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

Journal:	*Journal of the Royal Statistical Society*
Manuscript ID:	JRSS-OA-SA-Jan-11-0010.R1
Manuscript Type:	Original Article - Series A
Date Submitted by the Author:	14-Dec-2011
Complete List of Authors:	Tollenaar, Nikolaj; Ministry of Justice, WODC van der Heijden, Peter; Utrecht University,
Keywords:	recidivism, predictive performance, prediction, logistic regression, data mining, machine learning

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

Abstract. Using criminal population criminal conviction history information, 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 different techniques differs only slightly for general and violent recidivism, while differences are larger for sexual recidivism. For the general and violent recidivism data we present the results of logistic regression and for sexual recdivisim of linear discriminant analysis.

Keywords: recidivism; prediction; predictive performance; logistic regression; linear discriminant analysis; machine learning; data mining

1. Introduction

Risk assessment is a widespread phenomenon in forensic contexts. A plethora of violence, sexual and general recidivism risk instruments has 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 only change in 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 OGRS (Offender Group Reconviction Score), was developed in the United Kingdom: t. Probation staff and corrections researchers have been using it since the 1990's. The scale was required to be quick and 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 (two-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, Rogers, Watt, Amos, Gournay, and Skapinakis, 2005; Howard, Francis, Soothill, and Humphreys, 2009).
In the Netherlands, a similar instrument was developed to quickly assess recidivism risk and influencability of adult (e.g. 18 years or older) offenders, 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 that is available for probation officers. The Quickscan needed to consist of

a dynamic part to have a brief view of criminogenic needs and a static part to have an accurate estimate of recidivism. The latter led to the development of scale one of the quick scan that consisted of the assessment of the static recidivism risk, called the StatRec. This scale is similar to the original OGRS, but includes an additional risk factor, namely country of birth, and targets the four-year instead of the two-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 make a fist step towards such an instrument for the Netherlands. Taylor (1999) developed a revised OGRS that also predicts violent and sexual recidivism using static information. Two other important sexual recidivism scales, the popular static 99 (Hanson and Thornton, 1999) and its successor static 2002 als contain mainly static information. The risk matrix 2000 (Thornton et al. 2003) is another actuarial tool used by England and Wales' probation. This instrument has separate scales for violent and sexual recidivism risc. Besides static factors it also includes victim information.
Traditionally, the most suitable method for estimating and predicting the probability of an event at a point in time is standard linear logistic regression (Hosmer and Lemeshow, 2000). This model has been used both for the OGRS and the StatRec. It is a statistical model that is estimated via maximum likelihood, which models the logit of the probability to recidivate linearly. Since the sixties however, many different approaches have been developed in the field of machine learning and data mining for predicting a binary choice, that can be expressed in terms of 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 maximising the likelihood of a probability model, these methods typically tend to optimise 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 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 (Lim, Loh & Shih, 1998, 2000). Jamain & 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 LDA to be a good overall classifier as well as C4.5, a tree classifier. Linear discriminant analysis and logistic regression are known to give very similar results, even if LDA is used inappropriately, for instance when modeling non-normal data, or when the predictors are of a categorical measurement level (Hastie, Tibshirani and Friedman, 2009, p.128). Generally, the results of the different studies indicate that there is no 'best model' for all datasets and situations, and that with a specific dataset it is necessary to establish which model is best for the relevant data (Jamain et al, 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, Hastie & Tibshirani, 2000), random forests (Breiman, 2001) and support vector machines (SVM's, Cortes & 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 the HCR-20 to predict reconvictions for violence. They compared only a limited number of approaches on a binary outcome (logistic regression, CART 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 four years.

In machine learning and data mining, the performance is mainly measured by the classification error or its complement accuracy, whereas the performance of prediction scales is routinely evaluated with ROC analysis (see Mossman, 1994). The latter is a general nonparametric method of establishing the ability to discriminate between two classes 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 positives against the amount of false negatives, called the ROC curve. The higher the area under the curve (AUC) the 'better' the model. During the past ten years, the importance of this measure of discrimination has been discovered by the machine learning community as a means of assessing predictive validity (Provost & Fawcett, 2001; Ling, Huang & Zhang, 2003). This led to the development of algorithms optimising the AUC directly (Cortes & Mohri, 2003; Clemençon & Vayatis, 2010).

