Predicting reconviction from a dataset of convicted individuals

CFAS420 Statistical Learning I

Harry Baines

35315878

h.baines3@lancaster.ac.uk

Abstract

Extensive research has been undertaken in the area of predicting reconviction. In this paper, we seek to implement and evaluate machine learning models to predict reconviction from a range of criminological variables collected for offenders convicted in 1990. Tree models, notably classification trees, will enable a thorough exploration and interpretation of the covariates which significantly affect reconviction by classifying observations into distinct groups. Logistic regression, a model which will model the binary reconvicted variable (yes/no), will enable a statistical significance analysis of covariates to be undertaken. Finally, a neural network with a single hidden layer will be applied to the data. We conclude by comparing and discussing the predictive performance of these models, and the implications of employing these models in courts. We find the classification tree yields the greatest validation accuracy for the dataset, as well as a high level of interpretability in terms of the covariates which affect reconviction.

1 Introduction and Background

In 1996, a statistical model known as the offender group reconvinction scale (OGRS) was launched by the Home Office, which was based on a logistic regression analysis on a large sample of recently convicted offenders [1]. This model was intended to assist criminal courts in England and Wales by assessing the risk of an individual reoffending, and aid judges and magistrates with their sentencing decisions. It has since been the de facto standard for predicting reconviction for previous offenders. The OGRS model is built around 6 main covariates: age, sex, number of youth custody sentences (under age 21), total number of court appearances (total number of separate occasions when the offender has appeared in court and been found guilty), time in years since first conviction and the type of offence (offence expected to receive most serious sentence taken if multiple are given). The computed OGRS value yields an estimate of the probability of reconviction for a standard list offence (not including minor offences such as traffic violations) within a period of 2 years.

Continued development of this model resulted in OGRS2 and OGRS3, with the latter providing more accurate and valid predictions, using fewer risk factors. The covariates used are primarily static variables such as the offender's age, gender and criminal history [2]. OGRS3 focused on the creation of a parsimonious model, that is, a model which achieves great explanatory power with few explanatory variables and greater generalizability. OGRS4, a further iteration of the OGRS model, was developed to improve prediction of reconviction using temporal covariates such as the number of children and housing status of the offender [3].

The dataset used in this paper contains actuarial and static variables without time varying variables. Hence, the predictive performance of the models implemented in this paper will not surpass that of OGRS4.

2 Data Manipulation and Exploratory Data Analysis

The dataset we utilise in this paper contains observations of individuals who have been convicted of an offence in 1990. The conviction date is termed the *target date*, with the main convicted offence termed the *target offence*. Each offender possesses a unique identification number and data summarising their criminal history up to the target date. A single binary variable is also included which indicates whether the individual was reconvicted for any offence up to the end of 1999.

The dataset contains 3449 offenders with 21 variables describing each individual. 16 of these variables are categorical with the remaining 5 columns being numerical in nature. The dataset contained no missing values.

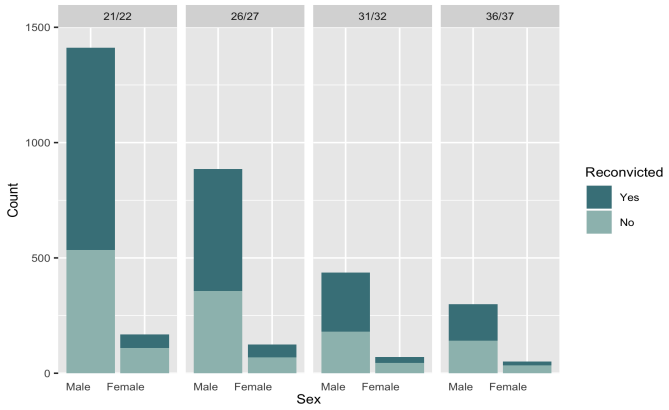


Figure 1: Bar plot of offenders reconvicted for each sex across age groups.

Figure 1 demonstrates a decreasing trend in the number of offenders who were reconvicted for both sexes across age groups, with age group 21/22 having the highest count and age group 36/37 having the smallest count. A larger proportion of males were reconvicted compared to females across the different age groups. Hence it would be reasonable to suggest both sex and age group of the offender play a key role in determining whether or not an offender is reconvicted. Across age groups, we observe a disproportionate number of offenders who are male (3034) compared to females (415) for the dataset.

It was also found the median number of convictions prior to target conviction for reconvicted individuals was higher than for those who were not reconvicted (see Appendix A). This could also suggest the number of convictions prior to target conviction would aid in the prediction of reconviction.

