

Exploring the Benefits of Data Mining on Juvenile Justice Data QCIF Final Report

Brett Gray, Daniel Birks, Troy Allard, James Ogilvie, Anna Stewart and Andrew Lewis

Justice Modelling @ Griffith (JMAG)



Exploring the Benefits of Data Mining on Juvenile Justice Data

QCIF Final Report

Chief Investigators:

Dr Brett Gray
Mr Daniel Birks
Dr Troy Allard
Mr James Ogilvie
A/Prof Anna Stewart
Dr Andrew Lewis

Produced by Justice Modelling @ Griffith, November 2008

Table of Contents

Chapter 1.	Introduction and Literature Review	3
1.1.	Risk Assessment in the Justice System.....	3
1.1.1.	Importance of Assessing Risk in the Justice System	3
1.1.2.	Actuarial Risk Assessment Tools	4
1.1.3.	Development of Risk Assessment Tools	5
1.2.	Knowledge Discovery and Data Mining.....	6
1.2.1.	Overview of Data Mining	6
1.2.2.	Specific Data Mining Methods	8
1.3.	KDD and Risk Assessment in the Justice System	9
1.4.	Conclusions and Aim of Current Project	10
Chapter 2.	Methodology	12
2.1.	Overview of the Models Developed	12
2.2.	The Dataset and the Data Pre-Processing	13
2.3.	Modelling Methodology	14
2.3.1.	The Estimation and Validation Data Sets	15
2.3.2.	Measurement of Comparative Performance	16
2.3.3.	Data Mining Techniques Adopted	16
Chapter 3.	Results	18
3.1.	Model 1: Predicting if an Offender will have a Subsequent Guilty Reappearance as a Juvenile.....	18
3.1.1.	Aim	18
3.1.2.	Base Cases	18
3.1.3.	Statistical Model	20
3.1.4.	Neural Network Model	20
3.1.5.	Decision Tree	23
3.1.6.	Model Performance.....	24
3.2.	Model 2: Predictive Models of the Most Serious Offence Type for the Next Appearance for Cases that Resulted in a Subsequent Juvenile Re-Appearance	27
3.2.1.	Aim	27
3.2.2.	Base Cases	27
3.2.3.	Statistical Model	32
3.2.4.	Neural Network Model	32
3.2.5.	Decision Tree Model.....	33
3.2.6.	Model Performance.....	33
3.3.	Model 3: Predicting Months to Re-Offence for those Cases that Result in a Subsequent Guilty Juvenile Appearance.....	36
3.3.1.	Aim	36
3.3.2.	Dependent Variable	36

3.3.3.	Base Cases	39
3.3.4.	Statistics Model.....	41
3.3.5.	Neural Network.....	41
3.3.6.	Decision Tree	42
3.3.7.	Model Performance.....	42
Chapter 4.	Discussion.....	44
References		46
Appendix 1: Statistical Analysis Models Developed for the Purpose of Comparison to Data Mining Techniques		50
Appendix 2: Representations of Resulting Decision Trees		62

List of Tables

Table 1: Accuracy of alternative methodologies for predicting re-offending in the juvenile justice system	2
Table 2: Base case for re-offending, probability of re-offending disaggregated by demographics	19
Table 3: Validation set performance for models predicting re-offence	26
Table 4: Base case for re-offending, probability of most common offence type disaggregated by demographics	29
Table 5: Validation set performance for models predicting most serious offence type	33
Table 6: Base case for re-offending, predicting months until re-offence disaggregated by demographics	40
Table 7: Validation set performance for models predicting time to re-offence.....	43
Table 8: Accuracy of alternative methodologies for predicting re-offending in the juvenile justice system	44

List of Figures

Figure 1: Classification performance broken down according to whether a re-offence is observed	27
Figure 2: The percentage of offenders classified correctly in the validation set for the multi-nomial regression, the neural network and the decision tree	35
Figure 3: Percentage of next offence types occurring in dataset and being predicted by the decision tree	35
Figure 4: Histogram of months to next offence	37
Figure 5: Histogram of the log of months to next offence	38

Executive Summary

Risk assessment procedures occupy a central role in the criminal justice system decision making process and typically involve a prediction about the likelihood that an individual will re-offend (Bonta, 1996, 2002; Gottfredson, 1987; Gottfredson & Moriarty, 2006; Hoge, 2002; Taxman, Cropsey, Young, & Wexler, 2007; Taxman & Thanner, 2006). The efficiency of risk assessment tools is determined largely by the predictive accuracy of the underlying analytical techniques and any improvement in accuracy is likely to result in significant benefits for public safety and offender rehabilitation. While risk assessment tools have typically been developed using traditional statistical techniques such as regression models that involve testing relationships to provide evidence for or against a given hypothesis, advances in the development of statistical computation techniques in the Information Technology (IT) fields of Artificial Intelligence (AI) and Knowledge Discovery and Data Mining (KDD) have the potential to improve the predictive accuracy of risk assessments in criminal justice (Caulkins, Cohen, Gorr, & Wei, 1996; Palocsay, Wang, & Brookshire, 2000).

The aim of the project was to apply techniques from the field of KDD (neural networks and decision trees) to criminal justice data to determine whether these techniques could be used to improve the predictive accuracy of models developed to predict risk of re-offending over base cases and commonly applied statistical methods. To accomplish this aim, the predictive accuracy of base cases, models developed using traditional statistical techniques, and models developed using neural networks and decision trees that predicted juvenile re-offending were compared. The predictive accuracy of these techniques was assessed for:

1. Prediction of recontact.
2. Prediction of offence type.
3. Prediction of time to recontact.

Findings indicated that the decision trees had more predictive accuracy than either the base cases adopted or the models developed using traditional statistical methods (Table 1). Additionally, the neural networks outperformed the base cases and traditional statistical models for both classification tasks (recontact and offence type), however was unable to form a model with reasonable predictive capability for the model predicting time to recontact. As

decision trees have consistently resulted in more accurate predictions and given the transparent nature of the outcomes produced by this technique, it has the most potential of the techniques applied.

Table 1: Accuracy of alternative methodologies for predicting re-offending in the juvenile justice system

Models	Predictive Accuracy of Cases Correctly Classified			
	Best Base Case	Statistical Model	Neural Network	Decision Tree
Recontact	67.51	72.40	76.35	76.45
Next offence type	30.30	33.20	34.01	33.69
	Standard Deviation of Residual			
Time to re-offence	1.11	1.12	1.17	1.09

The findings indicate that KDD techniques can be applied to improve predictive accuracy. There is considerable potential for greater application of these techniques to develop risk assessment tools and any improvement in prediction accuracy will result in more effective criminal justice system decision-making. The potential implications of improved accuracy are considerable given the wide ranging criminal justice system processes that require an assessment about risk of re-offending and the important role that such assessments have for public safety and the targeting of offender rehabilitation programs.

Chapter 1. Introduction and Literature Review

This chapter will highlight the importance of assessing risk in the justice system, outlining the benefits of actuarial risk assessment tools and how these tools have been developed using traditional statistical techniques. The advantages of data mining techniques that can be used to predict re-offending will be examined, along with specific techniques including neural networks and decision trees. The limited research that has assessed the usefulness of neural networks to predict re-offending and decision trees will then be investigated. The chapter will conclude noting that data mining techniques may improve the accuracy of models developed to predict re-offending and the considerable benefits that could result from improving the accuracy of prediction in the justice system.

1.1. Risk Assessment in the Justice System

This section will highlight the importance of assessing risk in the justice system for improving public safety and offender rehabilitation, outlining the advantages of actuarial risk assessment tools. The use of traditional statistical analyses such as regression analyses and survival analyses to develop risk assessment tools will be explored, along with the potential that new methodologies may have for improving the predictive accuracy of tools assessing risk of re-offending.

1.1.1. Importance of Assessing Risk in the Justice System

While the concept of ‘risk’ differs across contexts and may be considered the outcome of abstract factors that increase the likelihood of a particular outcome (Silver & Miller, 2002), such assessments in the justice system typically involve a prediction about the likelihood of *re-offending*. The importance of risk assessment in the justice system is highlighted by the broad ranging processes that require such an assessment and given its role in improving public safety and offender rehabilitation. Justice system processes that require an assessment about risk include bail, sentencing, prisoner classification, parole, the case management and supervision of community based orders and the provision of effective treatment (Gottfredson & Moriarty, 2006; Silver & Miller, 2002). Any improvement in the ability to accurately assess risk would improve the efficiency of criminal justice decision making. Risk

assessment provides a useful tool for the attainment of public safety by enabling the identification of offenders who pose an elevated risk of recidivism who require greater supervision. Consistent with the principles of best-practice for offender rehabilitation, risk assessments can also be used to target interventions, with high-risk offenders receiving intensive interventions and low-risk offenders receiving either none or minimal interventions (Andrews et al., 2006).

1.1.2. Actuarial Risk Assessment Tools

Research findings consistently indicate that decision-making based on actuarial risk assessment tools is more accurate, valid and reliable than clinical decision-making (Ægisdottir, White, Spengler et al., 2006; Dawes et al., 1989; Gambrill & Shlonsky, 2000; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Hanson, 2005). A risk assessment tool is a formalised and standardised assessment method to provide a uniform structure and a set of criteria for identifying and measuring risk among individuals in a population (Cicchinelli, 1995). Actuarial risk assessment tools help decision-makers focus on important information and structure clinical thinking and can take into account significantly more information than human experts in order to classify individuals on a continuum of risk, based on risk-related characteristics such as offence history, substance use, cognitive impairments and employment status. Actuarial tools also have the benefit that they do not require highly trained professionals in their implementation, which has the potential to conserve significant institutional staffing resources. Furthermore, actuarial risk assessment tools shift the burden of decision-making from judgements based on professional expertise that are susceptible to a range of biases to judgements derived from empirically established associations (Silver & Miller, 2002).

Risk assessment tools comprise a number of items that appraise a constellation of pertinent risk factors and, in some cases, protective factors. Risk factors are those variables that produce an elevated risk of recidivism while protective factors are those variables that buffer an individual from engaging in offending behaviour thereby diminishing the risk of recidivism (Andrews, Bonta, & Wormith, 2006; Bonta, 2002; DeMatteo & Marczyk, 2005; Douglas, Cox, & Webster, 1999; Farrington, 2002; Rogers, 2000). While it is typically argued that a holistic risk assessment incorporating both risk and protective factors is

required to obtain accurate assessments about risk (Rogers, 2000), very little is known about potential protective factors (Carr & Vandiver, 2001; DeMatteo & Marczyk, 2005; Morrison et al., 2002). Therefore, some assessments exclude these protective factors from contributing to overall risk/needs scores (e.g., Baker et al., 2003; Casey & Day, 2004; Thompson & Pope, 2005). Actuarial tools use aggregate group data in order to group a heterogenous collection of individuals into subgroups that vary along a continuum of risk.

1.1.3. Development of Risk Assessment Tools

Actuarial risk assessment tools incorporate criteria that have been demonstrated, through prior statistical assessment, to have a high association with recidivism, where the combination of multiple risk factors increase the likelihood of recidivism at the aggregate level (Cash, 2001; Dawes, Faust, & Meehl, 1989; Silver & Miller, 2002). Arguably the most important quality of a tool is its predictive validity and the primary criterion for assessing predictive validity is recidivism (Silver & Chow-Martin, 2002; Silver & Miller, 2002). Prediction errors include *false positive predictions*, which refer to cases predicted to re-offend that do not do so and *false negatives*, which refer to those cases predicted to not offend that do go on to offend. Errors of prediction have impeded the wide-spread adoption of actuarial tools to assess risk within offender and mental health populations (Monahan, Steadman, Appelbaum et al., 2000; Steadman, Silver, Monahan et al., 2000). Typically, actuarial risk assessment tools have had unacceptably high rates of false positive predictions. This situation highlights the need to consider new methodologies of constructing actuarial risk assessment tools that have the potential to improve the predictive accuracy of tools.

Many risk assessment tools that have been developed to assess risk of re-offending are based on simple linear regressions and the use of survival analysis to examine recidivism patterns over time based on specific constellations of risk factors is becoming increasingly common. For example, Visher, Lattimore and Linster (1991) used a multivariate survival model to examine recidivism patterns and develop risk assessment profiles for a sample of 1,949 serious juvenile offenders. The survival model allowed the researchers to examine the characteristics of subpopulations of youthful offenders based on the time to re-offence, and the covariates associated with an earlier or later time to such a failure. A range of criminal history and current commitment variables were found to be significant predictors of the

timing of rearrest for the youthful offenders. Furthermore, the survival model was demonstrated to estimate and assign statistically valid risk functions (i.e., risk assessment) to individuals based on their characteristics that had the potential to be used to guide decisions regarding treatment and supervision requirements.

Technical innovations in the analytical capabilities of IT may enable risk assessment tools with improved predictive accuracy to be developed (Brennan et al., 2004; Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). The methodologies have been applied to decision-making and risk assessment processes as they enable multiple factors that may affect an outcome to be considered in a nonlinear fashion, rather than focusing on single factors alone that are presumed to have linear associations with outcomes. In this sense, decision-making processes are made more ecologically valid, since real-world outcomes are affected by a range of factors with complex interactions and nonlinear relationships. As a result, decision-making may be more accurate and the error associated with predictions of future outcomes may be reduced.

1.2. Knowledge Discovery and Data Mining

This section will provide an overview of the advantages of methodologies such as data mining and the three main types of algorithms that may be used. Specific data mining methods including neural networks and decision trees will be examined, along with the limited research that has examined the usefulness of applying these techniques to criminal justice data to predict risk of re-offending and develop actuarial risk assessment tools.

1.2.1. Overview of Data Mining

KDD or data mining refers to the use of sophisticated multidimensional data analysis tools to discover latent, undiscovered, valuable, and nontrivial patterns and relationships where emphasis is on the semi-automated extraction of relationships from large datasets (Fayyad et al., 1996). It is best described as an iterative and exploratory process achieved through either automated or manual methods. The two primary roles of data mining are *prediction*, which involves the use of variables to predict unknown future events or values of a given outcome and *description* involving the identification of patterns that describe data in a meaningful manner.

The increased use of data mining techniques to examine and identify patterns in data is based on the ever-increasing quantities of data being generated by comprised databases. The analysis of large volumes of data using traditional hypothesis-based linear statistical analyses makes it difficult for analysts to identify relevant relationships. This is due to volumes of data being too large to manage, and data structures being too complicated to analyse effectively. While traditional statistical analyses involve the formulation of hypotheses that are then supported or rejected based on the findings of statistical tests, data mining techniques are used to extract relationships that are present in data that may not have been foreseen by an analyst, thus enabling a semi-automated knowledge discovery process (Seifert, 2004).

Data mining involves using a range of techniques using statistical models, mathematical algorithms, and machine learning methods to examine potential relationships in data sets and are often used to form predictive models of either continuous or categorical variables. Algorithms for data-mining can be classified according to the following distinction:

1. Methods use to discover predictive relationships for categorical variables (i.e.: classification methods).
2. Methods used to discover predictive relationships for numeric variables.
3. Methods of association rule discovery.

