DECISION TREE ANALYSIS: PREDICTION OF SERIOUS TRAFFIC OFFENDING

ABSTRACT

The objective was to predict whether an offender would commit a traffic offence involving death, using decision tree analysis. Four groups of predictive models were produced; two based on serious offending only (i.e. traffic offences that escalated to Court) and two based on time-bound (i.e. Period) data constructed from all traffic offending. The small number of "risk" offenders who committed a traffic offence involving death - 1.44% of the Court-based records and 0.14% of Period-based records - led to the use of a profit matrix in an attempt to mitigate against the small number of target records. While none of the predictive models were particularly useful in a practical sense, the non-aggregated Court data produced the best result as some individual nodes were more useful than others. The decision tree models generated for the Period data were least useful, although this could be due to the much lower number of "risk" offenders and the relatively short timeframe covered by the Period data. A key finding is that aggregated data based on predefined Police traffic offence groups appears unsuitable for decision tree analysis.

1. INTRODUCTION

With the availability of vast quantities of information via the implementation of data warehouses, increases in computing power, and the availability of appropriate software, data mining techniques can now be implemented by many large organisations. While many of the applications of data mining are obvious for private sector companies, such as the detection of fraud (e.g. banking and insurance organisations), and customer segmentation for special offers (e.g. credit card and telecommunications companies), government organisations have similar needs. For example, the accurate detection of fraud is also clearly of benefit to government departments such as the Inland Revenue Department. Regardless of type, any organisation that holds a large volume of data in a data warehouse/data mart environment will be interested in data mining applications, particularly as historical statistical techniques such as cluster analysis are not well suited to analyse thousands, if not millions, of records.

Police jurisdictions are also facing the new problem of “too much data”. Perhaps unsurprisingly therefore, articles are now being published showing the application of data mining techniques to Police-stored data. The use of data mining techniques to create predictive models increases the attractiveness of these methods, due to the potential for crime prevention and reduction. Within road policing, recent publications on data mining applications include the neural network modelling of the relationship between (mainly quantitative) driver characteristics and probability of traffic crash in Turkey (Kalyoncuoglu & Tigdemir, 2004), and a comparison of the accuracy of neural network, logistic regression, and decision tree models for predicting Korean traffic crash severity from a multitude of variables, such as road type, car shape, and speed of car (Sohn & Shin, 2001). One major area of road policing interest in data mining therefore arises from an interest in traffic crash reduction, particularly when applied to fatal and serious injury crashes.

In New Zealand, most fatal and serious injury traffic crashes are predominantly due to driver error. In fatal crashes, when the “at fault” driver dies, no Police prosecution occurs, and the matter falls within the jurisdiction of the Coroners Court. However, in the cases where the “at fault" driver lives, having killed another party, the offender is charged with the appropriate traffic offence and appears in District Court. While road deaths and serious injuries are a relatively rare event, they are outcomes with obviously severe consequences for the deceased or injured individual and their family, as well as being a source of economic burden. At the time this research was conducted, the social cost of each road crash fatality was approximately $2.63m (Land Transport Safety Authority, 2002).

The main purpose of this study was to produce a predictive model of the characteristics of traffic offenders who are at risk of causing a road traffic fatality to a person other than themselves. Due to a lack of previous research in this area, the analyses were conducted for aggregated, as well as non-aggregated, data. The pragmatic objective of the research was to produce decision tree models with nodes containing a preponderance of "risk" offenders, where the rules for the nodes could be used to identify potential "risk" offenders for intervention. Conversely, because "no-risk" offenders also represent a "no intervention" condition, models with nodes containing all or mostly "no risk" offenders, in combination with mixed "risk" and "no-risk" offender nodes, would be unhelpful due to the lack of practical, predictive ability for identifying likely "risk" offenders.

2. METHOD

2.1 DATA

There were two sources of data: Court-based traffic offences ("Court data") and period-based traffic offences ("Period data"). The Court data contains only traffic convictions that either were serious enough to automatically result in a Court appearance (e.g. drink driving offences), or were escalated to Court due to non-payment of traffic fines. The Court data represents a continuous period of approximately 35 years, so that most offenders within the dataset will have their entire Court-related traffic history included. Conversely, the Period data contains all traffic offences – minor as well as major – over a much smaller four-year period (January 2000 to December 2003), with the additional criterion that the driver had committed at least one traffic offence in the July 2003 to December 2003 period. This latter condition was to ensure that at least 3.5 years' potential driving offending history was included in the Period data to ensure sufficient driver history for the analyses.