In this paper we want to evaluate a broad set of prediction models on a broad set of performance metrics. We use more metrics so we will not only choose the best ranking or highest accuracy model. In this paper we want to find out if using data mining, machine learning and modern statistical models leads to a relevant 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 most 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 a 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.

2. Method

2.1. Data source

The StatRec scale uses only static information that is available in the Dutch Offender's Index. This index is an automated, encrypted and anonymised version of the judicial information that probation officers and others can request for an offender, the Judicial Documentation System (JDS). 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 persons from ages 12 and up, as the minimal age of criminal responsability 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 persons of 18 years or older. This means we restrict our scales to persons falling under adult law, with a custodial sentence, a non-custodial sentence or a disposalfrom 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 (PPD's) are usually for less serious crimes and supposedly similar to England and Wales' police cautions.

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, estimated date of release was subtracted from the recidivism duration. General recidivism includes motoring offences like leaving the scene of accident and driving under influence of alcohol or drugs. Cantonal court cases (i.e. misdemeanors) are not counted. Violent recidivism includes threatening, extortion, property using violence and/or threatening, assault and battery, murder, homicide, culpable homicide but excludes destruction of property. Sexual recidivism includes rape, violation, indecent assault (with minors, subordinates), sexual contact below age 16, age 12 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 would be needed, a 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.

2.3. Data selection

StatRec predicts the four 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 dataset consists of all adult perpetrators 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 this dataset (N=159,298): 10,000 for estimation of the models (training set), and 10,000 for validation of models (test set). For sexual recidivism, because the prevalence of sexual recidivism is relatively low, we include all sexual offenders of 2005 and only the sexual offenders of 2006 who have at least four years of follow up time, in order to obtain a larger data set. (N=1,392). Thirty percent of the sexual recidivism data is kept as a final test set. The recidivism is defined as the time from administration of the penal case to the first offence date in a recidivism case. As detention data is not yet well available, the follow-up time is corrected for estimated time in prison, based on the duration of the prison term. Cases that had less than four years follow-up time 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 datasets. The model of general recidivism needs to be very simple and fast to score. Therefore, small selection of variables should be included in the model. On the other hand, we allow the models for violent and sexual recidivism to be more exhaustive with respect to the information available in the judicial database because prediction these types of recidivism is inherently more difficult. Table 1 contains a list of all the variables used in the three submodels and their sample statistics.

To account for non-linearity the number of previous convictions will be included in the models transformed by a natural log. The possible nonlinear effect of age will be dealt with by including also the squared age, after centering 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 estimated

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

	General recidivism	Violent recidivism	Sexual recidivism
	(N=20,000)	(N=20,000)	(N=1,374)
4 year base rate (%)	37.7	22.4	5.2
Gender: female(%)	15.9	9.7	-
Age in years (mean)	35.2	34.7	38.3
Age of first conviction (mean)	27.1	24.8	29.5
Most serious offence type (%)			-
Violence	13.2	90,7	
Sexual	0.5	0,5	
Property with violence	1.5	1,2	
Property without violence	22.1	3,4	
Public order	10.7	3,2	
Drug offence	6.8	0,5	
Motoring offence	30.0	0,0	
Misc. offence	15.2	0,6	
Country of birth (%)			
Netherlands	70.3	71.2	69.7
Morocco	3.4	4.1	2.9
Neth. Antilles/Aruba	3.1	3.8	5.0
Surinam	4.6	5.2	5.3
Turkey	3.0	3.8	2.8
Other Western countries	8.4	4.9	6.1
Other non-Western countries	7.3	7.1	8.2
Offence type present in index case (%)			
Violence component (0/1)	-	-	12.7
Sexual component	-	0.6	-
Property with violence	-	1.4	2.0
Property without violence	-	4.5	3.4
Public order	-	13.9	5.0
Drug offence	-	1.0	0.9
Motoring offence	-	1.5	0.7
Misc. offence	-	6.9	9.4
Criminal history counts (mean)			
Reconviction density†	0.78	0.50	0.40
Number of previous convictions‡	5.0	6.06	4.31
previous violence offences	-	0.96	0.54
previous sexual offences	-	0.04	0.25
previous property with violence offences	-	0.22	0.54
previous property offences	-	2.49	1.68
previous public order offences	-	1.01	0.66
previous drug offences	-	0.19	0.11
previous motoring offences	-	0.78	0.60
previous misc. offences	-	0.07	0.26
previous prison terms	-	1.01	0.70
previous community service orders	-	0.43	0.29
previous fines	-	1.19	0.89
previous ppd's§	-	0.45	0.32

