0.074	0.324
Gender: male (0)/female (1)	−0.511	0.070	0.600
Age (years)	−0.039	0.004	0.961
Age (years^2)	0.0001	0.0001	1.0001
Age at first conviction	0.008	0.005	1.008
Conviction density	0.258	0.131	1.294
Most serious offence†			
Violence (reference)	0	1	1.00
Sexual	0.550	0.806	1.733
Property crime with violence	1.015	0.476	2.759
Property crime without violence	0.189	0.162	1.208
Public order	−0.001	0.094	0.999
Other offence	0.056	0.211	1.057
Offence type present in index conviction			
Sexual	−0.466	0.777	0.627
Property crime with violence	−0.632	0.459	0.532
Property crime without violence	0.043	0.144	1.044
Public order	0.147	0.049	1.158
Drug offence	−0.223	0.179	0.800
Motoring offence	0.176	0.118	1.193
Other offence	0.116	0.062	1.122
Country of birth			
Netherlands (reference)	0	1	1.00
Morocco	−0.153	0.080	0.858
Dutch Antilles or Aruba	0.268	0.079	1.307
Surinam	0.255	0.071	1.291
Turkey	−0.140	0.093	0.869
Other western countries	−0.095	0.085	0.910
Other non-western countries	0.159	0.065	1.172
Number of previous offences by type			
Violence	0.148	0.012	1.159
Sexual	0.012	0.060	1.012
Property crime with violence	0.048	0.023	1.049
Property crime without violence	−0.002	0.004	0.998
Public order	0.013	0.009	1.013
Drug offence	−0.012	0.021	0.988
Motoring offence	0.003	0.009	1.003
Other offence	0.018	0.017	1.019
$\log_e$(number of previous convictions)	0.528	0.078	1.695
Dummy for 11–20 previous convictions	1.154	0.200	3.171
Dummy for 21 or more previous convictions	1.197	0.271	3.310
Number of previous disposals by type			
Previous prison terms	−0.006	0.011	0.994
Previous community service orders	0.042	0.020	1.043
Previous fines	0.042	0.013	1.043
Previous Public Prosecutor's disposals	−0.035	0.019	0.966

†The parameter for motoring offence was not identified. Therefore, its category was joined with 'other'.

who were born in Dutch Antilles or Aruba or Surinam have approximately 30% higher odds of recidivism, regardless of the other characteristics.

The number of previous offences by type does not seem to predict much. Only the number of previous violence offences has a noticeable effect on violent recidivism. The number of previous disposals also has no large effect on this type of recidivism. Many predictors do not seem to have additional predictive power for these data. Many do not reach statistical significance.

Considering the ease of manual scoring of the risk scale that is used, these predictors could safely be removed from the model.

3.5.3 Sexual recidivism: linear discriminant analysis

Instead of a logistic regression, LDA performed notably better in the sexual recidivism data. The coefficients of the LDA are depicted in Table 8. There are three coefficients that immediately stand out for this model. The number of previous sexual offences and being born in Dutch Antilles or Aruba give the largest positive effect on the probability of sexual recidivism, whereas the number of previous Public Prosecutor's disposals has the largest negative effect. An interesting effect appears for the type of offence. If the sexual offence in a case is accompanied with a property theft with violence offence, the probability of sexually recidivism is larger. Another strong predictor seems to be the country of birth. People born outside the Netherlands

Table 8. Coefficients of the LDA model ($N = 1332$)

Predictor	*Coefficient of discriminant function*
Age (years)	−0.020
Age (years2)	−0.00004
Age at first conviction	0.023
Conviction density	0.366
Offence type present in index conviction	
Violence	−0.040
Property crime with violence	0.949
Property crime without violence	−0.178
Public order	0.025
Drug offence	−0.165
Motoring offence	−0.038
Other offence	−0.135
Number of previous offences by type	
Violence	−0.040
Sexual	0.949
Property crime with violence	−0.178
Property crime without violence	0.025
Public order	−0.165
Drug offence	−0.038
Motoring offence	−0.135
Other offence	−0.146
Country of birth	
Netherlands	−0.011
Morocco	0.415
Dutch Antilles or Aruba	1.645
Surinam	0.555
Turkey	0.645
Other western countries	0.448
Other non-western countries (reference category)	
$\log_e$(number of previous convictions)	0.097
Dummy for 11–20 previous convictions	0.681
Dummy for 21 or more previous convictions)	1.076
Number of previous disposals by type	
Previous prison terms	−0.023
Previous community service orders	0.022
Previous fines	0.250
Previous Public Prosecutor's disposals	−0.161

tend to have a higher risk of sexual recidivism. The number of previous convictions also has a large influence on the risk of sexual recidivism. Some disposals seem to be predictive of sexual recidivism: the number of previous fines is positively related to sexual recidivism whereas the number of Pulic Prosecutor's disposals is negatively related. We cannot explain what underlying factors cause this correlation.