For each data source, two datasets were constructed: an "non-aggregated" dataset and an "aggregated" dataset. The purpose of the non-aggregated datasets were to determine if the sequence of offending was predictive. The purpose of the aggregated datasets was to determine whether the volume of offending, rather than sequence, was more predictive. In both types of dataset, the record is at offender level, the difference in the datasets is simply the level of offence detail stored against each offender. The non-aggregated dataset retained information on each individual offence (e.g. offence code, age at offence) for each offender. Conversely, the aggregated dataset only contained summary information (e.g. total number of alcohol-related offences, total number of officer-issued speed offences) for each offender.

2.2 VARIABLES

Each person was classified as a "risk" or "no-risk" offender on the basis of whether they had committed one (or more) of ten traffic offences relating to causing death, creating the target variable for the models. As the occurrence of this type of offence is the reason for the risk, each of these offences, and all subsequent offences for the same offender, were deleted prior to analysis. Where an offender had committed more than one of these offences, the first offence was used as the point of deletion.

The other variables stored as standard in all four datasets were gender and total number of offences. The two non-aggregated datasets contained information about each offence, such as age of offender, and days between previous and immediately subsequent offence. Due to the volume of offending, the non-aggregated Court data contained 44 sets of offence-related data per offender, and the non-aggregated Period data contained 83 such sets. Because the non-aggregated Period data contained only four continuous years of data, whereas the Court data contained around 35 years, the order of offence information in the non-aggregated Period data was reversed. This meant that data on offence1 related to the most recent offence, offence2 was the second-latest offence, and so forth. This time reversal was made to compensate for the short duration of the data, as it was highly unlikely that the Period data would contain all offending for most offenders. The further assumption was made that the inability to produce an offending sequence moving forwards in time did not preclude the ability to produce such a sequence that moved backwards instead.

The two aggregated datasets were much simpler, with most of the variables simply being counts. Age of offender was limited to age at first and last offence. The number of offences were summarised into counts within general offence category. For example, this meant that all alcohol offences were counted together in one "alcohol" category, all police officer-issued speeding offences were summed together in one "speed" category and so forth. These aggregated datasets were used to determine whether it was the volume of offences that was more predictive than the exact nature of the offending. Clearly ordering of offending is unimportant for the aggregated datasets, with data completeness being the most important factor instead so that the offence summations are accurate.

2.3 DATA ANALYSIS

Four datasets were analysed: non-aggregated Court data; aggregated Court data; non-aggregated Period data, aggregated Period data. Due to the volume of data to analyse, the data mining was limited to one technique: decision tree analysis. As the non-aggregated datasets contain a high volume of categorical data (gender, offence code) and the aggregated datasets contain a high volume of count data, which also cannot be analysed parametrically, it was necessary to choose a method that was suitable for non-parametric data. The use of a parametric data mining technique, such as neural network or stepwise logistic regression, would have resulted in the creation of literally hundreds of binary indicator variables, hence these methods were not investigated. Three types of decision tree model were assessed for each dataset: Gini reduction; Entropy reduction; and Chi-square to determine if there was a clearly superior decision tree technique that should be recommended on the basis of this research.

SAS Enterprise Miner was used to perform the analyses. The social cost of a road fatality, mentioned earlier, was used to construct a profit matrix for the analysis (see Table 1 below). The rows of the matrix are the true outcome, and the columns are the predicted outcomes (from the decision tree model). Where there is no true risk, as represented by cells C and D in the table, the costs of both the correct and incorrect *prediction* of risk has been set to $0, to represent the minimal effect these classifications have on social cost. On the other hand, the major costs are associated with true risk (cells A and B). Where a predictive model fails to assign risk to an offender, the profit matrix assumes one death will occur as a result, causing a cost of $2,626,800 per false negative (cell B). Where the risky offender is accurately predicted, a saving of $656,700 is used, to suggest that any intervention with these offenders designed to prevent road fatalities is effective approximately 25% of the time (cell A).

Table 1. Profit matrix for traffic offence analyses.

Actual risk	Predicted:	
	Risk offence	No risk offence
Risk offender	$656,700 (A)	-$2,626,800 (B)
No-risk offender	$0 (C)	$0 (D)

The profit matrix decision was set to “Minimize loss”, to bias the model so that the probability of false negatives (cell B outcomes) are reduced in a situation where the vast majority of the dataset are true negatives. Finally, the Model Assessment Measure was set to “Average loss in top 10%”, as pre-testing showed that other Model Assessment Measure options caused the production of a only single node with all offenders classified as “no-risk”.

Due to the higher proportion (and higher absolute number) of “risk” offenders in the Court data (see Results section below), 50% of offenders were allocated to the training subset, and the remaining 50% were split equally between the validation and test subsets for the Court data. Conversely, the Period data was split only between training and validation subsets, with 65% in the former and 35% in the latter.