The target offence variable contains 11 distinct values which describe the principal offence at target conviction. Males dominate across all offence types compared to females, with theft being the most frequently occurring offence type for both sexes (see Appendix A).

Many machine learning algorithms rely on a roughly equal distribution of class labels. Observing heavily imbalanced classes (i.e. one class containing significantly more observations than another) can skew/bias the results, thus leading to poor predictive performance. Therefore, we split the dataset into training and testing sets with a 70%-30% split respectively. We ensure both sets contain balanced splits of the original data using a stratified random split with respect to the distribution of the binary output variable (RECONV).

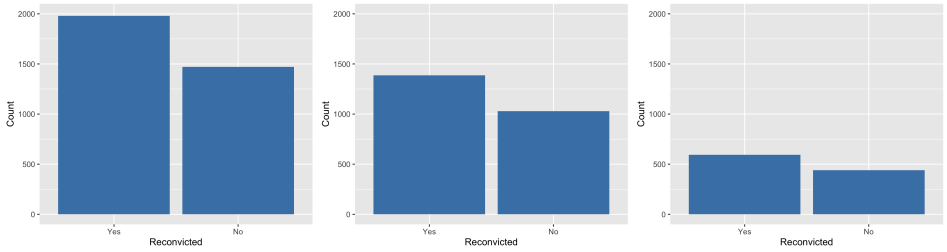


Figure 2: Counts of people reconvicted - whole dataset (left), training set (middle), testing set (right).

From Figure 2 we observe a slight majority of people were reconvicted (1979 people) compared to people who were not (1470 people). Although we observe a slight imbalance in the classes, the observed difference is not heavily disproportionate and we can proceed with our analysis without having to utilise undersampling or oversampling techniques.

3 Model Estimation

3.1 Baseline Model

We begin by implementing the simplest classification method, ZeroR, which simply predicts the majority class ignoring all predictors. This model will be used primarily for determining a baseline accuracy as a benchmark for more complex classification methods discussed in this paper. Any model with an accuracy

less than this baseline accuracy won't be considered as a better model. We begin by constructing a frequency table for the RECONV variable, and we find 1470 people weren't reconvicted (RECONV='No') and 1979 people who were (RECONV='Yes'). Hence, RECONV='Yes' is the ZeroR model for the dataset, yielding a baseline accuracy of $\frac{1979}{3449} \times 100 = 57.4\%$.

For the following models, we repeat 10-fold cross-validation 5 times in order to obtain optimal estimators based on the largest ROC value. We will evaluate the predictive performance of the final estimators returned on the validation set.

3.2 Tree Models

Given we are solving a binary classification problem, predicting reconviction will involve classifying observations into yes or no categories during training instead of obtaining continuous valued outputs like in regression problems. We begin by creating a tree containing all covariates (excluding the identification numbers) to determine which of the criminological variables are good predictors of whether an offender is reconvicted or not by the end of 1999. We utilise the caret package (using R) to facilitate the development of the model; it utilises rpart to construct the tree, whilst creating a tuned model from a predetermined range of maximum tree depth or complexity parameter values. Following cross-validation, the complexity parameter which yielded the highest ROC was 0.007. Repeating the same cross-validation procedure for the maximum tree depth parameter, the highest ROC yielded an optimal depth of 4 (see Appendix B). The maximum depth plot gave similar ROC values above a depth of 4, although 4 was chosen to reduce model complexity while maintaining a reasonable level of performance. We use the complexity parameter of 0.007 to prune the tree to avoid overfitting the model. The pruning procedure is carried out automatically using caret which minimises the cross-validated error.

Visualising the final tree model (see Appendix B), we notice the NUMCONV variable (number of convictions prior to target conviction) has been identified as the most important in determining reconviction. The second most important is LENPRECC (time in years from start of convictions to the target conviction). These covariates are the most significant in terms of determining reconviction, with all other covariates discarded following tree pruning. Following assessment of our model's predictive performance on the validation set we obtain an accuracy of 75.4%, with a 95% confidence interval of (72.69, 78.03).