†This variable is computed as $\sqrt{\#convictions/(careerlength+1)}$.
‡To aid fast scoring on the part of the probation officer, in the models the number of previous convictions is partially categorised. In the actual models, the log of 0 to 9 (+1) convictions are modeled linearly, and the categories of 11-20 and 21+ 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.
§ Public prosecutor's disposals.

by the models might differ over time posing a potential threat to temporal validity. Because the models will be updated regularly we doubt 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 has large N, small number of parameters and a high base rate, the violent recidivism data has a large N, a large number of parameters and a medium base rate, whereas the sexual recidivism has 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 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 generated probabilities correspond well with 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 is able to 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 the AUC, the area under the ROC-curve (Hanley & 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 .5 and the base proportion of the outcome in the data, also known as the 'base rate' in criminological literature. Some indicators of clinical usefulness include prediction accuracy (i.e. the percentage correctly classified), sensitivity (the percentage of observations with the positive outcome correctly classified) and 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). These authors proposed to combine three different indicators, namely the AUC, accuracy and root mean squared error (RMSE), into one performance measure, the SAR (Squared error, Accuracy, and ROC area), to establish optimality in different domains and create a more 'robust' measure. They found that the SAR correlated highest with a range of nine performance metrics.
For the purpose of this study, we have computed a range of performance criteria in order to be able to provide a detailed picture of which method should be preferred in a range of different situations. These criteria are:

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

- 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 positives and the true negatives 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 will only compare accuracies within datasets. The accuracy is the complement of the error rate often

used in classification studies;

- 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 the following formula:

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

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

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

$$SAR = \frac{AUC + ACC + (1 - RMSE)}{3};$$

- 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 with 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;

- Local calibration error (Calibration error around the cut-off, CAL(.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 only uses the domain around the threshold. This measure is calculated in the following way:

  (a) Ten percent of the data is selected around the chosen cut-off value;
  (b) 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;
  (c) 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. The SAR will be the decisive criterion on which we will base the selection of the final model, because it takes into account all three dimensions of predictive validity, namely discrimination, calibration and accuracy. The ACC and thus the 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 .5, it is more likely to have a positive outcome than a negative outcome.

2.6. Models

The following eleven models are compared with respect to their predictive performance:

(a) Linear logistic regression (see Hosmer & Lemeshow, 2000): logistic regression models the logit of a binomial probability linearly via maximum likelihood. It stems from the 1930's and is a special case of the family of generalised linear models (McCullagh and Nelder, 1989);

(b) Multivariate adaptive regression splines (MARS, Friedman, 1991): MARS 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 interaction degree need to be specified in advance. It can be fitted on the entire class of generalised linear models (McCullagh and Nelder, 1989). In our case the logit link function and the binomial family are used;

(c) Linear discriminant analysis (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 different 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, Tibshirani and Buja, 1993): this is essentially a nonparametric LDA. It uses MARS (see above) for estimating a non-linear decision boundary;

(e) Recursive partitioning (rpart): a classification tree is a method to recursively partition the covariate space into maximally separated groups with respect to the dependent variable (the CART algorithm by Breiman, Friedman, Olshen, and Stone, 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 $1 - \sum_j p^2$, 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 & 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 CART 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 shown to be an excellent way of obtaining precise and discriminative probabilities without having to explicitly specify a model and is often called the 'best off the shelf classifier' (Breiman, 1998);

(g) Logitboost (Friedman, Hastie & Tibshirani, 2000): this algorithm optimises the logistic loss function so it behaves like a generalized additive model. The base model used here is a decision stump (only the first split of a regression tree);

(h) Neural network with one single hidden layer (nnet): 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 e.g. Hastie et al., 2009, p. 398);

(i) Linear support vector machines (Linear SVM, Cortes & 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 maximised and misclassified examples are allowed/disallowed by varying the cost parameter C. The penalty of these errors is equal to the distance to the decision boundary times C. SVM's 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): K-nearest neighbours is a rather simple method of classification (Fix and Hodges, 1951): this algorithm finds a group of k observations closest to the test observation in terms of Euclidian distance. The proportion of the votes for the positive class is then returned as a probability;

(k) Partial least squares (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 optimise the covariance between the scores and the response variable. It has less strict model assumptions than for instance linear discriminant analysis.