4. Discussion

In this study we attempted to find the best predicting model for three types of recidivism data. These were a year's population for determining general recidivism, having a large N, a small number of parameters and a high base rate, a year's population of violent offenders for determining violent recidivism, having a large N, a large number of parameters and a medium base rate, and 2-year sexual recidivism data for a population of sexual offenders having a small N, a large number of parameters and a low base rate. In the first data set logistic regression analysis performs best overall, the second is outperformed marginally by calibrated logistic regression and LDA has the best performance in the last set. The differences in terms of performance between the best and the follow-up models are generally very small, however. Sexual recidivism was predicted best by the LDA and PLS models. Here logistic regression performed worse.

The conclusion is that using selected modern statistical, data mining and machine learning models provides no real advantage over logistic regression and LDA. If variables are suitably transformed and included in the model, there seems to be no additional predictive performance by searching for intricate interactions and/or non-linear relationships.

Regardless of the complexity of the final model, the formulation of the model can always be translated into a set of equations in an Excel spreadsheet so it can readily be used by a probation worker.

Our data suggest that sample size or low base rate has an effect on the accuracy that is achieved by different statistical models, whereas with a larger sample size or higher base rates no differences in accuracy between statistical models are found. To investigate this on a much larger number of studies, we performed an exploratory reanalysis of the meta-analysis data of Jamain *et al.* (2008) (see also Jamain *et al.* (2008) for an elaborate discussion of the validity of these data such as problems in defining metadata and publication bias). We performed a linear regression on the empirical error rate of the nine most prevalent methods from the studies with a two-class outcome. The predictors were the training and test sample size, the number of predictors, base rate and method. Interactions of method by base rate, training N and test N were included. The method proved to have at best only a very small unique effect, whereas the interactions with base rate, training and test data size were never significant. Therefore, the specific results in the sexual recidivism data seem to be caused by other characteristics of the data.

Caruana and Niculescu-Mizil (2004) proposed the squared error, accuracy and ROC area measure SAR as the simple average of RMSE, ACC and AUC. They advocated it as a summary measure when these three performance criteria lead to different conclusions. However, this has the obvious drawback that differences between these measures that exist are ignored, so it may be better to use SAR only when these measures point in the same direction. The results of this study may be limited because of the following points.

(a) There is an infinite set of potential transformations of the original variables that could potentially improve the performance. This could be so in the strictly linear models like logistic regression and LDA. This could give intrinsically non-linear methods an unfair

advantage. Given the negligible difference between the non-linear and linear models in terms of performance in these data, this is unlikely in this study, however.

(b) For many models that need tuning, there is an infinite set of tuning parameters that could hold subsets that have 'the' optimal performance for those classes of models. We could have missed the optimal tuning parameters in our tuning grids.

(c) Result may not generalize to more specific subsamples of the population that is used. This can be so when important predictors of recidivism are omitted from the model that are not highly correlated with the predictors included. Earlier results, however (Wartna *et al.*; Bogaerts, 2009), showed that the added value of predictors like addiction and work situation have only a limited additional effect on the prediction. It is, however, still possible that the relationship between predictors and the outcome can be different for different subpopulations.

In the special case where we consider sensitivity as important as specificity, the maximum obtainable accuracy is 70%. This is too low to rely solely on these models for decision making about individuals. We can use the predicted score for group-based predictions (see Wartna *et al.*, 2009) but, when individual predictions are concerned, additional information concerning dynamic factors is required.

References

Breiman, L. (1998) Combining predictors. *Technical Report*. Department of Statistics, University of California, Berkeley.

Breiman, L., Friedman, J. H., Olsen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. New York: Chapman and Hall.

Breiman, L. and Schapire, E. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Caruana, R. and Niculescu-Mizil, A. (2004) *Data Mining in Metric Space: an Empirical Analysis of Supervised Learning Performance Criteria*, pp. 69–78. New York: Association for Computing Machinery Press.

Caruana, R. and Niculescu-Mizil, A. (2006) An empirical comparison of supervised learning algorithms. In *Proc. 23rd Int. Conf. Machine Learning, Pittsburgh* (eds W. Cohen and A. Moore), pp. 161–168. New York: Association for Computing Machinery Press.

Caulkins, J., Cohen, J., Gorr, W. and Wei, J. (1996) Predicting criminal recidivism: a comparison of neural network models with statistical methods. *J. Crim. Just.*, **24**, 227–240.

Clemençon, S. and Vayatis, C. (2010) Tree-based ranking methods. *IEEE Trans Inform. Theor.*, **55**, 4316–4336.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Measmnt*, **20**, 31–46.

Copas, J. and Marshall, P. (1998) The offender group reconviction scale: a statistical reconviction score for use by probation officers. *Appl. Statist.*, **47**, 159–171.

Cortes, C. and Mohri, M. (2003) AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems* (eds L. S. B. S.Thrun and B. Schölkopf), pp. 313–320. Cambridge: MIT Press.

Cortes, C. and Vapnik, V. (1995) Support vector networks. *Mach. Learn.*, **20**, 1–25.

Dolan, M. and Doyle, M. (2000) Violence risk prediction: clinical and actuarial measures and the role of psychopathy checklist. *Br. J. Psychiatr.*, **177**, 303–311.