For example, a *predictive relationship* that could be discovered for a *categorical variable* could involve modelling the court orders (e.g. probation order, community service, or detention) assigned to juvenile offenders. Outcomes could indicate that the court outcome was dependent on the offenders age, the offence committed and the number of matters heard at a given court appearance. The modelling of a *numeric variable* may involve developing a model of time to re-offence after an offender's second court appearance based on the time between first and second court appearance and the offender demographics. *Association rule discovery* (Agrawal & Srikant, 1994; Gray & Orłowska, 1998) differs from the above two cases in that a dependent variable does not need to be specified. The algorithm searches for association rules where attributes of the data co-occur. Association rule discovery was formulated in the context of supermarket purchases where it may be found, for example, that 80% of people who purchase milk also purchase bread.

1.2.2. Specific Data Mining Methods

Some of the more common data mining methods include neural networks, decision trees, support vector machines and algorithms for mining association rules. A *neural network* is a form of statistical method that may be used to construct dynamic models of interactions among variables for the purposes of regression and classification (Paik, 2000). Neural networks are generally composed of a collection of elementary processing units interconnected by weighted connections or “relationships” of a particular strength (Paik, 2000). The most common neural network is known as the multi-layer perception (McClelland & Rumelhart, 1988). For the purpose of analysis, this technique can be viewed in terms of its function approximation capabilities. In the case of standard regression techniques (such as linear regression) a given surface is fitted to the data that best models a numeric dependent variable. In the case of linear regression, this surface is a linear function of the independent variables. A function approximation technique will mould such a surface to any continuous function that best models that data, as such complex non-linear relationships can be modelled. Neural networks can be used for both regression of a numeric dependent variable and classification of a categorical dependent variable.

Decision trees (Quinlan, 1986; Quinlan, 1987) form a tree of decisions that best segregates the data into bins that are predictive of a categorical attribute. Each decision point on the tree is a condition on an attribute of the data (e.g. gender = male). The tree may have many layers of such decision points based on different attributes. At each node of the tree, the distribution of a dependent variable is represented. The tree is formed by an algorithm that attempts to form nodes of the tree in which the distribution of the dependent variable for each node is maximally different to its sibling nodes (i.e. nodes with the same parent). This separation can be formed by a number of heuristics such as entropy or a p-value of a statistical test. Decision trees are usually used to model categorical variables, however, a similar algorithm can be used to form decision trees to model numeric variables (Breiman, Friedman, Olshen, & Stone, 1984).

1.3. KDD and Risk Assessment in the Justice System

The ability of AI and KDD methods to predict future events or output values is an important tool for many scientific disciplines. Given the central role of classification and prediction for the management of offender populations and the potential value of KDD methodologies for improving the accuracy of predictions, these techniques have been underutilised (for example, see Adderly & Musgrove, 2001; Brennan et al., 2004; Estivill-Castro & Lee, 2001 ; Lin & Brown, 2006; Strano, 2004; Zhang, Salerno, & Yu, 2003). This section will outline research that has used neural networks to predict re-offending and decision trees to develop actuarial risk assessment instruments.

There have been a number of applications of *neural network* methods to predict criminal recidivism in the criminological literature. In one of the earliest applications of a neural network methodology to the prediction of recidivism, Caulkins et al (1996), using a sample of 2,385 offenders released from prison, found that neural network models did not increase the predictive accuracy over multiple regression models. They argued that the failure of the more complex neural network models to improve predictive accuracy concerning recidivism was related to inadequate knowledge of the mechanisms linking risk factors to recidivism, and limitations in measuring these variables. Palocsay et al (2000) used neural network models to predict criminal recidivism by splitting an offender population into two groups: non-recidivists and eventual recidivists. Results suggested that the neural network models obtained significantly higher predictive accuracy in classifying offenders as recidivists and non-recidivists compared to logistic regression models. Palocsay et al (2000) highlighted the issue that the predictive accuracy of neural network models depends heavily on the choice of network topology, such as the number of hidden layers and nodes in each layer, the training methodologies used, and node activation functions. Few guidelines exist for researchers and practitioners to develop neural network models for the purposes of predicting criminal recidivism. However, it was also noted that neural network models provide a significant degree of flexibility and adaptability over traditional statistical models to produce superior performance in recidivism prediction.

Several researchers have used *decision trees* to construct actuarial risk assessment instruments. Monahan et al, (2000) utilised an Iterative Classification Tree (ICT) methodology to construct an actuarial instrument for violence risk using a sample of 939 civil

psychiatric patients. The ICT instrument was shown to partition 72.6% of the sample into either high or low risk of violence categories based on 106 risk factors derived from hospital records. The benefits of the ICT approach were argued to be its abilities to model violence as an outcome reached by multiple routes, and its use of two cut-off scores for identifying both high and low risk cases (Monahan et al., 2000; Steadman et al., 2000). Rosenfeld and Lewis (2005) applied a Classification and Regression Tree (CART) approach to violence risk assessment using a sample of 204 stalking offenders. The models constructed through the CART approach were found to have high levels of predictive accuracy comparable to logistic regression models. Of particular importance, a benefit of the CART approach was the relative simplicity of its application in clinical practice compared to logistic regression models. Additionally, Stalans, Yarnold, Seng, Olson and Repp (2004) found a decision tree methodology (classification tree analysis) to exhibit higher levels of predictive accuracy over logistic regression models at identifying violent recidivism risk among a sample of 1,344 violent offenders on probation. Therefore, decision tree methodologies display important advantages over more commonly used statistical models such as logistic regression for the development of risk assessment instruments.

1.4. Conclusions and Aim of Current Project

KDD and AI analytical techniques have great potential to assist in improving the predictive accuracy of decision-making processes and instruments aimed at assessing and predicting the risk of recidivism in criminal justice settings. Current empirical research strongly suggests that these advanced analytical techniques display higher levels of predictive accuracy over traditional statistical methods such as regression analyses in the prediction of criminal recidivism. Improvements in the predictive accuracy of assessments about risk of re-offending would result in more efficient criminal justice decision-making and improve public safety and offender rehabilitation. Furthermore, due to their abilities to more accurately model human decision-making processes in a more ecologically valid manner, these techniques have the added benefit of being more intuitively appealing to professionals in criminal justice practice.

To explore whether KDD techniques could improve the predictive accuracy of assessments about re-offending, the project explored and compared the predictive accuracy of base cases,

models developed using traditional statistical techniques, and models developed using neural networks and decision trees that predicted juvenile re-offending. The predictive accuracy of these techniques was assessed for the prediction of (i) recontact, (ii) next offence type, and (iii) time to recontact.

Chapter 2. Methodology

This chapter will provide an overview of the models that were developed, the dataset used to develop the models and pre-processing, and the methodology that was used to develop the models and assess their comparative predictive accuracy. Additionally, the data mining techniques adopted in the project (neural networks and decision trees) will be outlined.

2.1. Overview of the Models Developed

Models were developed to explore whether neural networks and decision trees produced more accurate predictions than base rates and models developed using traditional statistical techniques (logistic, multi-nomial and linear regressions) for:

1. Prediction of recontact.
2. Prediction of offence type.
3. Prediction of time to recontact.

The first set of models addressed the question: For a given guilty court appearance, does the offender reappear in court at a future date with a guilty outcome as a juvenile? These models have the highest operational relevance of the models developed as improvements in the accuracy about whether a juvenile will re-offend could result in considerable improvements in public safety and offender rehabilitation.

The question that the second models addressed was: For those cases where a subsequent guilty appearance as a juvenile occurs, what category of offence type is committed? At a statistical level, this model presents an additional difficulty over model one as rather than having a binary outcome (i.e. does vs. does not re-offend), the model is attempting to predict eight different outcomes in the dependent variable as offence type is classified into eight categories. Additionally, if such a model was accurate at predicting if a future crime was likely to be a more serious crime, such as a violent crime, this model could aid in targeted intervention decisions in a similar way to that described for model 1.

The third set of models addressed the question: For those cases where a subsequent guilty appearance as a juvenile occurs, how many months from a court appearance is the next crime

likely to occur? The importance of this model is that the statistical and data mining methods adopted to predict a variable are quite different depending on whether a dependent variable is numeric (e.g. how many months) versus categorical (e.g. does re-offend or offence type category). As such, the inclusion of this model ensures that data mining methods have been explored on both categorical (model one and two) dependent variables as well as numeric (this model). The third model also has a similar role to model two in aiding the assessment of how serious a re-offender may be, as a serious offender is much more likely to re-offend at a greater rate.

2.2. The Dataset and the Data Pre-Processing

The dataset adopted for the analyses contained information relating to Queensland juvenile court appearances to July 2007. This data included information relating to court appearances, offender demographic characteristics (such as age, gender and Indigenous status), the type of crime committed for the most serious matter heard at an appearance and court outcome information. An important part of any data analysis process is to pre-process the data to make it suitable for analysis.

The first step in the process was to subset the data to only appearances that did not have a court outcome of not guilty, as well as subsetting appearances to only contain records where the offender was aged between 10 years and 17 years and 1 month. Juveniles are not meant to be processed in the juvenile court system until age 10. The cut off age of 17 and 1 month was chosen due to the number of appearances of juveniles over this age declining rapidly. Even though there were still appearances over this age, the data indicated that an increasing proportion of appearances were being processed in the adult system. As the predictive model does not need to accommodate a probability for older children to have been processed in the adult system, the age 17 and 1 month was chosen as a cut off where the majority of children were still processed as juveniles.

After this subset, a number of variables in the dataset were created involving transformations of existing variables. All variables adopted in the models will be detailed in the section describing the model where they were used. The dependent variables were created to reflect if an offender had a subsequent appearance, and if so the type of crime committed at the

subsequent appearance and how many months would occur before this subsequent appearance.

Following this creation of variables, the data was then subset again to only contain appearances between the dates 1/7/1999 and 1/1/2007. There is a substantial drop in the rate of appearances before this date range, due to inferior data collection occurring prior to this date. As such, this missing data would create biases in the model and was better to subset all records preceding 1/7/1999. The data after the date 1/1/2007 was also subset. This resulted in subsetting out the most recent six months of data. The reason for this was because of an “edge effect”. The edge effect occurs when an offender may not have a subsequent appearance, however, if the data were to continue for a longer period of time, it is likely that a subsequent appearance would occur. For example, if an offender appeared in court in June 2007, even though that offender may be likely to re-offend, it is unlikely that the subsequent appearance would be observed before the 1/7/2007 end of the dataset, thus potentially creating a bias in a model that would be estimated on the complete dataset. It was observed that when an offender will re-offend, 70.8% of these offenders will re-offend within six months. As such by removing the last six months of data, the likelihood of this bias occurring is substantially decreased. It should be noted here that the dependent variables containing information relating to if the offender re-offends has been constructed before this subset, so if an offender had an appearance in December 2006 and March 2007, while the March record would be subset out here, the December record would display a subsequent re-appearance will occur in the dependent variables being modelled (as desired to reduce the discussed bias).

2.3. Modelling Methodology

The modelling methodology adopted for all of the models was based on a separation of data into estimation and validation data sets. The predictive performance of the models was also compared to a number of base cases and models developed from more traditional statistical methods.

2.3.1. The Estimation and Validation Data Sets

The processed juvenile court data was separated into two datasets, an estimation data set and a validation data set. The notion here is that models are estimated on the estimation data set and then the validation set is used to measure the predictive performance of models and compare this performance between models. Data mining models have a greater ability to extract relationships that may be present in data than traditional statistical methods which are usually based more on a hypothesis validation procedure.

Due to this ability to extract the relationships that are present in the data, data mining techniques can sometimes overfit to the data. This implies that the model may extract relationships in data to a degree that it is over fitting to peculiarities in the estimation dataset that will not generalise to predictive relationships outside the estimation set. For this reason, performance of a model on the estimation dataset is not a good indicator of predictive performance, and once estimated a models performance must be measured on data it has not seen during its estimation period. As such, the validation data set is used as an indicator of a models predictive performance and comparative performance between models.

The pre-processed juvenile court data contained 45,185 court records across 18,201 individual juveniles. Additionally, in this data, 28,103 appearances were by a juvenile that would have a subsequent guilty juvenile court appearance while 17,802 appearances had no subsequent appearance. To separate this data into estimation and validation sets, individual juveniles were randomly selected for either data set with 10,000 juveniles being placed in the estimation set and 8,201 being placed in the validation set. The separation of data was based on individual juveniles so that if different juveniles exhibited different re-offending behaviour, this would be reflected in the data separation and predictive performance measurements would be taking this into consideration.

This sampling procedure resulted in 24,794 guilty court appearances in the estimation set with 15,483 of these appearances resulting in a subsequent appearance. The validation set contained 20,391 appearances with 12,620 appearances resulting in a subsequent guilty appearance.

2.3.2. Measurement of Comparative Performance

The predictive performance of a model must be measured in a comparative manner. The question is not simply how well a model predicts the required outcome, but how strong is a prediction in comparison to other predictive measures. As such, to measure the performance of each of the models, a number of base cases are established. These cases range from a simple average of the data to what would be predicted when taking into account demographic information such as gender, Indigenous status, age group and appearance number. These base cases are quite simple and used to give a very basic assessment of the accuracy of data. In order to be of value, a model should provide greater predictive accuracy than the base cases.

To further assess the performance of the data mining models, models adopting more standard statistical procedures (logistic, multi-nomial and linear regressions) were also developed. By comparing the models developed to these standard statistical models, the predictive accuracy of neural networks and decision trees could be compared to what would be the status quo. These statistical models are documented in Appendix 1.

The base cases presented are both estimated and have their performance measured on the validation data set. Given the large amount of data used for estimation and the limited number of parameters used in these base case models this approach will not result in over fitting of estimations to the data. If anything, this method may provide these base cases with a slight advantage in accuracy over what would occur in real-world application. Therefore, any subsequent models which outperform them will have done so against the best case scenario for these base case models.

2.3.3. Data Mining Techniques Adopted

The data mining models adopted here include neural networks and decision trees. These methods were adopted due to their prevalence in the field of data mining as well as their proven ability to form models across a wide range of application areas. While data mining is comparatively new compared to traditional statistics and many other methods of analysis, the field is mature enough that a set of common techniques that have proven versatile and effective across many areas has emerged. Neural networks and decision trees have emerged

as some of the most versatile and accurate of the techniques available. As this project is exploratory and assessing the potential for data mining to be applied to criminological data and formulate predictive models that may be more accurate than other commonly used methods for analysing criminological data, it was considered important to adopt well established techniques.

It is important to note a few fundamental differences between neural networks and decision trees in their practical use to form predictive models. The model formed from a neural network is typically considered to be a 'black box'. If a neural network establishes an accurate predictive model, it is not possible to determine the nature of the relationship between the independent and dependent variables. As such, it is often the predictive performance established on a validation data set that is used as the justification for the use of the model. Decision tree models produce a tree of decisions based on the values of the independent variables and this is used to assess the predicted outcome. As such, the resulting model is completely transparent and an analyst can determine the exact structure of the model and how the independent variables are used to arrive at its prediction. Of course, measured accuracy on a validation data set is also an important part of the justification for the use of the model, but its transparency does tend to provide additional confidence in its use, as well as insight into the underlying relationships in the data.