Some of the aforementioned methods are not designed to generate an estimated probability as an outcome, like support vector machines, which instead give the distance to the decision boundary. It is, however, possible to transform these distances using a nonlinear 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 dataset with respect to the actual outcome on the estimation dataset. 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 on request.

2.7. Model selection within method

Many of the algorithms and models 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 10-fold cross validation using accuracy as the main criterion. We chose accuracy because many of the models are built to optimise accuracy. The final selected model is tested on the test set of the data, using the SAR as the final criterion. We will 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 datasets:

- Recursive partitioning tree: maximum split depth of the tree in 2, 3, 4, 6, 10, 13, 15, 16, 28, threshold of 0.001;
- Single layer feed forward neural networks: 1, 2, 4, 8, 16 hidden nodes and decay of 0, 0.01, 0.001, 0.0001, 0.00001. All combinations of the resulting grid are tried;
- MARS/FDA: maximum number of terms in the model of 15, 20, 25, 30, 35, 40, 45, maximum interaction degree: 1, 2, 3. All combinations of the resulting grid are evaluated;
- Adaptive boosting (adaBoost): 50, 150, 200, 250, 300, 350, 400, 450, 500 boosting iterations, tree maximum depth of 30, step size of 0.1.;
- LogitBoost: 50, 150, 200, 250, 300, 350, 400, 450, 500 boosting iterations;

- Linear support vector machine: a cost parameter of 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 2, 10, 100 and class weights of (2,3), (1,3) and (1,20) for general, violent and sexual recidivism respectively;
- K nearest neighbours classification: k=5, 7, 9, 11, 13, 15, 17, 19, 21 and 23;
- Partial least squares: number of latent components=1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Logistic regression and linear discriminant analysis 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 using the ROCR package version 1.0-2 (Sing, Sander, Beerenwinkel & Lengauer, 2005). The actual training of the different models was performed using the caret package (Kuhn, 2010).

3. Results

We will successively discuss the benchmarks for the general recidivism model, the violent recidivism model and the sexual recidivism model and then make an overall judgment on the three domains of the predictive performance.

3.1. *General recidivism*

Table 2 shows the performance indicators of the 11 seven-variable compared models on the test data. For the SVM and PLS models the results are based on the calibrated probabilities. In each column, the score of most optimal model is printed in boldface. We will first describe the performance with respect to the overall SAR(.5) measure, followed by a treatment of the other metrics.

Following the SAR at cut-off .5, the logistic regression model seems to outperform all other models. Its SAR(.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, the 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.

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 .5 is zero for logistic regression, MARS and PLS. The largest error is made by the logitBoost.

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 the SAR at cut-off .5, the calibrated logistic regression model seems to outperform all other models. It is however almost identical to the SAR(.5) of discriminant analysis, adaBoost

Table 2. Test data performance of eleven *general* recidivism models on nine indicators

	Criterion					
	cut-off independent			cut-off = 0.5		
	AUC	RMSE	CALerr	ACC(.5)	SAR(.5)	CAL(.5)
Model						
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 cal.	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 cal.	0.710	0.473	0.036	0.680	0.645	0.000
pls cal.	0.773	0.432	0.036	0.726	0.689	0.000
knn cal.	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 cal.	0.768	0.436	0.054	0.725	0.686	0.011

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

and the calibrated PLS models. These differences can be attributed to only very slight differences in the AUC, ACC(.5) and RMSE. Exceptionally poorly performing models, judging by the SAR(.5), are the rpart, logitBoost, K-nn, and SVM.
The classification accuracy is highest on the LDA but 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 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 kappa (Cohen, 1960) as the model selection criterion in finding the optimal tuning parameters in the sexual recidivism models. The results on the test data of the sexual recidivism models show again the
Judging by the SAR(.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 components of the SAR, we can see that the variation is almost exlusively due to the variations in the 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. The AUC of the logistic regression is a lot smaller than the AUC of LDA and PLS. The relatively low RMSE does not make up for this. MARS, FDA, rpart, and K-nn seem to fail completely. Surprisingly, the overall calibration error is lowest in the logistic regression model model. This model also has the lowest calibration around the .5 cut-off. The 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.