3.3 Logistic Regression

Next, we implement a logistic regression model, a common method for solving binary classification problems. We begin by modelling the RECONV variable in terms of all covariates. In this way, we can identify any covariates with high levels of statistical significance. We follow a similar procedure as carried out with classification trees, by training a model with 5 repetitions of 10-fold cross-validation to assess predictive performance. Observing the summary output we observe NUMCONV (number of convictions prior to target conviction), being previously convicted of theft (i.e. THEFT = Yes), and males in age groups 31/32 and 36/37 at target conviction are all statistically significant at the 99.9% level. We can interpret the coefficient obtained for NUMCONV as follows - for every one unit increase in the number of previous convictions, the log odds of reconviction increase by 0.15. Converting this 'logit' to a probability, this is equivalent to saying the odds of an individual reconvicting increases by 16.2% for a single unit increase in the number of previous convictions. Performing a similar calculation for those who have previously been convicted of theft, the odds of an individual reconvicting increases by 65.1%.

Being convicted of a previous drugs offence (DRUGS = Yes) and males in age group 26/27 at target conviction are both statistically significant at the 99% level. Interestingly, older age groups in TARGAGE are considered more significant in determining reconviction, with age group 36/37 yielding the smallest p-value. Those who were previously convicted of burglary or fraud and forgery, having a theft offence at target conviction of either criminal damage or drugs offences, and being male are all statistically significant at the 95% level.

We proceed to obtain probability outputs for both classes and classify them into discrete outputs based on a predefined threshold value to obtain the predicted classes. Assessing the predictive performance on the validation set with a threshold value of 0.5, we obtain an accuracy of 74.7% with a 95% confidence interval of (71.89, 77.29).

3.4 Neural Networks

Finally, we implement a shallow neural network (i.e. containing only a single hidden layer) containing a varied number of hidden units. Using the nnet package in R we can easily test across a range of hyperparameter values with 5 repeats of 10-fold cross-validation as previously described. The optimal

model based on these hyperparameters will be returned following cross-validation and will be available for use on the validation set for predicting reconviction.

A range of 1-5 hidden units was explored in the analysis including a range of small decay parameter values (0, 0.1, 0.2 and 0.3) which specify the regularization strength to prevent overfitting. Following cross-validation, the optimal model returned contained 2 hidden units and a decay value of 0.2 (see Appendix C). Assessing performance on the validation set we obtain an accuracy of 75.1% with a 95% confidence interval of (72.29, 77.66).

4 Comparison and Discussion

Model	Accuracy (%)	95% Confidence Interval	Misclassification Rate (%)
Classification Tree	75.4	(72.69, 78.03)	24.6
Logistic Regression	74.7	(71.89, 77.29)	25.3
Shallow Neural Network	75.1	(72.29, 77.66)	25.0

Table 1: Summary metrics for each of the implemented models.

In Table 1 we notice the classification tree yields the greatest accuracy on the validation set with a value of 75.4%, with logistic regression achieving the worst accuracy of 74.7%. In the classification tree, NUMCONV and LENPRECC were identified as the key variables for determining reconviction following automatic pruning. Logistic regression identified further covariates with varying levels of statistical significance, with the key contributors being NUMCONV and being previously convicted of theft. Interestingly, LENPRECC was not identified as statistically significant in the logistic regression model like in the classification tree.

An important aspect of model development is interpretability, that is, how easy is it for people without implementation specific knowledge to understand why the algorithms make the decisions they do. Classification trees are vastly easier to interpret which variables contribute to the outcome, whereas with the neural network model it is notoriously difficult to understand the decision making process. Interpretability is of paramount importance to judges in making their sentencing decision in courts, as well as achieving a desired level of accuracy for predicting reconviction. It is therefore recommended that the classification tree is employed in this context.

Another important consideration is deciding on a threshold value used to obtain predictions. Increasing this value would increase the number of offenders classified as not reconvicted when they were (false negatives). Likewise if the threshold value was to decrease, the number of offenders classified as reconvicted would increase, meaning offenders who didn't reconvict would be classified as being reconvicted (false positives). This is an interesting trade-off, given that a threshold too high would mean only a few offenders who are likely to reconvict would be reconvicted (although many would be missing), and a threshold too low would mean more offenders would be predicted to reconvict when many will not. Given this trade-off, a value of 0.5 provides a suitable compromise.

Despite a worse validation accuracy compared to the classification tree, the neural network model may be likely to achieve higher predictive performance on unseen observations due to the model's complexity. The model accounted for all covariates, whereas the classification tree contained only 2. Hence with a larger sample, it is possible the neural network will surpass that of the classification tree in terms of predictive performance. However, this comes at a cost of a decrease in interpretability as observed in the classification tree.