A final distinction between the two types of techniques is the nature of the models they establish. In the case of a neural network, as detailed above, a continuous form of a regression surface is moulded to the data. As such, the internal workings of the network are continuous in nature, rather than binary decisions. When a categorical outcome is being predicted, it is usually based on a discrete decision process on the output of the potentially multi-dimensional continuous regression surface. In the case of a decision tree, each point of the models decision process is a discrete (or binary) decision point. This means the tree is ultimately slicing the independent variable space into regions of different predictions (or distributions of predictions). This is the reason for the transparency of the resulting model, but may also mean that different predictive problems are more suited to one or the other approach.

Chapter 3. Results

This chapter will present the results of models developed to predict recontact, offence type and time to recontact. For each of the main models developed, the aim of the model, base cases, standard statistical model, neural network model, decision tree, and model performance will be examined.

3.1. Model 1: Predicting if an Offender will have a Subsequent Guilty Reappearance as a Juvenile

3.1.1. Aim

The aim of this model is to predict if an offender will have a subsequent guilty court appearance as a juvenile.

3.1.2. Base Cases

Two base cases were constructed for this model. The first case was simply to compute that on the validation data set 61.9% of cases resulted in a re-offence. As such, to simply predict that any offender will re-offend will achieve a classification accuracy of 61.9%. This can be considered an “unintelligent” base case, the accuracy of which can be achieved with no real analysis, and as such should form a floor to the level of accuracy we are hoping for from the model.

The second base case was to make this estimate more sophisticated. Previous work relating to simulation modelling of the juvenile justice system (Stewart, Spencer, O'Connor, Palk, Livingston, & Allard, 2004) used logistic regression and found that the best predictors of re-offending were gender, Indigenous status, age group broken into two groups (10-14 vs. 15+) and the number of prior appearances with an upper limit (upper limit is the 5th appearance or greater). The combination of these variables creates a number of bins (or groupings) for the data. Table 2 presents the values resulting from breaking the data into these groupings and displays the percentage of appearances that result in a re-appearance.

Table 2: Base case for re-offending, probability of re-offending disaggregated by demographics

Age Group	Indigenous/Gender	Appearance Number	Re-offends?			
			No		Yes	
			n	Row %	n	Row %
10 to 14	Non-Indigenous Male	1	403	35.14	744	64.86
		2	77	15.98	405	84.02
		3	29	10.55	246	89.45
		4	12	6.70	167	93.30
		5+	11	3.20	333	96.80
	Indigenous Male	1	113	16.67	565	83.33
		2	39	9.38	377	90.63
		3	13	4.13	302	95.87
		4	13	5.22	236	94.78
		5+	21	2.68	764	97.32
	Non-Indigenous Female	1	219	48.67	231	51.33
		2	32	26.67	88	73.33
		3	8	15.38	44	84.62
		4	4	14.81	23	85.19
		5+	3	6.38	44	93.62
	Indigenous Female	1	53	21.46	194	78.54
		2	20	13.70	126	86.30
		3	8	8.08	91	91.92
		4	2	2.90	67	97.10
		5+	13	9.49	124	90.51
15+	Non-Indigenous Male	1	2011	66.70	1004	33.30
		2	720	52.52	651	47.48
		3	347	42.16	476	57.84
		4	203	36.06	360	63.94
		5+	563	31.47	1226	68.53
	Indigenous Male	1	353	51.01	339	48.99
		2	195	38.16	316	61.84
		3	139	33.82	272	66.18
		4	109	31.87	233	68.13
		5+	463	26.78	1266	73.22
	Non-Indigenous Female	1	713	77.75	204	22.25
		2	178	55.11	145	44.89
		3	77	42.54	104	57.46
		4	49	39.20	76	60.80
		5+	89	43.20	117	56.80
	Indigenous Female	1	181	61.15	115	38.85
		2	86	45.99	101	54.01
		3	41	32.03	87	67.97
		4	37	34.26	71	65.74
		5+	124	30.24	286	69.76

For the second base case, the most likely outcome (of re-offend or does not re-offend) for each appearance based on the demographic of the offender in Table 2 was adopted as the prediction of whether the young person re-offends. This resulted in a classification accuracy of 67.51% on the validation set.

3.1.3. Statistical Model

A logistic regression was performed to formulate a statistical model to predict re-offending (Appendix 1). The estimation dataset was used to estimate the model and then resulted in a predictive accuracy of 72.40% of cases classified correctly on the validation set.

3.1.4. Neural Network Model

The neural network model adopted the following independent variables:

Indigsex: This variable represented the Indigenous status and gender of the offender, adopting the following values:

Value	Represents
1	Non-Indigenous male
2	Indigenous male
3	Non-Indigenous female
4	Indigenous female

Offtype: This variable represented the type of offence committed for the most serious matter at the current appearance and adopted the values:

Value	Represents
1	Offences against the person
2	Break and enter, burglary
3	Theft and related offences
4	Drug offences
5	Traffic offences
6	Public order offences
7	Property damage
8	Other offences

Crtorder: This variable represented the court outcome of the current offence, adopting the values:

Value	Represents
3	Divert from formal order
4	Non supervised order
5	Community supervision
6	Immediate release
7	Detention served

AppNumMaxFactor: This variable represented the appearance number of the current appearance with a maximum bound of 5. So if, for example, an offender had two prior offences to the current offence, this variable would adopt the value 3.

MonthsTillEdgeEffectCategorical: This is a categorical variable that is a representation of how much longer the offender has in the dataset to be able to offend again. This is the minimum of one of the following, the number of months until the offender turns 17 years and 1 month, or the number of months between the current appearance date and the 1/7/2007, beyond which a subsequent appearance would not have appeared in our dataset. This

number of months was then categorised into the following values that the variable may adopt:

Value	Number of Months
1	< 0.25
2	0.25 to 0.5
3	0.5 to 0.75
4	0.75 to 1
5	1 to 1.5
6	1.5 to 2
7	2 to 3
8	3 to 4.5
9	4.5 to 6
10	6 to 7.5
11	7.5 to 9
12	9 to 12
13	12 to 18
14	18 to 24
15	> 24

AgeContinuousAtAppearance: This represents the age of the offender in years at the current appearance. Note that the variable is continuous so an example of a value may be 16.34.

AgeContinuousAtFirstAppearance: This variable is a continuous representation of the age of the offender at their first appearance. This variable would potentially add useful information relating to the age of initiation, as an early onset can be an indicator for continued criminal activity.

All variables in the model except *AgeContinuousAtAppearance* and *AgeContinuousAtFirstAppearance* were adopted as factors, meaning that for each variable there were a number of inputs into the neural network, each adopting the value 0 or 1 representing a single possible value of a variable (for example, one of the neural network inputs is an input that adopts the value 1 if the offence type is an offence against the person and a 0 otherwise, with separate inputs for each other possible value).

The neural network was tested in different cases with different numbers of hidden nodes (this determines some of the architecture in neural networks and is well known in the neural network literature). The final number of hidden nodes adopted was 13. The network performance on the validation set achieved a classification accuracy of 76.35%, exceeding both the base cases and the logistic regression.

3.1.5. Decision Tree

The decision tree was estimated with the same independent variables as for the neural network as well as the following additional variables:

TimeSincePreviousAppearanceMonths: This variable represented the number of months since the offenders' previous appearance (as a continuous variable). If this is the offenders' first appearance, the variable adopts the value 0, however the tree should be able to determine this from the input AppNumMaxFactor which would adopt the value 1.

MostSeriousCourtOutcomeToDate: This variable adopted the same values as the variable crtorder above, but represented the most serious of these outcomes for all appearances to date for the given offender.

AverageMonthsBetweenAppearances: The average number of months between appearances for the offender to date, or the value -1 if this is the offender's first appearance.

AverageMattersPerAppearance: The average number of matters heard per court appearance for the offender to date.

Mattnum: The number of matters heard at the current court appearance.

It was speculated that adopting too many input variables in the neural network model could present difficulties to the network as each additional variable increases the dimensionality of the input space, however it was not believed this would be a problem for the decision tree, due to the nature of how it computes the model by dissecting the independent variable space into regions with delimiters.

The resulting tree resulted in a classification accuracy of 76.45% on the validation set, thus showing the strongest performance of all the models. As detailed above, the resulting model of a decision tree is transparent to the analyst and a representation of the tree formed is shown in Appendix 2. As can be seen from the tree, the most significant decisions were based on the age of the offender at the current appearance and the appearance number, with further decisions being based on the offenders demographic characteristics (Indigenous status and gender), the resulting court order, the offence type and information relating to the number of matters at the appearance.

3.1.6. Model Performance

Two metrics were applied to each of the models on the validation dataset to assess the performance of the models:

1. Percent of Court Appearances Classified Correctly.
2. A metric derived from the maximum likelihood metric used in the estimation procedure of many analysis methods.

The percent of court appearances classified correctly is fairly self explanatory. To understand the motivation for the second metric it is useful to consider the information content in the results of these models. One use of these models is to return a classification of whether an offender at a given court appearance will re-offend as a juvenile; however a possibly more useful result is a probability if they will re-offend. For instance, consider a case where there are two models both of which are predicting an offender will re-offend, as such giving the same outcome. However, assume that one model is assigning a 63% probability of re-offending while another model is assigning a 74% chance of re-offending. The probabilities of one of these models will be more accurate than the other when applied to a large number of court appearances. This is potentially very useful information. Such probabilities would allow a model to move beyond a simple classification to a metric that could be used to assess the risk level for a given offender, beyond a yes/no classification.

To assess the accuracy of the probabilities produced by the various models, a metric based on the likelihood metric adopted in statistical estimation was used. Maximum likelihood involves summing the log of the likelihood a model assigns to an observed point in the data.

To derive this metric, the likelihood is the probability a model assigns to the observed outcome for whether an offender will re-offend as a juvenile. A log to base is then applied to this probability, as performed in standard maximum likelihood estimation and then rather than summing these values, an average of this value over all cases observed in the validation set is taken. Adopting an average will be monotonically equivalent to a sum (that is, if the sum of the log likelihoods would be greater for a model, then the average would also be greater). The values observed for this metric will be negative, and the greater the value, the greater the fit of the model.

There are several points to consider about the calculation of this metric. The output of a neural network in its standard architecture is not well suited to measuring a probability of the outcome rather than a classification, and as such this metric is not applied to the neural network. Additionally, decision trees sometimes estimate a zero or 100% probability of an outcome, which will not work well with a log operation. As such, probability estimates for all models were bound to be a maximum of 95% or a minimum of 5%.

For our base cases, as the classification is based on the percent of observed offenders re-offending in the validation dataset (or this percentage broken into data bins for base case 2), this percentage was adopted as the percentage estimate of the model.

The performance of the various models on the validation set according to these metrics is presented in Table 3. The logistic regression outperforms both base cases for the percentage of appearances classified correctly while both the neural network and the decision tree outperform the logistic regression and the base cases. While the classification performance of the decision tree is slightly higher than the neural network, the difference is marginal. As the decision tree produces a transparent model and is able to accurately produce percentage chances for outcome, it would be the preferred model in these circumstances over the neural network. The neural network would need to quite substantially outperform a decision tree to overcome these relative deficiencies

Table 3: Validation set performance for models predicting re-offence

	% Classified Correctly	Avg Log Likelihood
Base Case 1	61.90	-0.67
Base Case 2	67.51	-0.56
Statistics Model	72.40	-0.56
Neural Network	76.35	Unknown
Decision Tree	76.45	-0.53

In terms of the likelihood measurement adopted, the logistic regression shows a comparable performance to the second base case, both of which outperform base case one. The decision tree outperforms the logistic regression and all base cases, indicating more accurate probability assignment to outcomes.

To further assess the classification accuracy of the models, the classification performance of the models has been broken down according to the classification accuracy of appearances that result in a re-offence as well as the accuracy of appearances that do not result in a further juvenile offence (Figure 1). The data mining models outperform both the logistic regression and base case 2 for both possible outcomes with the decision tree achieving its greater accuracy for cases that do re-offend.

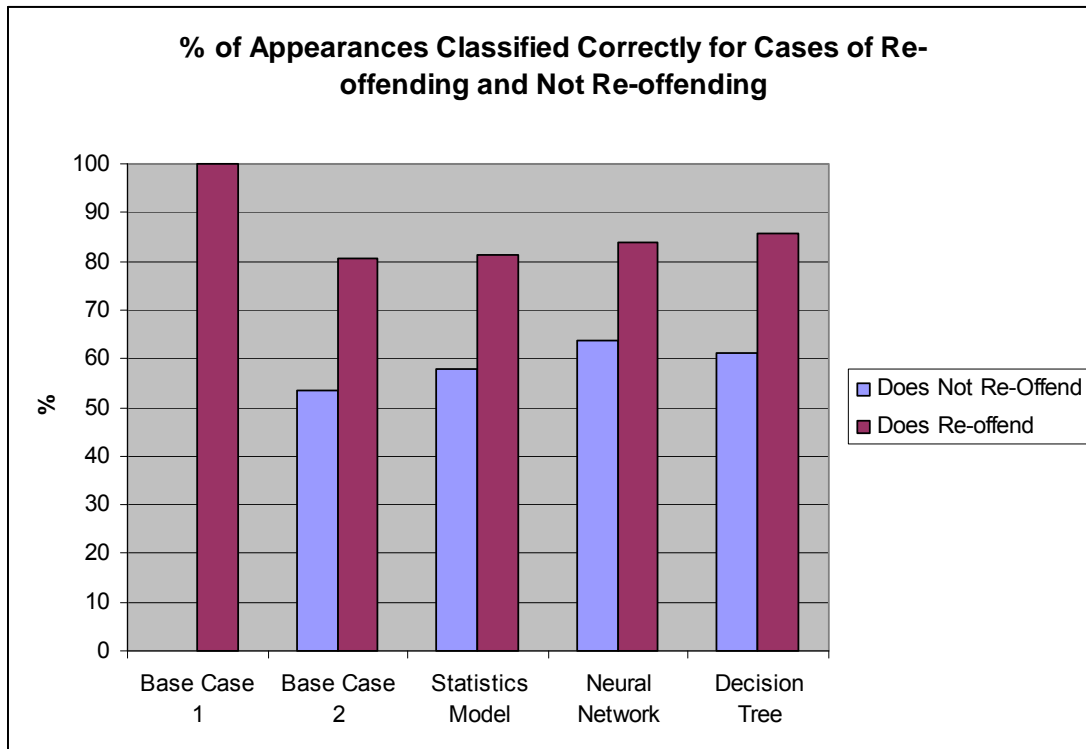


Figure 1: Classification performance broken down according to whether a re-offence is observed

3.2. Model 2: Predictive Models of the Most Serious Offence Type for the Next Appearance for Cases that Resulted in a Subsequent Juvenile Re-Appearance

3.2.1. Aim

This model is applied to those cases that did result in a subsequent guilty appearance as a juvenile. The model aims to predict the most serious offence type heard at this subsequent appearance.