Table 3. Test data performance of eleven *violent* recidivism models on nine indicators

	Criterion					
	cut-off independent			cut-off = 0.5		
	AUC	RMSE	CALerr	ACC(.5)	SAR(.5)	CAL(.5)
Model						
logreg cal.	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 cal.	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 cal.	0.671	0.408	0.035	0.772	0.678	0.025
pls cal.	0.741	0.396	0.038	0.777	0.708	0.020
knn	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
SvmLinear	0.726	0.405	0.058	0.770	0.697	0.037

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

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

	Criterion					
	cut-off independent			cut-off = 0.5		
	AUC	RMSE	CALerr	ACC(.5)	SAR(.5)	CAL(.5)
Model						
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 cal.	0.725	0.202	0.021	0.955	0.826	0.094
fda cal.	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 cal.	0.722	0.203	0.018	0.955	0.824	0.271
knn	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 cal.	0.711	0.199	0.022	0.958	0.823	0.234

Note: 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. Summary of the results

The assertion of many authors that there is no best model for all data holds true with respect to these three datasets, although some do well on all models. Following the dimensions of prognostic models from Vergouwe, we will 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. Its discrimination is less well with respect to sexual recidivism. As others have found, overall the LDA models are good discriminators on all three datasets. Another good overall discriminator is the partial least squares model.
The single tree recursive partitioning method should not be used if discrimination is important. It is most suitable for optimising classification accuracy.

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 cutoff. The logitBoost and 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 by 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, LDA) and MARS seem to do well across datasets on this criterion. The 'classify all in the largest class' scheme showed to obscure 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 cut-off point choice and calibration of the probabilities. The comparability of clinical usefulness can be assessed by taking the following point of departure. If we consider false positives and false negatives equally costly we will want to make an equal percentage of errors in both types of predicted events. In this case the amount of false positives is as large as the amount of false negatives. 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. The table shows again different models are optimal on different datasets. 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 PLS, and MARS.
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 would be the only goal, PLS seems to be the best. The LDA model is the second best on classification accuracy. The logitBoost model is again the worst performing model, now achieving an accuracy worse than flipping a coin. The adaBoost model, relying on the rpart model, also suffers from poor performance in these data.

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

	General recidivism	Violent recidivism	Sexual recidivism
Model			
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
linear SVM	0.699	0.671	0.602

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 of comparison of the models was to choose for logistic regression or LDA when they did not perform worse than their modern counterparts. This turned out to be the case for all data sets. In subsequently making a choice for either logistic regression or LDA, we choose for 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 ultimately chosen were fitted on the complete data-sets. We will describe these definitive models here.

3.5.1. General recidivism: logistic regression

The odds ratios of the seven predictor model based on 159,298 observations are provided in table 6. Although the N is very large, some of the categories are nonsignificant. The table shows that the conditional odds of reconviction are 32 percent less for women than for men. Recidivism tends to decline with age; the odds lower three percent for each extra year. the quadratic effect of age is small but nevertheless significant. The reconviction density is a very powerful predictor in general recidivism. A one-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 property offence (for instance, armed robbery) has the largest effect of all offence types. Property offences and public order offences have the next highest odds ratios. The number of previous offences, parameterised as a log term for 0 to 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 opposite to what would be expected. This can be explained by the implicit effect of this variable in the calculation of the reconviction density. If the career length is long, the age at first conviction is early. Perpetrators born in Netherlands Antilles/Aruba and Surinam have respectively a 51% and 34% higher odds of recidivating than perpetrators born in

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

	Coefficient	Std. error	Odds ratio
Predictor			
Constant	-0.15	0.03	0.86
Gender: male(0)/female(1)	-0.39	0.02	0.68
Age in years	-0.06	0.00	0.94
Age in years squared	0.00	0.00	1.00
Age at first conviction	0.02	0.00	1.02
Reconviction density	0.45	0.05	1.57
Offence type (exclusive)			
Violence (reference)	0		1.00
Sexual	-0.53	0.08	0.59
Property with violence	0.27	0.05	1.31
Property without violence	0.07	0.02	1.07
Public order	-0.01	0.02	0.99
Drug offence	-0.19	0.03	0.82
Motoring offence	-0.03	0.02	0.97
Other offence	-0.31	0.02	0.74
Country of birth			
Netherlands (reference)	0		1.00
Morocco	0.03	0.03	1.03
Neth. Antilles/Aruba	0.41	0.03	1.51
Surinam	0.29	0.03	1.34
Turkey	0.11	0.03	1.12
Other Western countries	-0.26	0.02	0.77
Other non-Western countries	-0.05	0.02	0.95
Log number of previous convictions	0.97	0.03	2.63
Dummy for 11-20 previous convictions	2.58	0.06	13.20
Dummy for 21 or more previous convictions	3.25	0.09	25.79

the Netherlands. This might be partially explained by the fact that of immigrants we miss part of their judicial history in the country of birth.