Fisher, R. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Fix, E. and Hodges, Jr, J. (1951) Discriminatory analysis: non-parametric discrimination: consistency properties. *Technical Report*. School of Aviation Medicine, Randolph Field.

Freund, Y. and Schapire, R. E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. 2nd Eur. Conf. Computational Learning Theory, London*, pp. 23–37. New York: Springer.

Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, **28**, 337–374.

Hanley, J. and McNeil, B. (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**, 29–36.

Hanson, R. and Thornton, D. (1999) *Static 99: Improving Actuarial Risk Assessments for Sex Offenders*. Ottawa: Department of the Solicitor General and Her Majesty's Prison Service.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, vol. 2. New York: Springer.

Hosmer, D. and Lemeshow, S. (2000) *Applied Logistic Regression*. New York: Wiley.

Howard, P., Francis, B., Soothill, K. and Humphreys, L. (2009) OGRS 3: the revised offender group recon-

viction scale. *Technical Report*. Ministry of Justice, London. (Available from `http://www.justice.gov.uk/downloadsoasys-research-summary-07-09.pdf`.)
Jamain, A. and Hand, D. (2008) Mining supervised classification performance studies: a meta-analytic investigation. *J. Class.*, **25**, 87–112.
King, R., Feng, C. and Sutherland, A. (1995) Statlog: comparison of classification algorithms on large real-world problems. *Appl. Artif. Intell.*, **9**, 259–287.
Kuhn, M. (2010) caret: classification and regression training. *R Package Version 4.62*.
Lim, T.-S., Loh, W.-Y. and Shi, Y.-S. (1998) An empirical comparison of decision trees and other classification methods. *Technical Report*. University of Wisconsin, Madison.
Lim, T.-S., Loh, W.-Y. and Shi, Y.-S. (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn.*, **40**, 203–229.
Ling, C., Huang, J. and Zhang, H. (2003) AUC: a statistically consistent and more discriminating measure than accuracy. In *Proc. 18th Int. Jt Conf. Artificial Intelligence*, *Menlo Park* (eds G. Gottlob and T. Walsh), pp. 329–341. San Francisco: Morgan Kaufmann.
Maden, A., Rogers, P., Watt, G., Amos, T., Gournay, P. and Skapinakis, P. (2005) *Assessing the Utility of Offenders Group Reconviction Scale-2 in Predicting the Risk of Reconviction within 2 and 4 Years of Discharge from English and Welsh Medium Secure Units (MRD 12/58)*. London: Academic Unit of Psychiatry.
McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Mossman, D. (1994) Assessing predictions of violence: being accurate about accuracy. *J. Consultng Clin. Psychol.*, **6**, 783–792.
Platt, J. (2000) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifers*. Cambridge: MIT Press.
Provost, F. and Fawcett, T. (2001) Robust classification for imprecise environments. *Mach. Learn.*, **42**, 203–231.
R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
de Ruiter, C. and van Dorsellaar, S. (2010) De quick-scan reclassering: betrouwbaarheid en bruikbaarheid verslag van een pilot-onderzoek (The probation service quick scan: reliablity and utility). *Technical Report*. Trimbos Instituut, Utrecht.
de Ruiter, C. and de Jong, E. (2006) *Handleiding QuickScan Reclassering Nederland*. Trimbos Instituut, Utrecht.
Shapiro, D. (1999) The interpretation of diagnostic tests. *Statist. Meth. Med. Res.*, **8**, 113–134.
Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in r. *Bioinformatics*, **21**, 3940–3941.
Taylor, R. (1999) Predicting reconvictions for sexual and violent offences using the revised offender group reconviction scale. *Research Finding 104*. Research, Development and Statistics Directorate, London.
Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C. and Erikson, M. (2003) Distinguishing and combining risks for sexual and violent recidivism. In *Understanding and Managing Sexually Coercive Behavior* (eds R. Prentky, E. Janus and M. Seto), pp. 225–235. New York: New York Academy of Sciences.
Vergouwe, Y. (2003) Validation of clinical prediction models: theory and applications in testicular germ cell cancer. *PhD Thesis*. Erasmus University, Rotterdam.
Wartna, B., Tollenaar, N. and Bogaerts, S. (2009) Statrec: inschatting van het recidivegevaar van verdachten van een misdrijf. *Tijdschft. Crim.*, **51**, 211–227.
Webster, C. D., Douglas, K., Eaves, D. and Hart, D. (1997) HCR-20: assessing risk for violence. Mental Health, Law and Policy Institute, Simon Fraser University, Vancouver.
Wold, H. (1985) Partial least squares. In *Encyclopedia of Statistical Sciences* (eds S. Kotz, N. L. Johnson and C. B. Read), pp. 581–591. New York: Wiley.
Yang, M., Liu, Y. and Coid, J. (2010) *Applying Neural Networks and Other Statistical Models to the Classification of Serious Offenders and the Prediction of Recidivism*, vol. 6/10. London: Ministry of Justice.
Yang, M., Liu, Y. and Coid, J. (2011) A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *J. Quant. Crim.*, **27**, 547–573.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supporting information to: Which method predicts recidivism best?'.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.