3.2.2. Base Cases

Three base cases were adopted for this model. The first base case is the simple case of predicting the most commonly occurring offence type. The probability distribution of offence types observed in the validation set were as follows:

Offence Category	n	%
Offences against the person	1,869	9.17
Break and enter, burglary	3,333	16.35
Theft and related offences	6,224	30.52
Drug offences	1,274	6.25
Traffic offences	1,729	8.48
Public order offences	1,740	8.53
Property damage	1,283	6.29
Other offences	2,932	14.38

Theft and related offences were the most common offence category, so the first base case was to simply predict this category, resulting in a 29.81% predictive accuracy. The reason this accuracy is slightly different than the 30.52% is that the data representing current offences and the data representing next offences are not exactly the same (due to the sub-setting of data according to date range in the pre-processing stage). As described for model 1, this case is considered to be a case that is constructed with no real analysis, and as such provides a floor on the accuracy we hope to achieve with other models.

The second base case was used to explore the possibility that past offences were the best predictor of future offences; this model assumes that the offence type of the current appearance will be the offence type of the next appearance. Under this model, predictive accuracy was measured at 29.5%. As this is less than the first simplistic base case, there is no need to consider the predictive accuracy of the second base case any further.

The final base case adopted a similar concept to model one where it disaggregated the data into a number of categories according to gender, Indigenous status, age group and appearance number and predicted the most common offence type for the demographic category of the offender. This data is represented in Table 4. The highlighted cells represent the most frequent offence type for the category. As can be seen, the model is similar to the first base case with theft and related offences being the most common category for most demographic categories. However, break and enter, burglary is the most common offence type for a few categories. The predictive accuracy of this base case is 30.3%, showing only a marginal improvement over the first base case.

Table 4: Base case for re-offending, probability of most common offence type disaggregated by demographics

Indigenous Status / Gender	Age Group	Appearance Number	Offences against the person		Break and enter, burglary		Theft and related offences	
			n	%	n	%	n	%
Non Indigenous Male	10 to 14	1	115	10.03	195	17.00	461	40.19
		2	33	6.85	97	20.12	170	35.27
		3	18	6.55	55	20.00	105	38.18
		4	8	4.47	34	18.99	68	37.99
		5+	37	10.76	74	21.51	135	39.24
	15+	1	281	9.33	296	9.82	757	25.12
		2	99	7.23	170	12.41	377	27.52
		3	64	7.79	91	11.07	269	32.73
		4	39	6.93	72	12.79	166	29.48
		5+	136	7.60	267	14.92	559	31.25
Indigenous Male	10 to 14	1	42	6.19	205	30.24	216	31.86
		2	33	7.93	136	32.69	122	29.33
		3	21	6.67	96	30.48	97	30.79
		4	16	6.43	80	32.13	79	31.73
		5+	69	8.79	234	29.81	245	31.21
	15+	1	65	9.39	136	19.65	157	22.69
		2	51	9.98	125	24.46	112	21.92
		3	36	8.78	102	24.88	91	22.20
		4	32	9.36	91	26.61	82	23.98
		5+	189	10.94	390	22.57	387	22.40
Non Indigenous Female	10 to 14	1	38	8.46	25	5.57	232	51.67
		2	21	17.50	7	5.83	51	42.50
		3	6	11.54	3	5.77	28	53.85
		4	6	22.22	2	7.41	16	59.26
		5+	7	14.89	2	4.26	19	40.43
	15+	1	87	9.49	46	5.02	301	32.82
		2	32	9.91	22	6.81	123	38.08
		3	18	9.94	9	4.97	72	39.78
		4	11	8.80	9	7.20	51	40.80
		5+	23	11.17	13	6.31	74	35.92
Indigenous Female	10 to 14	1	21	8.50	46	18.62	94	38.06
		2	16	10.96	26	17.81	59	40.41
		3	13	13.13	20	20.20	33	33.33
		4	9	13.04	15	21.74	32	46.38
		5+	23	16.79	25	18.25	46	33.58
	15+	1	31	10.47	23	7.77	84	28.38
		2	24	12.83	16	8.56	56	29.95
		3	29	22.66	12	9.38	36	28.13
		4	13	12.04	13	12.04	34	31.48
		5+	57	13.90	53	12.93	128	31.22

Table continued over page

Indigenous Status / Gender	Age Group	Appearance Number	Drug offences		Traffic offences		Public order offences	
			n	%	n	%	n	%
Non Indigenous Male	10 to 14	1	55	4.80	60	5.23	64	5.58
		2	33	6.85	22	4.56	29	6.02
		3	18	6.55	11	4.00	15	5.45
		4	12	6.70	9	5.03	15	8.38
		5+	13	3.78	16	4.65	19	5.52
	15+	1	260	8.63	507	16.83	292	9.69
		2	145	10.58	184	13.43	121	8.83
		3	78	9.49	89	10.83	73	8.88
		4	59	10.48	59	10.48	41	7.28
		5+	178	9.95	177	9.89	138	7.71
Indigenous Male	10 to 14	1	6	0.88	33	4.87	46	6.78
		2	5	1.20	10	2.40	33	7.93
		3	8	2.54	7	2.22	30	9.52
		4	8	3.21	6	2.41	12	4.82
		5+	15	1.91	21	2.68	47	5.99
	15+	1	32	4.62	72	10.40	70	10.12
		2	26	5.09	45	8.81	57	11.15
		3	22	5.37	33	8.05	37	9.02
		4	16	4.68	26	7.60	30	8.77
		5+	83	4.80	115	6.66	160	9.26
Non Indigenous Female	10 to 14	1	18	4.01	12	2.67	33	7.35
		2	3	2.50	2	1.67	6	5.00
		3			2	3.85	3	5.77
		4						
		5+	1	2.13	1	2.13	2	4.26
	15+	1	65	7.09	110	12.00	111	12.10
		2	25	7.74	15	4.64	31	9.60
		3	15	8.29	7	3.87	23	12.71
		4	12	9.60	7	5.60	12	9.60
		5+	17	8.25	5	2.43	16	7.77
Indigenous Female	10 to 14	1	4	1.62	4	1.62	18	7.29
		2	1	0.68	3	2.05	7	4.79
		3	2	2.02	1	1.01	9	9.09
		4					5	7.25
		5+	4	2.92	2	1.46	12	8.76
	15+	1	13	4.39	27	9.12	36	12.16
		2	6	3.21	11	5.88	18	9.63
		3	3	2.34	2	1.56	12	9.38
		4	5	4.63	4	3.70	17	15.74
		5+	8	1.95	12	2.93	40	9.76

Table continued over page

Indigenous Status/Gender	Age Group	Appearance Number	Property damage		Other offences	
			Count	Row %	Count	Row %
Non Indigenous Male	10 to 14	1	102	8.89	95	8.28
		2	59	12.24	39	8.09
		3	28	10.18	25	9.09
		4	21	11.73	12	6.70
		5+	25	7.27	25	7.27
	15+	1	158	5.24	462	15.33
		2	91	6.64	183	13.36
		3	40	4.87	118	14.36
		4	42	7.46	85	15.10
		5+	112	6.26	222	12.41
Indigenous Male	10 to 14	1	63	9.29	67	9.88
		2	33	7.93	44	10.58
		3	31	9.84	25	7.94
		4	15	6.02	33	13.25
		5+	52	6.62	102	12.99
	15+	1	46	6.65	114	16.47
		2	16	3.13	79	15.46
		3	23	5.61	66	16.10
		4	19	5.56	46	13.45
		5+	94	5.44	310	17.94
Non Indigenous Female	10 to 14	1	31	6.90	60	13.36
		2	10	8.33	20	16.67
		3	1	1.92	9	17.31
		4	2	7.41	1	3.70
		5+	7	14.89	8	17.02
	15+	1	30	3.27	167	18.21
		2	15	4.64	60	18.58
		3	7	3.87	30	16.57
		4	5	4.00	18	14.40
		5+	10	4.85	48	23.30
Indigenous Female	10 to 14	1	24	9.72	36	14.57
		2	12	8.22	22	15.07
		3	8	8.08	13	13.13
		4	1	1.45	7	10.14
		5+	5	3.65	20	14.60
	15+	1	7	2.36	75	25.34
		2	13	6.95	43	22.99
		3	6	4.69	28	21.88
		4	2	1.85	20	18.52
		5+	17	4.15	95	23.17

3.2.3. Statistical Model

The statistical model developed for comparative purposes was based on a multi-nomial regression (Appendix 1). The predictive accuracy of this model was 33.2% on the validation set showing approximately a 9.58% increase in base case accuracy. That is, the increase in accuracy was approximately 9.58% of the accuracy observed by the best base case.

3.2.4. Neural Network Model

The independent variables adopted in the neural network for this model included all variables adopted for the neural network model in model 1, as well as the variable `MostSeriousCourtOutcomeToDate` detailed in the decision tree model for model 1. As well as these variables, additional variables were created to more explicitly represent the distribution of past offence types for the offender. As the model is attempting to predict the next offence type, a representation of the probability distribution of current/past offence types for an offender was added as this would provide useful information for the model that may allow improved predictive accuracy.

To represent this distribution, eight variables were created representing the eight different offence types. Each of these variables then adopted the value that is the proportion of current/past offences for the individual that were of this offence type. For example, if an offender was currently at their fourth court appearance and, say, the first two offences were theft and related offences followed by a drug offence and then followed by a property damage offence, the variable representing theft and related offences would adopt the value 0.5 (representing 50% of offences) while the variables representing drug offences and property damage offences would each adopt the value 0.25 (25%). All variables representing the other offence types would adopt the value 0. These variables should provide quite rich information relating to the offending history.

The neural network was estimated with 13 hidden units and was provided a 34.01% classification accuracy on the validation set, showing a 2.44% improvement in accuracy over the multi-nomial regression.

3.2.5. *Decision Tree Model*

The decision tree model was estimated with the same set of independent variables as the neural network model. The resulting model showed a classification accuracy of 33.69%, thus lying between the accuracy of the multi-nomial regression and the neural network. The structure of the tree formed can be seen in Appendix 2. As can be seen, the model relies heavily on the input representation of the distribution of past offence types and the current offence type as well as incorporating information about the offenders' demographic characteristics (Indigenous status and gender) as well as their age. It is interesting to note here that the model chooses to focus on the distribution of past offence types and saw no role to incorporate appearance number information into the model.

3.2.6. *Model Performance*

Table 5 presents the models' classification accuracy on the validation set. The overall classification accuracy of all models is quite low. This simply implies that there may not be the information content in the dataset to provide a strong prediction of the next offence type, or that the next offence is inherently quite random and, as such, it may only be possible to achieve a low level of accuracy.

Table 5: Validation set performance for models predicting most serious offence type

Model	Validation Set Classification Accuracy
Base 1: Predict Most Common Offence	29.81%
Base 2: Predict Current Offence	29.50%
Base 3: Most Common Offence by Demographic	30.30%
Stats Prediction: Multi-Nomial Regression	33.20%
Neural Network	34.01%
Decision Tree	33.69%

The actual level of predictive accuracy is not of great importance for this study as the modelling methodology of constructing base cases and statistical models from which to compare the performance of the data mining models allows comparisons between the data mining accuracy in comparison to more commonly used methods. If the accuracy of the

neural network (the highest performing model) is compared with that of the first base case (indicating a floor on the level of accuracy that is potentially achievable), there was a 14.06% gain in accuracy, in that the increase in accuracy observed is 14.06% of the original accuracy of the first base case. While the strongest performing model here is the neural network, the decision tree would be the preferable model as it demonstrates an increase in accuracy while producing a transparent model at the same time, allowing an analyst to understand exactly how a prediction is arrived at.

Figure 2 breaks down this classification accuracy according to the type of next offence for the mult-nomial regression, the neural network and the decision tree. As displayed, all of these models are achieving comparatively high classification accuracy on theft and related offences, with the break and enter displaying the next highest level of accuracy. These are the most commonly occurring offence types as shown above, indicating that classification accuracy is easier to achieve on the more prevalent offence types. This would potentially be due to the models placing a higher degree of predictions according to the more prevalent offence types. This is confirmed in Figure 3 below, and would be due to a greater probability of accuracy in predicting the more prevalent offence types, while requiring a degree of supportive evidence to place a prediction in a less frequent offence type.

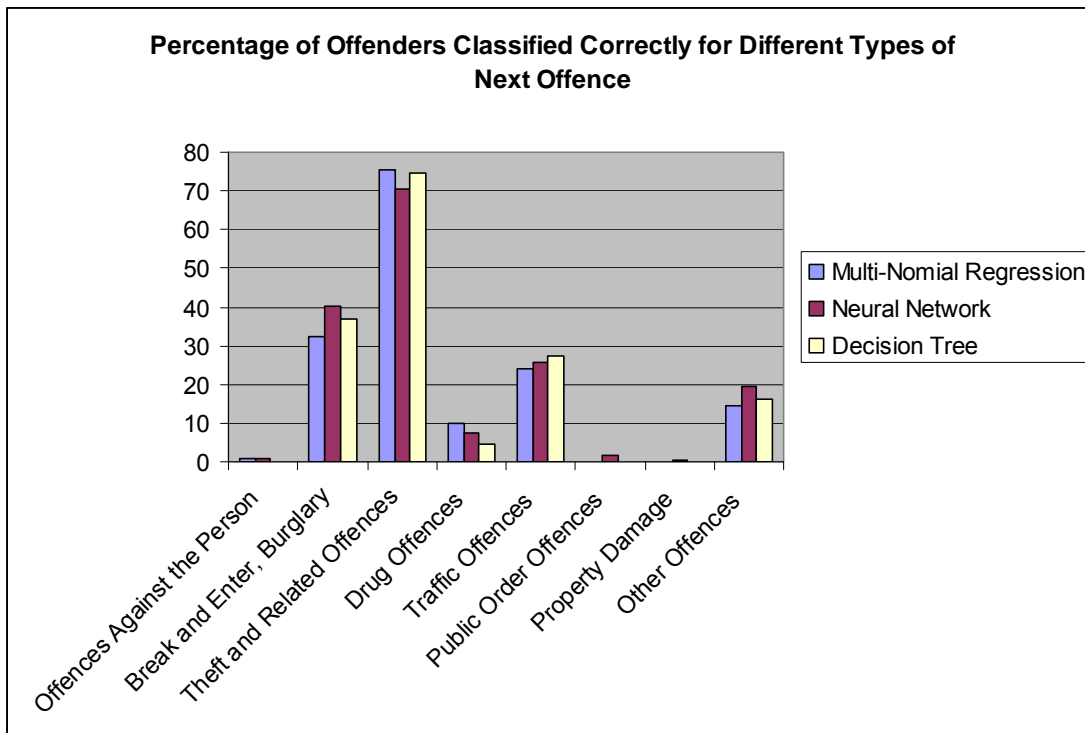


Figure 2: The percentage of offenders classified correctly in the validation set for the multi-nomial regression, the neural network and the decision tree