3.5.2. Violent recidivism: logistic regression

Just as for general recidivism a logistic regression model was chosen. The coefficients of this model are in table 7. Just 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 with violence, the odds of violent recidivism is 176% larger, keeping all other variables constant. The offence type (any) shows that if the case also includes a property 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 born in the Netherlands Antilles/Aruba or Surinam have a approximately 30% higher odds of recidivating, regardless of the other characteristics.

The number of previous offence types does not seem to predict a lot. 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. A lot do not reach statistical significance. Considering the ease of manual scoring of the risk scale used, these predictors could safely be removed from the model.

Table 7. Regression coefficients of the violent recidivism logistic regression model (N=25,041)

	Coefficient	Std. error	Odds ratio
Predictor			
Constant	-1.13	0.07	0.32
Gender: male(0)/female(1)	-0.51	0.07	0.60
Age in years	-0.04	0.00	0.96
Age in years squared	0.00	0.00	1.00
Age at first conviction	0.01	0.00	1.01
Reconvicion density	0.26	0.13	1.29
Offence type (exclusive)†			
Violence (reference)	0		1.00
Sexual	0.55	0.81	1.73
Property with violence	1.02	0.48	2.76
Property without violence	0.19	0.16	1.21
Public order	0.00	0.09	1.00
Other offence	0.06	0.21	1.06
Offence type (any)			
Sexual	-0.47	0.78	0.63
Property with violence	-0.63	0.46	0.53
Property without violence	0.04	0.14	1.04
Public order	0.15	0.05	1.16
Drug offence	-0.22	0.18	0.80
Motoring offence	0.18	0.12	1.19
Other offence	0.12	0.06	1.12
Country of birth			
Netherlands (reference)	0		1.00
Morocco	-0.15	0.08	0.86
Neth. Antilles/Aruba	0.27	0.08	1.31
Surinam	0.26	0.07	1.29
Turkey	-0.14	0.09	0.87
Other Western countries	-0.09	0.08	0.91
Other non-Western countries	0.16	0.07	1.17
Number of previous offence by type			
Violence	0.15	0.01	1.16
Sexual	0.01	0.06	1.01
Property with violence	0.05	0.02	1.05
Property without violence	0.00	0.00	1.00
Public order	0.01	0.01	1.01
Drug offence	-0.01	0.02	0.99
Motoring offence	0.00	0.01	1.00
Other offence	0.02	0.02	1.02
Log number of previous convictions	0.53	0.08	1.70
Dummy for 11-20 previous convictions	1.15	0.20	3.17
Dummy for 21 or more previous convictions	1.20	0.27	3.31
Number of previous disposals by type			
Previous prison terms	-0.01	0.01	0.99
Previous community service orders	0.04	0.02	1.04
Previous fines	0.04	0.01	1.04
Previous ppd's	-0.04	0.02	0.97

†Motoring offence as a category was not identified and joined with 'other'.

3.5.3. Sexual recidivism: linear discriminant analysis

Instead of a logistic regression, LDA performed notably better in sexual recidivism data. The coefficients of the linear discriminant analyisis are depicted in table 7. There are three coefficients that immediately stand out for this model. The number of previous sexual offences and being born in the Netherlands Antilles/Aruba give the largest positive effect on the probability of sexual recidivism, whereas the number of previous ppd's has the largest negative effect. An interesting effect appears for the offence type. If the sexual offence in a case is accompanied with a property with violence offence, the probability of sexually recidivating is larger. Another strong predictor seems to be the country of birth. People born outside the Netherlands 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 ppd's is negatively related. It is 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 years population for determining general recidivism, having a large N, a small number of parameters and a high base rate; a years population of violent offenders for determining violent recidivism, having a large N, a large number of parameters and a medium base rate; a sexual recidivism data of two year population of sexual offenders having a small N, large number of parameters and a low base rate. In the first dataset logistic regression analysis performs best overall, the second is outperformed marginally by calibrated logistic regression whereas 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 linear discriminant analysis and partial least squares models. Here logistic regression performed worse.