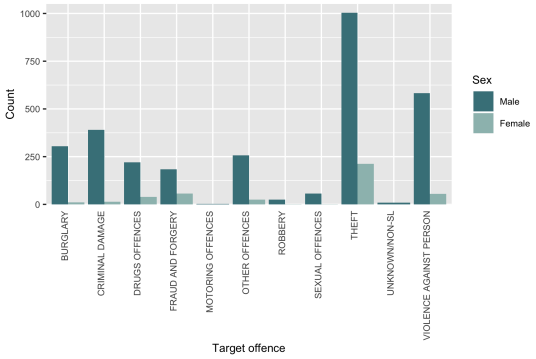
References

[1] J. Copas and P. Marshall, "Royal statistical society publications," Jan 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00104

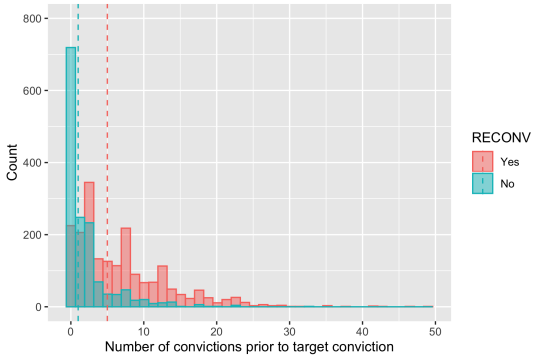
[2] P. Howard, B. Francis, K. Soothill, and L. Humphreys, "Ogrs3: the revised offender group reconviction score," Mar 2009. [Online]. Available: https://www.researchgate.net/publication/258365273_OGRS3_the_revised_Offender_Group_Reconviction_Score

[3] M. of Justice, "Research and analysis on the offender assessment system," Jul 2015. [Online]. Available: https://www.gov.uk/government/publications/research-and-analysis-on-the-offender-assessment-system

A Further Exploratory Plots

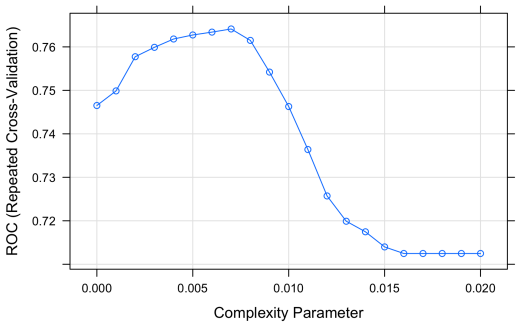


(a) Bar plot of offences at target conviction for each sex.

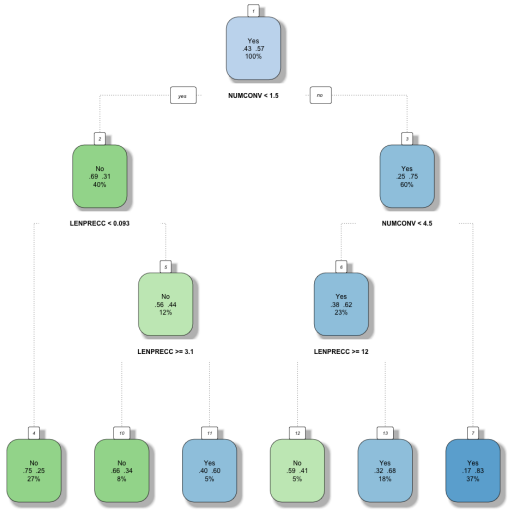


(b) Distribution of reconvicted offenders based on number of convictions prior to target conviction.

B Tree results following cross-validation



(a) ROC values for different complexity parameters.



(b) Final classification tree following complexity pruning.

C Neural network cross-validation results

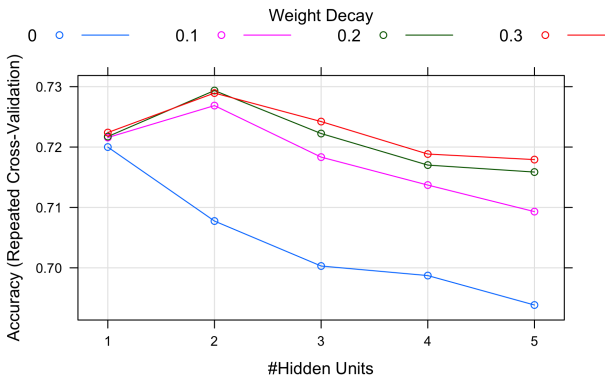


Figure 5: Neural network cross-validation accuracies for hidden units and decay parameters.