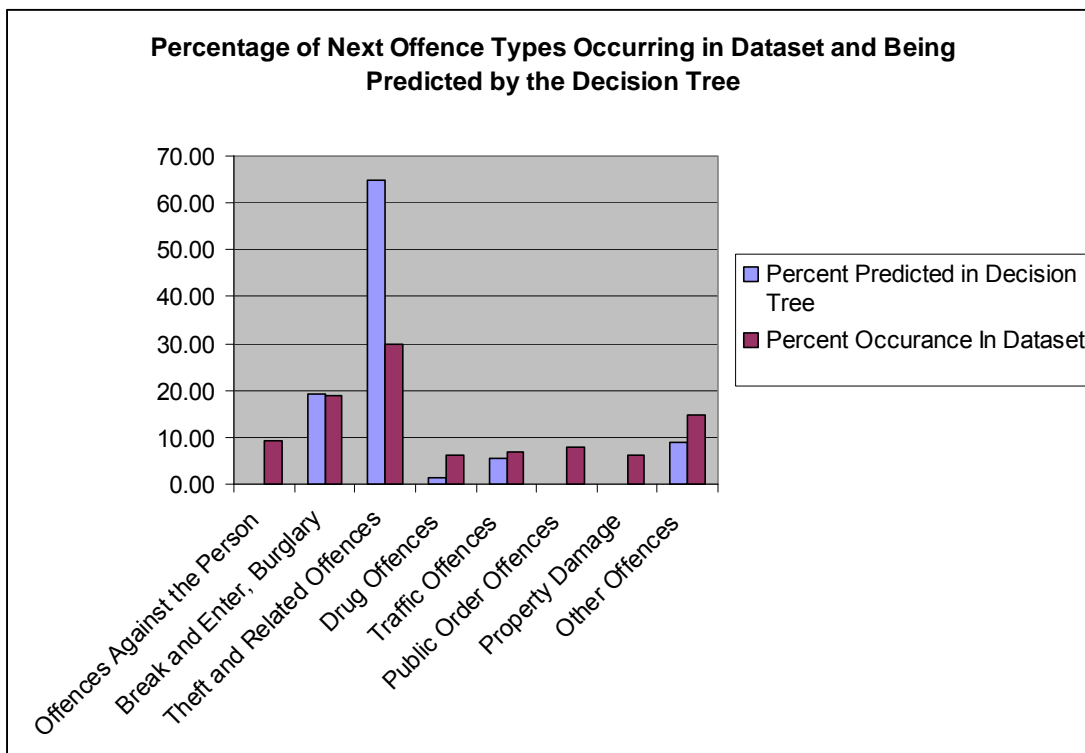


Figure 3: Percentage of next offence types occurring in dataset and being predicted by the decision tree

3.3. Model 3: Predicting Months to Re-Offence for those Cases that Result in a Subsequent Guilty Juvenile Appearance

3.3.1. Aim

This model is applied to those cases that did result in a subsequent guilty appearance as a juvenile. The model aims to predict how long (in months) until the subsequent guilty appearance occurs. This enables modelling of a numeric dependent variable in the study. The above models are based on a categorical dependent variable, and have indeed shown an improvement of predictive accuracy for the data mining models. However, it is possible that the mechanics of formulation of a model with a numeric prediction may be different and as such need to be assessed in a separate model. That is one of the primary purposes of this model for our study.

3.3.2. Dependent Variable

Figure 4 presents a histogram of months to reappearance on the validation set. As can be seen, the distribution of this variable is highly skewed and as such may present problem for the modelling processes.

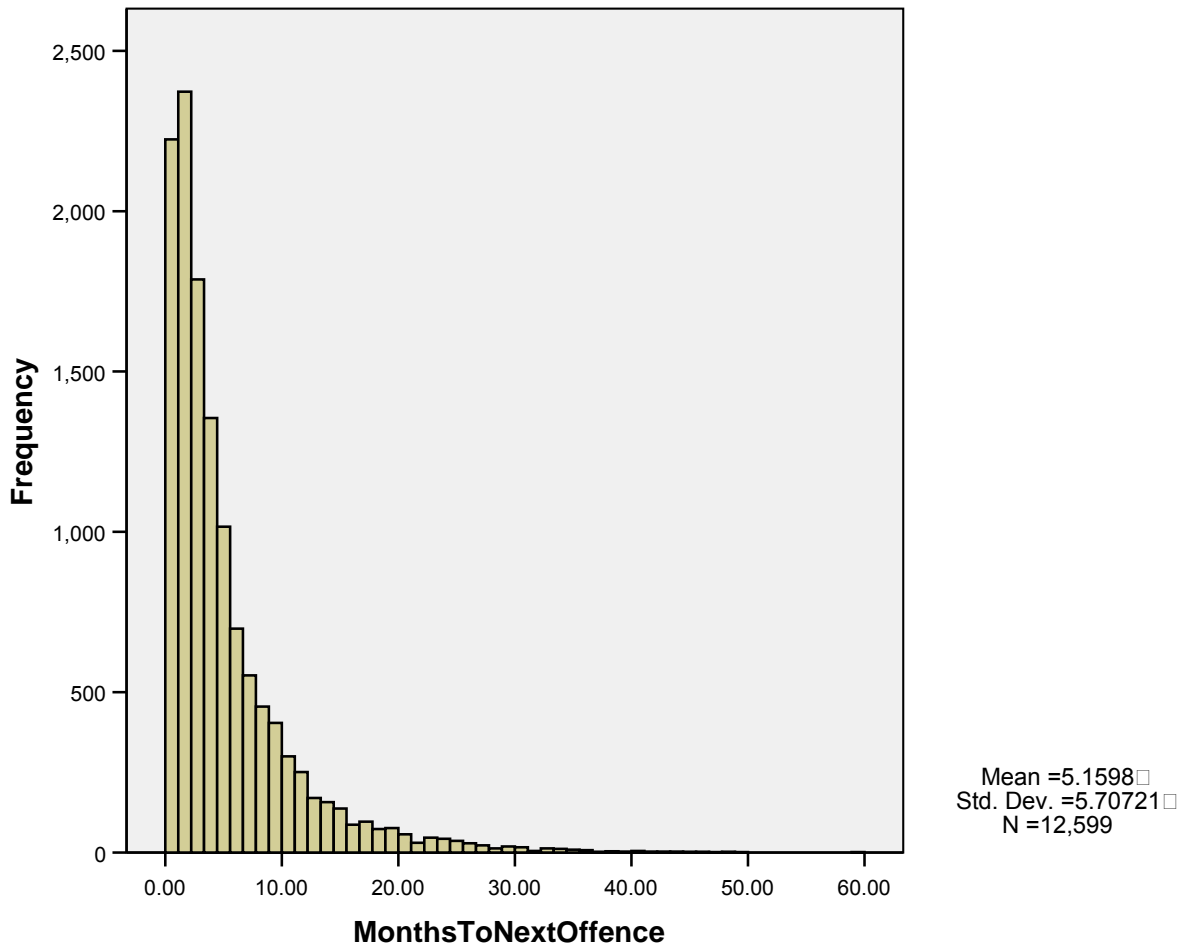


Figure 4: Histogram of months to next offence

To solve this, a log of this variable was performed. As can be seen in Figure 5, the distribution has become more symmetrical by applying the log, and while not quite normally distributed, should be more reasonable to use as a dependent variable.

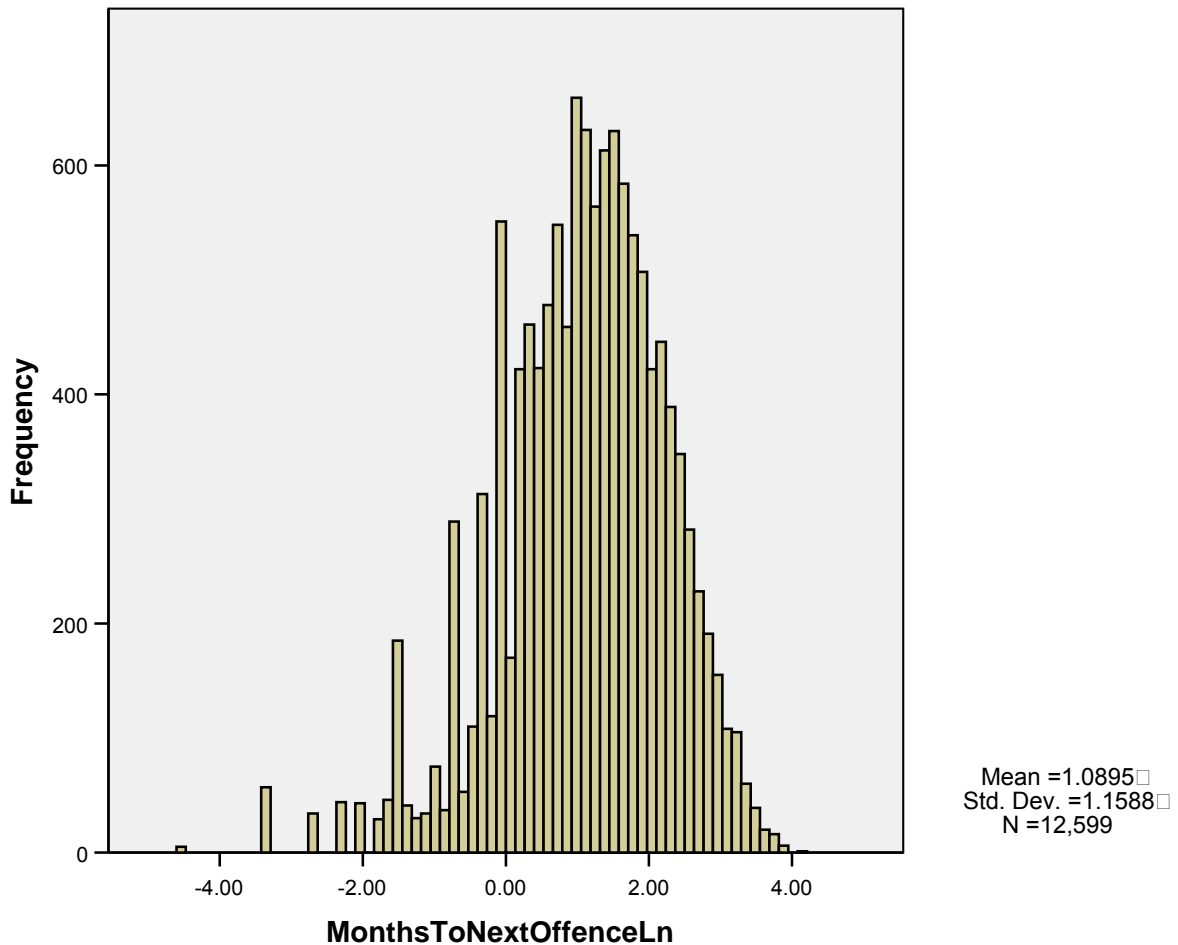


Figure 5: Histogram of the log of months to next offence

The log of the number of months to the next offence was therefore adopted as the dependent variable for the model. Note that in this histogram there were a number of negative outlier points with a value less than negative 4. Most traditional statistical methods would require the removal of these outliers in order not to violate the assumptions of the model. For both the base cases and the data mining methods, these points have been included in the modelling. Data mining methods should be more robust to forming models with outliers present. Additionally, this touches on a very important methodological difference of this approach that is of benefit. While many traditional analysis methods require the removal of outliers, this process is often removing the very information in the dataset that is of most interest. In the cases here, the outliers are offenders who have almost an immediate reappearance. To remove these cases may indeed remove the cases that are of most interest in such a study. An inability to properly study outliers is a major deficiency of many more traditional statistical methods.

3.3.3. Base Cases

Three base cases were adopted for this model. The first base case simply adopted the average of the dependent variable on the validation set as its prediction. The mean log of months to re-offence was 1.09 with a standard deviation of 1.16.

For the second base case, the ‘edge effect’ needed to be considered and taken into account. As all appearances at which the age of the offender was greater than 17 years and one month were subset, this provides a limit on the possible time to re-offence for an offender. For example, if the offender’s current offence is right on their 17th birthday, this means their next offence will have to occur within one month as registered within the data. Additionally, even if an offender is young, as our data ended in June 2007, if a subsequent offence occurred, it must have been before this date. These two factors enable computation of the maximum number of months that the re-offence must have occurred within, which may be termed the months to the edge effect. Therefore, the second base case simply involves taking a log of the months to the edge effect and assuming the prediction is either 1.09 months (derived from base case 1) or the log months to the edge effect (whichever is smaller).

The third base case computes separate average log months to re-offence for each demographic category. These averages are presented in Table 6. The third base case assumed that the time to re-offence was the minimum of the average months to re-offence for that offender’s demographic category or the log months to the edge effect for that offender.

Table 6: Base case for re-offending, predicting months until re-offence disaggregated by demographics

Indigenous Status / Gender	Age Group	Appearance Number	Mean Log Months Till Re-offence
Non Indigenous Male	10 to 14	1	1.81
		2	1.45
		3	1.18
		4	1.05
		5+	0.97
	15+	1	1.22
		2	0.97
		3	0.94
		4	0.87
		5+	0.73
Indigenous Male	10 to 14	1	1.57
		2	1.23
		3	1.12
		4	1.15
		5+	0.91
	15+	1	1.27
		2	1.07
		3	1.08
		4	0.96
		5+	0.78
Non Indigenous Female	10 to 14	1	1.76
		2	1.21
		3	1.18
		4	0.67
		5+	0.73
	15+	1	1.13
		2	0.97
		3	1.09
		4	0.9
		5+	0.77
Indigenous Female	10 to 14	1	1.43
		2	1.15
		3	1.1
		4	1
		5+	1.02
	15+	1	1.18
		2	1.12
		3	1.04
		4	0.89
		5+	0.76

3.3.4. Statistics Model

For the statistics model, a linear regression was adopted. Details of this linear regression can be found in Appendix 1. The model that was developed accounted for 6.4% of the variance in the prediction of time to re-offence.

3.3.5. Neural Network

The neural network was used to estimate the data with 13 hidden units. Additionally, the architecture of the network was different from the previous two models as in these models the output nodes of the neural network adopted a logistic activation function. This will result in values being returned from the network that are between zero and one. This is a desirable property when a network is being used for categorization; however, many of the values being predicted here do not fit in that range. To address this, the output node adopted a linear activation function, thus allowing a full range of values to be produced.

The independent variables adopted included the following variables as previously defined:

IndigSex
OffType
Crtorder
MostSeriousCourtOutcomeToDate
AppNumMaxFactor
AgeConinuousAtAppearance
AgeContinousAtFirstAppearance

In addition to these variables, the following independent variables were also used:

MonthsTillEdgeEffectLn: The log of the number of months to the edge effect for the offender, as previously described.

DetentDurationMonths: Number of months of detention ordered in the current court appearance multiplied by 2/3. This variable is included as any detention served will be a period the offender is not able to commit another offence, and as such may extend the time to

re-offence. The sentence is multiplied by $2/3$ as most detention sentences are not served to their full sentence and $2/3$ is an estimate of the average portion of a sentence served.

3.3.6. Decision Tree

The decision tree was estimated with the same independent variables as adopted above. The structure of this tree is presented in Appendix 2. It is apparent that the variables adopted to form its decisions include age at appearance, the log of the months to the edge effect and the appearance number. It is interesting to note the model did not determine that Indigenous status or gender were significant enough to include, but based its decisions on age (months to edge effect) and appearance number.

3.3.7. Model Performance

The residuals of the predictions on the validation set were computed to compare the results of the different methods. The residual is calculated as the observed log of months to the next offence minus the predicted log of the months to the next offence. It should be observed that these residuals should have a mean that is close to zero. Assuming the mean is sufficiently close to zero, the more accurate models should have a lower standard deviation of residuals, indicating the predictions are not deviating around the observed values to a large degree.