The conclusion is that using selected modern statistical, data mining and machine learning models provide 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 nonlinear relations. Regardless of the complexity of the model chosen, the formulation of the model can always be translated to a set of equation in an excel sheet so it can readily be used by the probation worker.

Our data suggest that sample size or low base rate has an impact on the accuracy 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 re-analysis of the meta-analysis data of Jamain et al (2008, see also this publication for an elaborate discussion of the validity of these data such as problems in defining meta-data and publication bias). We performed a linear regression on the empirical error rate of the nine most prevalent methods from the studies with a 2-class outcome. Predictors are the training and test sample size, the number of predictors, base rate and method and included interactions of method by base rate, training N and test N. Method proved to have at best only a very small unique effect, while 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 measure SAR (Squared error, Accuracy, and ROC area) as the simple average of 1-RMSE, ACC and AUC. They advocate 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

Table 8. coefficients of the LDA model

	coefficients of discriminant function
Age in years	-0.02
Age in years squared	0.00
Age at first conviction	0.02
Reconvicion density	0.37
Offence type (any)	
Violence	-0.04
Property with violence	0.95
Property without violence	-0.18
Public order	0.03
Drug offence	-0.17
Motoring offence	-0.04
Other offence	-0.13
Number of previous offence by type	
Violence	-0.04
Sexual	0.95
Property with violence	-0.18
Property without violence	0.03
Public order	-0.17
Drug offence	-0.04
Motoring offence	-0.13
Other offence	-0.15
Country of birth	
Netherlands	-0.01
Morocco	0.42
Neth. Antilles/Aruba	1.64
Surinam	0.55
Turkey	0.64
Other Western countries	0.45
Other non-Western countries (reference category)	
Log number of previous convictions	0.10
Dummy for 11-20 previous convictions	0.68
Dummy for 21 or more previous convictions	1.08
Number of previous disposals by type	
Previous prison terms	-0.02
Previous community service orders	0.02
Previous fines	0.25
Previous ppd's†	-0.16

† Public prosecutor's disposals.

be better to use the SAR only when these measures point into 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 the case in the strictly linear models like logistic regression and LDA. This could give intrinsically nonlinear methods an unfair advantage. Given the negligible difference between the nonlinear 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 generalise to more specific sub-samples of the population used. This can be the case when important predictors of recidivism are omitted from the model that are not highly correlated with the included predictors. Earlier results however (Wartna, Tollenaar & Bogaerts, 2009) show 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 relation between predictors and the outcome can be different for different sub-populations.

In the special case where we consider sensitivity as important as specificity, the maximum obtainable accuracy is 70 percent. 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 (compare 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, Dept. Statistics, Univ. California, Berkeley.

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

Breiman, L. and E. Schapire (2001). Random forests. In *Machine Learning*, pp. 5–32.

Caruana, R. and A. Niculescu-Mizil (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. pp. 69–78. ACM Press.

Caruana, R. and A. Niculescu-Mizil (2006). an empirical comparison of supervised learning algorithms. In W. Cohen and A. Moore (Eds.), *Proceedings of the 23rd International Conference on Machine learning, Pittsburgh, PA 2006*, IMCL '06, pp. 161–168. ACM.

Caulkins, J., J. Cohen, W. Gorr, and J. Wei (1996). Predicting criminal recidivism: a comparison of neural network models with statistical methods. *Journal of Criminal Justice 24*(3), 227–240.

Clemençon, S. and C. Vayatis (2010). Tree-based ranking methods. *IEEE Transactions on Information Theory 55*(9), 4316–4336.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 31–46.

Copas, J. and P. Marshall (1998). The offender group reconviction scale. *Appl. Stat. 47*, 159–171.

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

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

de Ruiter, C. and E. de Jong (2006). *Handleiding QuickScan Reclassering Nederland*. Trimbos instituut.

de Ruiter, C. and S. van Dorsellaar (2010). De quick-scan reclassering: betrouwbaarheid en bruikbaarheid. verslag van een pilot-onderzoek. [the probation service quick scan: Reliablity and utility]. Technical report, Utrecht.

Dolan, M. and M. Doyle (2000). Violence risk prediction: Clinical and actuarial measures and the role of psychopathy checklist. *British Journal of Psychiatry 177*, 303–311.