The observed mean and standard deviation of residuals for the various models are presented in Table 7. It is apparent that the residuals are sufficiently close to zero. Base case 3 displayed better accuracy than the other base cases; however, the decision tree did slightly outperform the third base case. The neural network and linear regression were not able to outperform the third basic base case indicating they were not able to form models that could effectively generalise to the validation dataset.

Table 7: Validation set performance for models predicting time to re-offence

	Mean of residual	Standard deviation of residual
Base 1: Mean of data is prediction	0.000	1.16
Base 2: Minimum of base case 1 of the months till the edge effect	0.009	1.15
Base 3: Base case by demographic	0.006	1.11
Linear Regression	-0.026	1.12
Neural Network	0.014	1.17
Regression Tree	0.005	1.09

Further experimentation was performed to determine if additional accuracy could be formed with the decision tree by focusing only on court appearances where the offender had a prior appearance and including independent variables representing the time to the previous offence and the average time between offences. It was speculated that these variables may add information that would be highly related to predicting time to next offence; however no improvement of accuracy was found. Therefore, the best model was found to be the decision tree, which would be the preferred model because of its performance on the validation set and because of the transparency of the model produced.

Chapter 4. Discussion

This project applied neural networks and decision trees to criminal justice data to determine whether these techniques could be used to improve the predictive accuracy of models developed to predict risk of juvenile re-offending over base cases and commonly applied statistical methods. The predictive accuracy of these techniques was assessed for (i) prediction of recontact, (ii) prediction of offence type, and (iii) prediction of time to recontact.

Findings indicated that the decision trees had more predictive accuracy than either the base cases adopted or the models developed using traditional statistical methods (Table 8). Additionally, the neural networks outperformed the base cases and traditional statistical models for both classification tasks (recontact and offence type), however was unable to form a model with reasonable predictive capability for the model predicting time to recontact. As decision trees have consistently resulted in more accurate predictions and given the transparent nature of the outcomes produced by this technique, it has the most potential of the techniques applied.

Table 8: Accuracy of alternative methodologies for predicting re-offending in the juvenile justice system

Models	Predictive Accuracy of Cases Correctly Classified			
	Best Base Case	Statistical Model	Neural Network	Decision Tree
Recontact	67.51	72.40	76.35	76.45
Next offence type	30.30	33.20	34.01	33.69
	Standard Deviation of Residual			
Time to re-offence	1.11	1.12	1.17	1.09

The findings indicate that KDD techniques can be applied to improve predictive accuracy. There is considerable potential for greater application of these techniques to predict risk of re-offending as any improvement in prediction accuracy will result in more effective criminal justice system decision-making. The potential implications of improved accuracy are considerable given the wide ranging criminal justice system processes that require an assessment about risk of re-offending and the important role that such assessments have for public safety and the targeting of offender rehabilitation programs.

Future research could consider applying data mining-survival analysis techniques. *Survival analysis* is suited to model time to probabilistic events and may be used to develop risk assessment models of recidivism. As the technique provides a distribution of time to re-offence, there is more information than a simple prediction of re-offence or a classification of an individual as high, moderate, or low risk. Additionally, there has been some experimental research in integrating survival analysis with decision trees and neural networks. More accurate predictive models of re-offending could be obtained if the improvements obtained in the current project were to occur in survival analysis. These techniques should be considered of a more experimental nature than the established methods that have been adopted in the current project and as such this research would be considered more exploratory and would require the work here to justify its movement.

References

- Adderly, R., & Musgrove, P. B. (2001). *Data Mining Case Study: Modelling the Behaviour of Offenders Who Commit Serious Sexual Assaults*. Paper presented at the Conference on Knowledge Discovery in Data. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Databases*.
- Ægisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, 34(3), 341-382.
- Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, 52(1), 7-27.
- Baker, K., Jones, S., Roberts, C., & Merrington, S. (2003). *The Evaluation of the Validity and Reliability of the Youth Justice Board's Assessment for Young Offenders: Findings for the First Two Years of Use of the ASSET*. Oxford: Centre for Criminological Research, University of Oxford & Youth Justice Board for England and Wales.
- Bonta, J. (1996). Risk-Needs Assessment and Treatment. In A. T. Harland (Ed.), *Choosing Correctional Options That Work: Defining the Demand and Evaluating the Supply* (pp. 18-32). Thousand Oaks: SAGE Publications.
- Bonta, J. (2002). Offender Risk Assessment: Guidelines for Selection and Use. *Criminal Justice and Behavior*, 29(4), 355-379.
- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth.
- Brennan, T., Wells, D., & Alexander, J. (2004). *Enhancing Prison Classification Systems: The Emerging Role of Management Information Systems*. Washington, DC: U.S. Department of Justice, National Institute of Corrections.
- Carr, M. B., & Vandiver, T. A. (2001). Risk and protective factors among youth offenders. *Adolescence*, 36(143), 409-426.
- Casey, S., & Day, A. (2004). *Development of Juvenile Justice Risk Needs Tool*. Victoria: Victorian Department of Human Services.

- Cash, S. (2001). Risk Assessment in Child Welfare: The Art and Science. *Children and Youth Services Review*, 23, 811-830.
- Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting Criminal Recidivism: A Comparison of Neural Network Models with Statistical Methods. *Journal of Criminal Justice*, 24(3), 227-240.
- Cicchinelli, I. F. (1995). Risk Assessment Expectations and Realities. *The APSAC Advisor*, 8, 3-8.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, 243, 1668-1674.
- DeMatteo, D., & Marczyk, G. (2005). Risk Factors, Protective Factors, and the Prevention of Antisocial Behavior Among Juveniles. In K. Heilbrun, N. E. S. Goldstein & R. E. Redding (Eds.), *Juvenile Delinquency: Prevention, Assessment, and Intervention* (pp. 19-44). New York: Oxford University Press.
- Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence Risk Assessment: Science and Practice. *Legal and Criminological Psychology*, 4, 149-184.
- Estivill-Castro, V., & Lee, I. (2001). *Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-Referenced Crime Data*. Paper presented at the Proceedings of the Sixth International Conference on Geocomputation, Brisbane, Australia,
- Farrington, D. P. (2002). Developmental Criminology and Risk-Focused Prevention. In M. Maguire, R. Morgan & R. Reiner (Eds.), *The Oxford Handbook of Criminology* (3rd ed., pp. 657-701). Oxford: Oxford University Press.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996). *Advances in Knowledge Discovery and Data Mining*: AAAI/MIT Press.
- Gambrill, E., & Shlonsky, A. (2000). Risk Assessment in Context. *Children and Youth Services Review*, 22(11/12), 813-837.
- Gottfredson, D. M. (1987). Prediction and Classification in Criminal Justice Decision Making. *Crime and Justice*, 9, 1-20.
- Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical Risk Assessment: Old Problems and New Applications. *Crime & Delinquency*, 52(1), 178-200.
- Gray, B., & Orłowska, M.E. (1998). *The Use of Clustering to Mine Interesting Association Rules*. Technical Report TR425. Queensland: School of Computer Science and Electrical Engineering, University of Queensland.

- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment, 12*(1), 19-30.
- Hanson, R. K. (2005). Twenty Years of Progress in Violence Risk Assessment. *Journal of Interpersonal Violence, 20*(2), 212-217.
- Hoge, R. D. (2002). Standardized Instruments for Assessing Risk and Need in Youthful Offenders. *Criminal Justice and Behavior, 29*(4), 380-396.
- Lin, S., & Brown, D. E. (2006). An Outlier-Based Data Association Method for Linking Criminal Incidents. *Decision Support Systems, 41*, 604-615.
- McClelland, J., & Rumelhart, D. (1988). *Explorations in Parallel Distributed Processing*. Cambridge, MA: MIT Press.
- Monahan, J., Steadman, H. J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Silver, E., et al. (2000). Developing a Clinically Useful Actuarial Tool for Assessing Violence Risk. *British Journal of Psychiatry, 176*, 312-319.
- Morrison, G. M., Robertson, L., Laurie, B., & Kelly, J. (2002). Protective Factors Related to Antisocial Behavior Trajectories. *Journal of Clinical Psychology 58*(3), 277-290.
- Paik, H. (2000). Comments on Neural Networks. *Sociological Methods & Research, 28*(4), 425-453.
- Palocsay, S. W., Wang, P., & Brookshire, R. G. (2000). Predicting Criminal Recidivism Using Neural Networks. *Socio-Economic Planning Sciences, 34*, 271-284.
- Quinlan, J. R., (1986). Induction of decision trees. *Machine Learning, 1*, 81-106.
- Quinlan, J. R., (1987). Simplifying decision trees. *International Journal of Man-Machine Studies, 27*(3), 221-234.
- Rogers, R. (2000). The Uncritical Acceptance of Risk Assessment in Forensic Practice. *Law and Human Behaviour, 24*(5), 595-605.
- Rosenfeld, B., & Lewis, C. (2005). Assessing Violence Risk in Stalking Cases: A Regression Tree Approach. *Law and Human Behaviour, 29*(3), 343-357.
- Seifert, J. W. (2004). *Data Mining: An Overview* (CRS Report for Congress). Washington, DC: Congressional Research Service.
- Silver, E., & Chow-Martin, L. (2002). A Multiple Models Approach to Assessing Recidivism Risk: Implications for Judicial Decision Making. *Criminal Justice and Behavior, 29*(5), 538-568.
- Silver, E., & Miller, L. L. (2002). A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control. *Crime & Delinquency, 48*(1), 138-161.

- Stalans, L. J., Yarnold, P. R., Seng, M., Olson, D. E., & Repp, M. (2004). Identifying Three Types of Violent Offenders and Predicting Violent Recidivism While on Probation: A Classification Tree Analysis. *Law and Human Behaviour*, 28(3), 253-271.
- Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., et al. (2000). A Classification Tree Approach to the Development of Actuarial Violence Risk Assessment Tools. *Law and Human Behaviour*, 24(1), 83-100.
- Stewart, A., Spencer, N. O'Connor, I. Palk, G. Livingston, M., & Allard, T. (2004). Juvenile Justice Simulation Model: A Technical Report. Justice Modelling at Griffith.
- Strano, M. (2004). A Neural Network Applied to Criminal Psychological Profiling: An Italian Initiative. *International Journal of Offender Therapy and Comparative Criminology*, 48(4), 495-503.
- Taxman, F. S., Cropsey, K. L., Young, D. W., & Wexler, H. (2007). Screening, Assessment, and Referral Practices in Adult Correctional Settings. *Criminal Justice and Behavior*, 34(9), 1216-1234.
- Taxman, F. S., & Thanner, M. (2006). Risk, Need, and Responsivity (RNR): It All Depends. *Crime & Delinquency*, 52(1), 28-51.
- Thompson, A. P., & Pope, Z. (2005). Assessing Juvenile Offenders: Preliminary Data for the Australian Adaptation of the Youth Level of Service/Case Management Inventory. *Australian Psychologist*, 40(3), 207-214.
- Visher, C. A., Lattimore, P. K., & Linster, R. L. (1991). Predicting the Recidivism of Serious Youthful Offenders Using Survival Models. *Criminology*, 29(3), 329-366.
- Zhang, Z., Salerno, J. J., & Yu, P. S. (2003). *Applying Data Mining in Investigating Money Laundering Crimes*. Paper presented at the Conference on Knowledge Discovery in Data. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 747-752.

Appendix 1: Statistical Analysis Models Developed for the Purpose of Comparison to Data Mining Techniques

There were a total of 24,794 court appearances in the estimation dataset relating to 9,936 unique offenders. In terms of the base-rate of reoffending, 9,311 (37.55%) court appearances were not associated with a subsequent reoffence, while 15,483 (62.45%) of court appearances were associated with a subsequent reoffence. In the validation dataset there were a total of 20,391 court appearances relating to 8,265 unique offenders. The base-rate for reoffending was 61.9 percent ($N = 12,620$), and 38.1 percent ($N = 7,771$) for non-reoffending.

Model 1: Predicting Reoffending

A binary logistic regression analysis was performed in order to predict the likelihood of reoffending for each unique court appearance. The variables included in the analysis were:

1. Reoffend (Dependent Variable)– binary variable where: 0 = no reoffence; and 1 = reoffence.
2. Gender – binary variable where: 0 = female; and 1 = male
3. Indigenous status – binary variable where: 0 = non-Indigenous; and 1 = Indigenous
4. Age – continuous variable representing the age in years that an individual appeared in court.
5. Appearance number – continuous variable representing the total number of court appearances for each individual up to and including the current appearance
6. Offence type – binary variable where: 0 = all other offences; and 1 = offence against a person.

Descriptives

Nonparametric bivariate correlations among the variables were obtained prior to conducting the main analysis (Table 1).

Table 1: Nonparametric bivariate correlations among the dependent and independent variables entered into the model predicting re-offending (N = 24794)

Variable	1	2	3	4	5	6
1. Reoffend						
2. Gender	-.07**					
3. Indigenous status	.23**	.04**				
4. Age at appearance	-.39**	-.02**	-.18**			
5. Appearance number	.28**	-.11**	.31**	.09**		
6. Offence type	-.05**	.03**	.01*	.01*	-.02**	

* $p < .05$, ** $p < .01$

All independent variables were significantly associated with reoffending in the expected directions, except for offence type, which was significantly negatively associated with a subsequent reoffence. This indicated that more serious offences against persons were associated with a lowered likelihood of reoffending. This association was most likely a result of an imprisonment effect, where more serious offences are likely to result in harsher court outcomes (i.e., detention) that remove an individuals' opportunities to reoffend. The largest correlation ($\rho = -.39$) was between age at appearance and reoffending, indicating that offenders who appeared in court at a younger age were more likely to reoffend.

Results

A binary logistic regression analysis with the binary DV of reoffending was performed, with the variables of gender Indigenous status, age at appearance, appearance number, and offence type entered as predictors. Table 2 displays the regression coefficients (B), the regression coefficients' standard errors, Wald statistics, degrees of freedom, odds ratios, and 95% odds ratio confidence intervals for the predictors in the model.

Table 2: Results of the binary logistic regression with re-offending as the dependent variable

Variable	B	SE (B)	Wald χ^2	df	Odds Ratio	95% CI for Odds Ratio	
						Lower	Upper
Gender	.48	.04	162.43†	1	1.62	1.50	1.74
Indigenous status	.55	.03	257.79†	1	1.73	1.62	1.85
Age at appearance	-1.01	.02	3296.68†	1	.37	.35	.38
Appearance number	.20	.01	1167.20†	1	1.22	1.21	1.23
Offence category	-.40	.05	55.29†	1	.67	.67	.75
Constant	14.59	.27	3027.99†	1			

* $p < .05$, ** $p < .01$, *** $p < .001$, † $p < .0005$

Although an initial run of the logistic regression, inspection of standardised errors of prediction indicated that 211 cases were outliers in the solution, exceeding the criterion of $z = \pm 3.3$ (Tabachnick & Fidell, 2007), and were excluded from the analysis. This resulted in $N = 24,583$ court appearances being submitted for analysis ($N = 9,100$ {37%} non reoffend; and $N = 15,483$ {63%} reoffend). The Hosmer-Lemeshow goodness-of-fit test was significant ($\chi^2(8) = 160.13, p = .0005$), indicating that the logistic model provided a poor fit to the data (-2 Log Likelihood = 25273.88). The poor model fit was most likely the result of a significant degree of dependency in the data. Individuals were able to have multiple court appearances in the data (there were a total of $N = 15,483$ reappearances in the data set), which resulted in cases being significantly related to each other since they could share the same characteristics (i.e., be the same unique offender). Therefore, it is likely the model is overfitted to the data, meaning that all results should be interpreted with caution.

The test of the full model containing the eight predictors against a constant-only model was statistically significant, $\chi^2(5, N = 24583) = 7128.90, p < .0005$, indicating that the variables as a set were reliable in distinguishing between court appearances that did and did not result in a reoffence. The variables accounted for a moderate amount of variance in reoffending, with Nagelkerke's $R^2 = .34$, and 95% confidence intervals ranging from .32 to .36. Classification in the model was reasonable, with 60.5 of court appearances not resulting in a subsequent reoffence and 82.2 percent of court appearance resulting in a subsequent reoffence classified correctly, resulting in an overall classification success rate of 74.2 percent (Table 3).

Table 3: Classification Table for Logistic Regression Predicting Reoffending after a Court Appearance.

Observed	Predicted		Percentage Correctly Classified
	Does not Reoffend	Does Reoffend	
Does not Reoffend	5509	3591	60.5
Does Reoffend	2758	12725	82.2
Overall Classification Percentage			74.2

Using the Wald criterion, all variables emerged as significant predictors of reoffending after a court appearance. Individuals with a younger age at appearance were more likely to have a subsequent re-offence. Indigenous youth were approximately 1.7 (70%) times more likely to have a subsequent re-offence compared to non-Indigenous males. Males were approximately 1.6 (60%) times more likely to have a subsequent re-offence compared to non-Indigenous females. Individuals with an offence type of offences against a person were approximately 0.4 (60%) times less likely to have a subsequent re-offence compared to individuals with any other type of offence.

In order to validate the logistic regression model’s classification accuracy, the equation derived from the estimation set was applied to the validation set:

$$\text{Prob(Reoffend)} = \frac{e^{14588 + (.479)(\text{gen}) + (.550)(\text{ind}) + (-1.01)(\text{age}) + (.199)(\text{AppearanceNumber}) + (-.396)(\text{catoff})}}{1 + e^{14588 + (.479)(\text{gen}) + (.550)(\text{ind}) + (-1.01)(\text{age}) + (.199)(\text{AppearanceNumber}) + (-.396)(\text{catoff})}}$$

The application of the estimation dataset logistic regression equation to the validation dataset resulted in a comparative level of classification accuracy. Classification in the model was reasonable, with 57.8 of court appearances not resulting in a subsequent reoffence and 81.4 percent of court appearance resulting in a subsequent reoffence classified correctly, resulting in an overall classification success rate of 72.4 percent (Table 3).

Table 3: Classification table for logistic regression predicting re-offending after a court appearance on validation dataset.

Observed	Predicted		Percentage Correctly Classified
	Does not Reoffend	Does Reoffend	
Does not Reoffend	4492	3279	57.8
Does Reoffend	2349	10271	81.4
Overall Classification Percentage			72.4

Model 2: Predicting Offence Type for Reoffending

A multi-nomial logistic regression analysis was performed in order to predict the likelihood of offence type for each unique court reappearance. In both the estimation and validation datasets, only those appearances that were associated with subsequent reoffences were included ($N = 15,483$ and $N = 12,620$ for the estimation and validation sets respectively).

The predictors included in the analysis were the same as the first model except for offence category type. Rather than use a binary indicator of offence type, a variable that included eight offence types was used as a predictor. The eight categories of offence type included:

1. Offences against the person
2. Break and enter, burglary
3. Theft and related offences
4. Drug offences
5. Traffic offences
6. Public order offences
7. Property damage
8. Other offences

Future offence type was the dependent variable of the multinomial model, using the predictors of Indigenous status, gender, age at appearance, appearance number, and previous offence type.

Results

Since multi-nomial logistic regression produces a regression equation for each outcome category, results will be presented for each predicted offence type. This will be followed by a summary of the predictive accuracy of the equations considered together.

Overall, the model provided a poor fit to the data (Pearson χ^2 (12327) = 12287.594, p = .60), though the final model was significant (-2 Log Likelihood = 19141.147, χ^2 (77) = 2738.50, p < .0005). This indicated that the predictors as a set were reliable in predicting offence type for subsequent reoffences. The variables accounted for a small amount of variance in future offence type, with Nagelkerke's R^2 = .17, and 95% confidence intervals ranging from .14 to .18.

A Table for each of the eight offence type models will not be presented. The multi-nomial model exhibited an overall offence type classification accuracy level of 32.6%, which may be considered poor (Table 5). In other words, the model was only correct in classifying 32.6% of court appearance cases according to their next offence. The application of the offence type model equations to the validation dataset produced similar results (Table 6), with an overall offence type classification accuracy level of 33.2%. However, the multi-nomial models continued to provide a poor fit to the data in the validation dataset.

Table 5: Classification accuracy (%) and percentage predicted by next offence type for estimation dataset

Offence Type	Observed	Percentage Classified	
	Percentage of Current Offences	Correctly by Offence Type	Percentage Predicted
Offences against the person	7.7	-	-
Break and enter, burglary	19.9	31.0	17.3
Theft and related offences	32.9	75.3	68.1
Drug offences	5.8	7.5	1.9
Traffic offences	6.0	23.6	5.0
Public order offences	7.9	0.1	>.01
Property damage	6.4	-	-
Other offences	13.5	13.7	7.7
Overall Percentage	100	32.6	100

Table 6: Classification accuracy (%) and percentage predicted by next offence type for validation dataset

Offence Type	Observed Percentage of Current Offences	Percentage Classified Correctly by Offence Type	Percentage Predicted
Offences against the person	8.7	0.9	.3
Break and enter, burglary	18.9	32.5	16.6
Theft and related offences	32.3	75.6	67.8
Drug offences	5.3	9.8	2.6
Traffic offences	5.8	23.9	4.9
Public order offences	8.1	-	-
Property damage	6.8	-	-
Other offences	14.0	14.6	7.7
Overall Percentage	100	33.2	100

Model 3: Predicting Time to Re-offence

A multiple regression model was produced in order to estimate the time to re-offence for court appearance cases that were associated with a subsequent re-offence ($N = 15483$). Months to next offence was entered as the dependent variable, with the variables of Indigenous male, Indigenous female, age at first appearance, appearance number, offence category (binary variable used in model 1), and months since previous appearance. Nonparametric bivariate correlations among the variables are presented in Table 7. All variables except binary offence type were significantly associated with the dependent variable of months to re-offence. The largest correlation of 0.5 was between appearance number and months since previous appearance. Multicollinearity was not an issue.

Table 7. Nonparametric bivariate correlations among the dependent and independent variables entered into the model predicting months to re-offence (N = 15,483)

Variable	1	2	3	4	5	6	7
1. Months to re-offence							
2. Indigenous status	-.05**						
3. Gender	-.04**	-.06**					
4. Age at appearance	-.21**	-.15**	.02**				
5. Appearance number	-.23**	.27**	.09**	.20**			
6. Offence type	.01	.02*	-.04**	.01	.00		
7. Months since previous appearance	-.09**	.11**	.05**	.14**	.50**	.03**	

* $p < .05$, ** $p < .01$

The dependent variable of months to next offence was highly positively skewed, with the majority of re-offenders taking less than 5 months to reappear in court (Figure 1). Due to assumptions of linearity in the regression model and the skewed nature of the dependent variable, months to reoffend was logarithmically transformed (Figure 2). All cases with log-transformed scores lower than -4 were removed from the analysis.

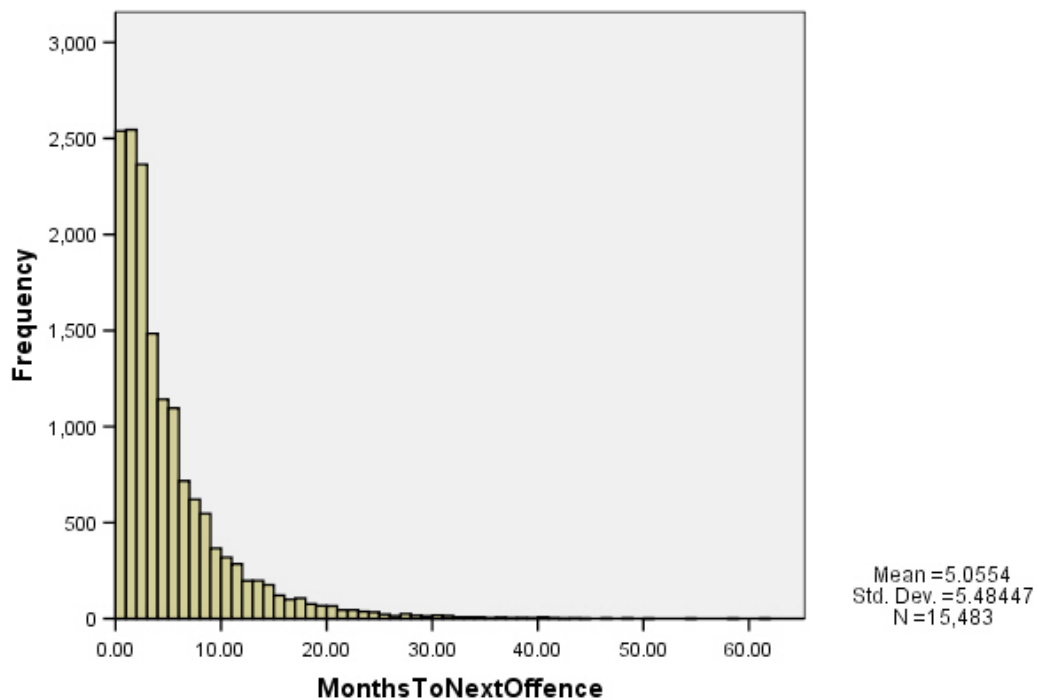


Figure 1: Distribution of the dependent variable of months to next offence

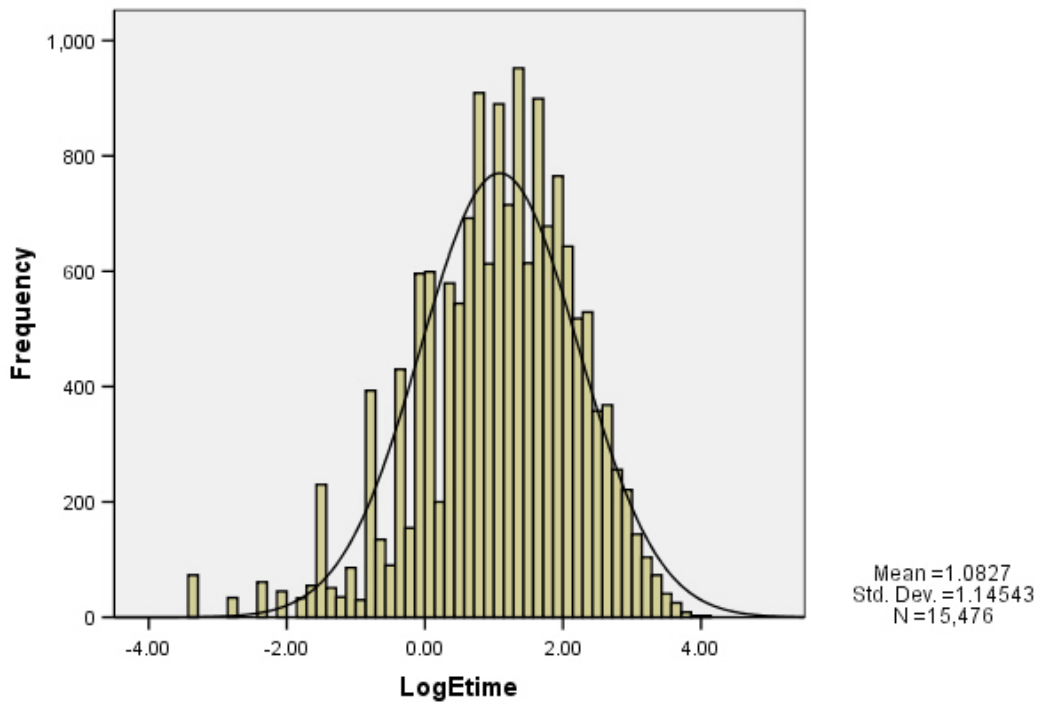


Figure 2: Distribution of the logarithmically transformed dependent variable of months to next offence

All independent variables were significantly skewed, with the variables of appearance number and months since previous appearance positively skewed, and age at first appearance negatively skewed. The variables were left untransformed. Examination of the standardized residuals of prediction and Mahalanobis distance through an initial run of the model identified 96 univariate and 275 multivariate outliers ($\chi^2(6) = 22.46$) that were removed from subsequent analyses, resulting in a total of 15,104 cases included in the regression model.

Table 8 displays the unstandardised regression coefficients (B), standard error of the unstandardised regression coefficients (*se*), standardized regression coefficients (β), the semipartial correlations (sr_i^2), R^2 , and adjusted R^2 for each variable in the time to reoffence model.

Table 8: Parameter estimates for the regression model predicting time to re-offence (N = 15,104)

Variable	B	se	β	sr_i^2
Indigenous status	-.09	.02	-.04	>.01
Gender	-.06	.02	-.02	>.01
Age at appearance	-.14	.01	-.17	.03
Appearance number	-.05	.00	-.15	.02
Offence category	.06	.03	.01	>.01
Months since previous appearance	.00	.00	.00	>.01
Constant	3.489	.11	R ² = .06 Adjusted R ² = .06 R = .25	

* $p < .05$, ** $p < .01$, *** $p < .001$, † $p < .0005$

R for regression was significantly different from zero, $F(6, 15097) = 173.46, p < .0005$, with an R^2 of .06 (95% confidence intervals from .05 to .07). This indicated that the variables as a set accounted for approximately 6.4% of the variance in time to reoffence. All variables except offence category and months since previous appearance were significant predictors of the time in months to re-offence. However, while the variables as a set were significant predictors, they only accounted for a small proportion of the variance in court appearance cases' times to re-offence. Age at appearance was the most powerful predictor, accounting for 3.0% of unique variance in months to reoffend. Figure 3 displays a histogram of the distribution of the standardised residuals of prediction for the regression model. Inspection of the residuals revealed that they were slightly negatively skewed.