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

Fix, E. and J. Hodges Jr (1951). discriminatory analysis: non-parametric discrimination: Consistency properties. Technical report, Randolph Field, TX.

Freund, Y. and R. E. Schapire (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, London, UK, pp. 23–37. Springer-Verlag.

Friedman, J., T. Hastie, and R. Tibshirani (2000, April). Additive logistic regression: A statistical view of boosting. *Ann. Stat. 28*(2), 337–374.

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

Hanson, R. and D. Thornton (1999). *Static 99: Improving actuarial risk assessments for sex offenders.* Ottawa, ON: Department of the Solicitor General and Her Majesty's Prison Service.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: Data mining, inference and prediction.*, Volume second. New York: Springer-Verlag.

Hosmer, D. and S. Lemeshow (2000). *Applied logistic Regression*. New York: John Wiley & Sons Inc.

Howard, P., B. Francis, K. Soothill, and L. Humphreys (2009, March). OGRS 3: the revised offender group reconviction scale. Technical report, http://www.justice.gov.uk/downloadsoasys-research-summary-07-09.pdf.

Jamain, A. and D. Hand (2008, June). Mining supervised classification performance studies: A meta-analytic investigation. *Journal of Classification 25*(1), 87–112.

King, R., C. Feng, and A. Sutherland (1995, June). Statlog:comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence 9*(3), 259–287.

Kuhn, M. (2010). caret: Classification and regression training. R package version 4.62.

Lim, T.-S., W.-Y. Loh, and Y.-S. Shi (1998). An empirical comparison of decision trees and other classification methods. Technical report.

Lim, T.-S., W.-Y. Loh, and Y.-S. Shi (2000). A comparison of prediction accuracy, complexity, and training trime of thirty-three old and new classification algorithms. *Machine Learning 40*(1), 203–229.

Ling, C., J. Huang, and H. Zhang (2003). AUC: a statistically consistent and more discriminating measure than accuracy. In G. Gottlob and T. Walsh (Eds.), *Proceedings of the 18th international joint conference on artificial intelligence (IJCAI 2003)*, Menlo Park, pp. 519–524. AAAI Press.

Maden, A., P. Rogers, G. Watt, T. Amos, P. Gournay, and P. Skapinakis (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 J. Nelder (1989). *Generalized Linear Models*. Boca Raton: Chapman and Hall/CRC.

Mossman, D. (1994). Assessing predictions of violence. being accurate about accuracy. *Journal of Consulting and Clinical Psychology 6(4)*, 783–792.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Large Margin Classifers. Cambridge, MA, MIT Press.

Provost, F. and T. Fawcett (2001, March). Robust classification for imprecise environments. *Machine Learning 42*(3), 203–231.

R Development Core Team (2008). R: A language and environment for statistical computing. ISBN 3-900051-07-0.

Shapiro, D. (1999, June). The interpretation of diagnostic tests. *Statistical Methods in Medical Research 8*(2), 113–134.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005). ROCR: Visualizing classifier performance in r. *Bioinformatics 21*(20), 3940–3941.

Taylor, R. (1999). *Predicting Reconvictions for sexual and violent offences using the revised offender group reconviction scale*. Number 104 in Research findings. Research, Development and Statistics Directorate.

Thornton, D., R. Mann, S. Webster, L. Blud, R. Travers, C. Friendship, and M. Erikson (2003). Distinguishing and combining risks for sexual and violent recidivism. In R. Prentky, E. Janus, and M. Seto (Eds.), *Understanding and managing sexually coercive behavior*, Volume 989, pp. 225–235. New York: Annals of the New York Academy of Sciences.

Vergouwe, Y. (2003). *Validation of Clinical Prediction Models: Theory and Applications in Testicular Germ Cell Cancer*. Ph. D. thesis, Erasmus University.

Wartna, B., N. Tollenaar, and S. Bogaerts (2009). Statrec: inschatting van het recidivegevaar van verdachten van een misdrijf. *Tijdschrift voor Criminologie 51*(3), 211–227.

Wold, H. (1985). Partial least squares. In *Encyclopedia of Statistical Sciences*, pp. 581–591. New York: Wiley.

Yang, M., Y. Liu, and J. Coid (2010, March). *Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism*, Volume 6/10. Ministry of Justice.

Yang, M., Y. Liu, and J. Coid (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. *Journal of Quantitative Criminology Online First, 31 March 2011*.