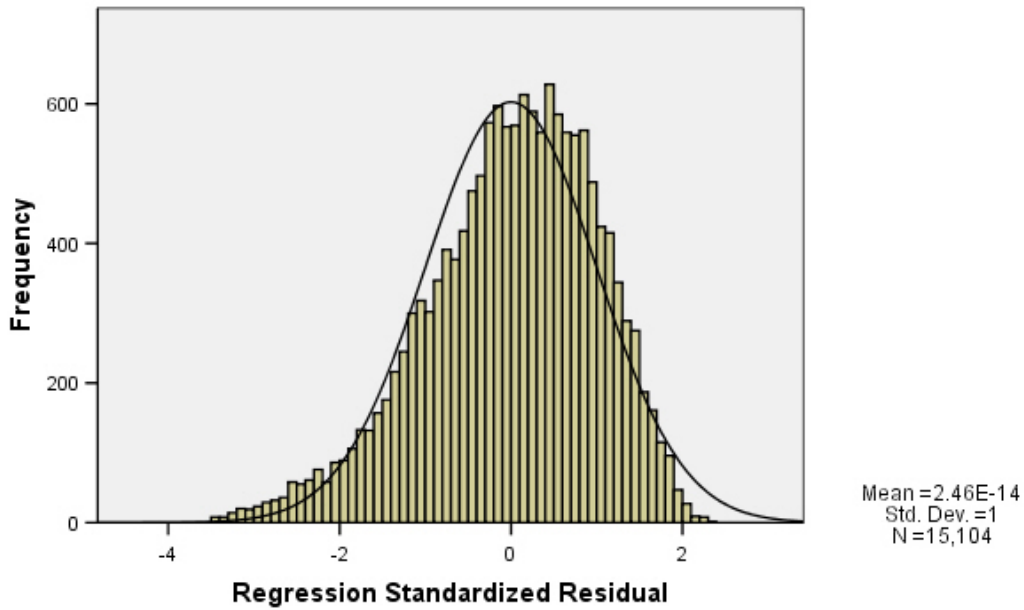


Figure 3: Histogram of the standardised residuals of prediction for the regression model predicting time to re-offence

In order to validate the regression model’s predictive accuracy, the equation derived from the estimation set was applied to the validation set:

$$\text{Log of Months to reoffence} = 3.489 + (-.088)(\text{ind}) + (-.064)(\text{gen}) + (-.144)(\text{age}) + (-.045)(\text{appearance number}) + (.056)(\text{catoff}) + (-.001)(\text{months since previous appearance})$$

Displayed in Figure 4 is histogram of the distribution of the standardised residuals of prediction for the validation regression model. Inspection of the residuals revealed that they were negatively skewed. This indicated that the regression model derived from the estimation data set provided a poor fit to the validation data due to outliers in the solution.

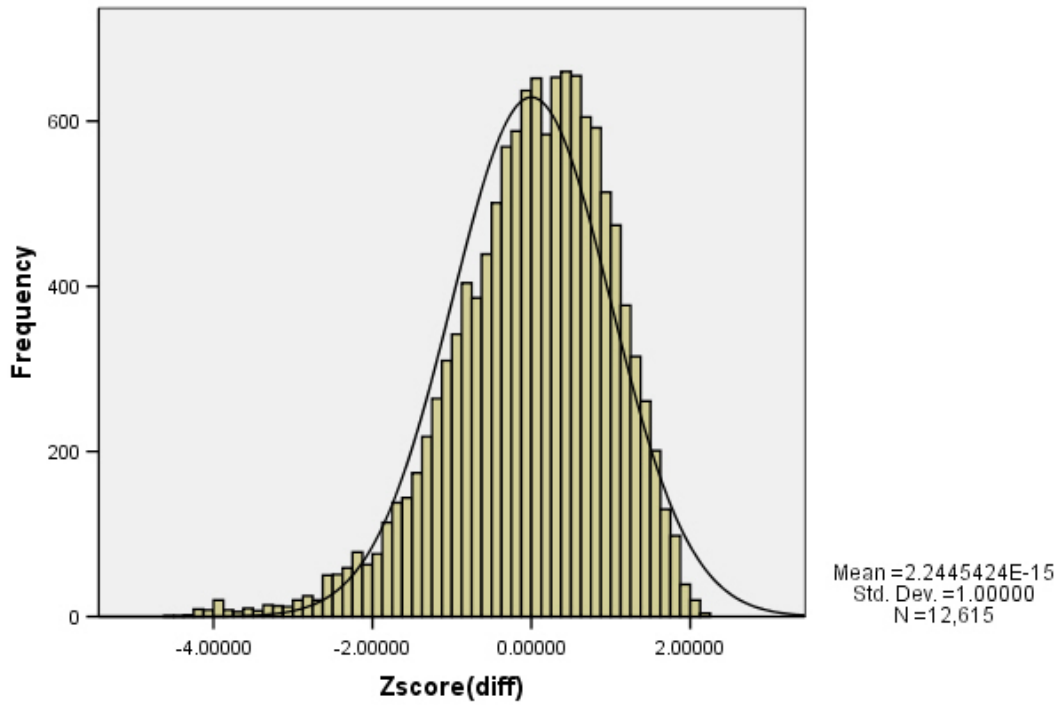


Figure 4: Histogram of the standardised residuals of prediction for the validation regression model predicting log of time to re-offence

Appendix 2: Representations of Resulting Decision Trees

Decision Tree for Model 1, Predicting re-offence

node), split, number of cases, loss, predicted does reoffend (0 = no, 1 = yes), (probability of reoffending outcomes)

* denotes terminal node

- 1) root 24794 9311 1 (0.3755344 0.6244656)
- 2) AgeContinuousAtAppearance \geq 16.54346 5497 1392 0 (0.7467710 0.2532290)
- 4) AgeContinuousAtAppearance \geq 16.88296 2063 211 0 (0.8977218 0.1022782) *
- 5) AgeContinuousAtAppearance $<$ 16.88296 3434 1181 0 (0.6560862 0.3439138)
- 10) AppNumMaxFactor=1,2 1730 377 0 (0.7820809 0.2179191) *
- 11) AppNumMaxFactor=3,4,5 1704 804 0 (0.5281690 0.4718310)
- 22) AgeContinuousAtAppearance \geq 16.74059 735 258 0 (0.6489796 0.3510204)
- 44) AverageMonthsBetweenAppearances \geq 3.180287 524 150 0 (0.7137405 0.2862595) *
- 45) AverageMonthsBetweenAppearances $<$ 3.180287 211 103 1 (0.4881517 0.5118483)
- 90) AgeContinuousAtFirstAppearance \geq 15.19644 128 53 0 (0.5859375 0.4140625)
- 180) AverageMattersPerAppearance $<$ 5.212121 108 38 0 (0.6481481 0.3518519) *
- 181) AverageMattersPerAppearance \geq 5.212121 20 5 1 (0.2500000 0.7500000) *
- 91) AgeContinuousAtFirstAppearance $<$ 15.19644 83 28 1 (0.3373494 0.6626506) *
- 23) AgeContinuousAtAppearance $<$ 16.74059 969 423 1 (0.4365325 0.5634675)
- 46) AverageMonthsBetweenAppearances \geq 7.008507 243 100 0 (0.5884774 0.4115226) *
- 47) AverageMonthsBetweenAppearances $<$ 7.008507 726 280 1 (0.3856749 0.6143251) *
- 3) AgeContinuousAtAppearance $<$ 16.54346 19297 5206 1 (0.2697829 0.7302171)
- 6) AppNumMaxFactor=1 7057 3242 1 (0.4594020 0.5405980)
- 12) MonthsTillEdgeEffectCategorical=10,11,12,13 3230 1294 0 (0.5993808 0.4006192)
- 24) AgeContinuousAtFirstAppearance \geq 16.16564 1181 359 0 (0.6960203 0.3039797) *
- 25) AgeContinuousAtFirstAppearance $<$ 16.16564 2049 935 0 (0.5436798 0.4563202)
- 50) indigsex=3 374 107 0 (0.7139037 0.2860963) *
- 51) indigsex=1,2,4 1675 828 0 (0.5056716 0.4943284)
- 102) offtype=1,5 336 121 0 (0.6398810 0.3601190) *
- 103) offtype=2,3,4,6,7,8 1339 632 1 (0.4719940 0.5280060)
- 206) crtorder=3,4,6 1108 553 1 (0.4990975 0.5009025)
- 412) indigsex=1 748 350 0 (0.5320856 0.4679144) *
- 413) indigsex=2,4 360 155 1 (0.4305556 0.5694444) *
- 207) crtorder=5 231 79 1 (0.3419913 0.6580087) *
- 13) MonthsTillEdgeEffectCategorical=14,15 3827 1306 1 (0.3412595 0.6587405)
- 26) indigsex=1,3 2616 1048 1 (0.4006116 0.5993884)
- 52) AgeContinuousAtFirstAppearance \geq 13.86858 2004 879 1 (0.4386228 0.5613772)
- 104) indigsex=3 538 244 0 (0.5464684 0.4535316) *
- 105) indigsex=1 1466 585 1 (0.3990450 0.6009550) *
- 53) AgeContinuousAtFirstAppearance $<$ 13.86858 612 169 1 (0.2761438 0.7238562) *
- 27) indigsex=2,4 1211 258 1 (0.2130471 0.7869529) *
- 7) AppNumMaxFactor=2,3,4,5 12240 1964 1 (0.1604575 0.8395425) *

Decision Tree for Model 2, Predicting Next Offence Type

node), split, n, loss, yval, (yprob)

* denotes terminal node

- 1) root 15454 10884 3 (0.083 0.2 0.3 0.068 0.071 0.079 0.059 0.15)
- 2) OffType2Distn \geq 0.1961538 5722 4006 2 (0.065 0.3 0.26 0.066 0.06 0.069 0.054 0.13)
- 4) indigsex=2 3150 2048 2 (0.068 0.35 0.24 0.047 0.042 0.074 0.054 0.13) *
- 5) indigsex=1,3,4 2572 1826 3 (0.062 0.24 0.29 0.089 0.082 0.063 0.054 0.12)
- 10) OffType5Distn $<$ 0.2111111 2389 1672 3 (0.065 0.24 0.3 0.088 0.067 0.063 0.054 0.12)
- 20) offtype=1,3,4,5,6 864 564 3 (0.078 0.19 0.35 0.097 0.069 0.068 0.037 0.12) *
- 21) offtype=2,7,8 1525 1106 2 (0.058 0.27 0.27 0.083 0.066 0.06 0.064 0.12)
- 42) OffType8Distn $<$ 0.03125 1144 803 2 (0.052 0.3 0.29 0.073 0.065 0.065 0.062 0.099)
- 84) crtorder=4,5 968 673 3 (0.054 0.28 0.3 0.075 0.066 0.068 0.055 0.096) *
- 85) crtorder=3,6,7 176 107 2 (0.04 0.39 0.19 0.062 0.057 0.045 0.1 0.11) *
- 43) OffType8Distn \geq 0.03125 381 292 3 (0.079 0.2 0.23 0.11 0.068 0.045 0.068 0.19) *
- 11) OffType5Distn \geq 0.2111111 183 131 5 (0.016 0.19 0.16 0.093 0.28 0.066 0.049 0.15) *
- 3) OffType2Distn $<$ 0.1961538 9732 6649 3 (0.094 0.14 0.32 0.069 0.078 0.085 0.062 0.16)
- 6) OffType3Distn \geq 0.3798077 5103 3070 3 (0.078 0.15 0.4 0.063 0.056 0.064 0.057 0.13) *
- 7) OffType3Distn $<$ 0.3798077 4629 3579 3 (0.11 0.12 0.23 0.075 0.1 0.11 0.067 0.19)
- 14) OffType5Distn \geq 0.2071429 702 474 5 (0.08 0.11 0.18 0.068 0.32 0.083 0.031 0.13) *
- 15) OffType5Distn $<$ 0.2071429 3927 3002 3 (0.12 0.12 0.24 0.077 0.062 0.11 0.074 0.2)
- 30) OffType4Distn \geq 0.1602564 677 512 3 (0.074 0.12 0.24 0.21 0.074 0.083 0.049 0.14)
- 60) OffType8Distn $<$ 0.3166667 572 426 3 (0.066 0.12 0.26 0.23 0.077 0.084 0.047 0.12)
- 120) offtype=2 9 3 2 (0 0.67 0.11 0 0 0 0 0.22) *
- 121) offtype=1,3,4,5,6,7,8 563 418 3 (0.067 0.11 0.26 0.23 0.078 0.085 0.048 0.12)
- 242) AgeContinuousAtAppearance $<$ 16.15195 382 271 3 (0.063 0.12 0.29 0.2 0.068 0.094 0.045 0.12) *
- 243) AgeContinuousAtAppearance \geq 16.15195 181 125 4 (0.077 0.088 0.19 0.31 0.099 0.066 0.055 0.12) *
- 61) OffType8Distn \geq 0.3166667 105 76 8 (0.11 0.11 0.18 0.12 0.057 0.076 0.057 0.28) *
- 31) OffType4Distn $<$ 0.1602564 3250 2490 3 (0.13 0.12 0.23 0.048 0.059 0.12 0.079 0.21)
- 62) AgeContinuousAtAppearance $<$ 14.6872 945 677 3 (0.15 0.17 0.28 0.02 0.038 0.088 0.1 0.15) *
- 63) AgeContinuousAtAppearance \geq 14.6872 2305 1759 8 (0.12 0.1 0.21 0.06 0.068 0.13 0.069 0.24)
- 126) OffType8Distn $<$ 0.218254 1204 939 3 (0.13 0.11 0.22 0.065 0.074 0.15 0.071 0.17) *
- 127) OffType8Distn \geq 0.218254 1101 763 8 (0.099 0.095 0.21 0.054 0.062 0.11 0.066 0.31) *

Decision Tree for Model 3. Predicting Log of Months to Re-offence

node), split, n, deviance, yval

* denotes terminal node

- 1) root 15454 20504.26000 1.0788830
- 2) MonthsTillEdgeEffectLn < 2.187708 2732 2943.94000 0.5169473
 - 4) AgeContinuousAtAppearance >= 16.79535 440 498.80690 -0.2184670
 - 8) AgeContinuousAtAppearance >= 16.99247 38 83.50575 -1.7229620 *
 - 9) AgeContinuousAtAppearance < 16.99247 402 321.15740 -0.0762510 *
 - 5) AgeContinuousAtAppearance < 16.79535 2292 2161.48300 0.6581263
 - 10) AgeContinuousAtAppearance >= 16.61739 636 534.21000 0.4191058 *
 - 11) AgeContinuousAtAppearance < 16.61739 1656 1576.98300 0.7499241 *
- 3) MonthsTillEdgeEffectLn >= 2.187708 12722 16512.37000 1.1995560
 - 6) AppNumMaxFactor=3,4,5 6908 8553.32300 0.9931898
 - 12) AppNumMaxFactor=5 4100 4957.13000 0.8941837 *
 - 13) AppNumMaxFactor=3,4 2808 3497.32400 1.1377500 *
 - 7) AppNumMaxFactor=1,2 5814 7315.30600 1.4447540
 - 14) MonthsTillEdgeEffectLn < 2.887284 1786 1816.93900 1.2077370 *
 - 15) MonthsTillEdgeEffectLn >= 2.887284 4028 5353.54900 1.5498460
 - 30) AppNumMaxFactor=2 1504 2022.93700 1.3617660 *
 - 31) AppNumMaxFactor=1 2524 3245.70800 1.6619190
 - 62) MonthsTillEdgeEffectLn < 3.667029 1888 2371.69900 1.5976610 *
 - 63) MonthsTillEdgeEffectLn >= 3.667029 636 843.07070 1.8526730 *