3. RESULTS

The constructed datasets were very large, with a very small proportion of “risk” offenders in each. The Court data is based on the records of 530,605 offenders, of which “risk” offenders represent 1.44% of the total. The Period data is based on 321,474 offenders, of which “risk” offenders represent an even smaller proportion at 0.14%. The proportion of offences relating to “causing death” are very rare when examined over the entire range of traffic offending, but are somewhat more common when considered only in the light of serious offending that proceeds to Court. Because the focus of this paper is on the correct classification of offenders, the results presented here are a subset of the output produced, and are those primarily concerned with misclassification rates.

3.1 COURT DATA

Table 2 summarises the misclassification rates for both the non-aggregated and aggregated Court datasets respectively. A misclassification rate of 1.44% (i.e. 0.0144) could be achieved simply by classifying all offenders into a single “no-risk” category. Of particular interest are the misclassification rates for the “test” subsets, as the results for the test subsets (by definition) resemble the misclassification rates that would be achieved in practice by each model. In no model was there even a single node that only contained “risk” offenders; in every case there was a mixture of “risk” and “no-risk” offenders.

Table 2. Misclassification rates for the Court datasets.

Model	Non-aggregated Court data			Aggregated Court data		
	Misclassification rate			Misclassification rate		
	Training	Validation	Test	Training	Validation	Test
Chi square	0.0142177058	0.0148057685	0.0148057685	0.0143921008	0.0143883806	0.0143883806
Entropy	0.0143760153	0.0143760695	0.0143760695	0.0143921008	0.0143883806	0.0143883806
Gini	0.0143760153	0.0144514553	0.014421301	0.0143883806	0.0143883806	0.0143883806

Table 2 suggests that all three decision tree methods are equally suitable for the modelling of aggregated Court data, whereas the Entropy tree appears slightly superior in the case of the non-aggregated data. However, there is a clear difference in the complexity of the decision trees produced for each of these datasets, with the non-aggregated data requiring over 100 nodes per tree. This leads to complex decision rules for prediction, although these can be easily automated in software. For example, the rule for Node 64 of the Entropy decision tree for the non-aggregated Court data, where between 5% and 8% of drivers are "risk" offenders is:

1. if offence1 is one of: (A302 B101 B123 B154 B201 C101 K213 N309 R516) AND PREC2 IS ONE OF: (A301 A302 A313 A314 A322 A332 B101 B107 B108 B131 B141 B207 C101 D101 D701 H203 H211 H501 H708 K110 L204 L413 N309 V222); and
2. offence3 is one of: (A101 A102 A109 A305 A306 A309 A311 A313 A323 A330 A331 A501 A507 A514 A515 A518 A530 B109 B110 B126 B127 B128 B132 B140 B141 B142 B184 B203 B208 B301 B401 B501 C101 C201 C214 D101 D201 D301 D351 D401 D502 D701 D703 D710 D804 D902 E101 E102 F102 F201 F301 F501 G101 H101 H202 H211 H301 H501 H714 K101 L111 L112 L141 L142 L143 L144 L201 L202 L403 L404 L406 L408 L411 L416 L417 L427 M120 M121 M123 M128 M131 M204 M206 M207 M209 M401 N353 N459 T802 U101 V408 V419 Y901); and
3. offence4 is one of: (A101 A102 A109 A130 A300 A301 A303 A305 A306 A309 A315 A320 A322 A323 A330 A501 A507 A514 A515 A518 A530 B101 B105 B106 B108 B109 B110 B111 B126 B128 B131 B133 B140 B141 B142 B143 B184 B201 B203 B205 B211 B301 C101 C201 C401 D101 D201 D301 D502 D703 D715 D902 D903 E101 E102 E301 F101 F102 F201 G101 G201 H101 H206 H214 H501 H707 H711 H718 K101 L111 L120 L128 L141 L142 L143 L201 L202 L230 L406 L407 L408 L413 M122 M123 M125 M126 M128 M205 M206 M207 M401 N302 N353 O105 S101 V206 V418 Y901); and
4. age at offence1 is greater than or equal to 21.408 years; and
5. the fine for first offence is greater than or equal to $23.

While the decision trees for the aggregated data appear to have performed better in comparison to those for the non-aggregated data, based on the misclassification results, these models were still not very helpful in practice. Of the relatively few nodes which achieved a higher concentration of "risk" offenders, typically only around 6% of each node were "risk" offenders. In addition, the higher performing nodes also typically contained a smaller number of offenders; in a number of cases each of the better differentiating nodes contained less than 1% of the dataset.

3.2 PERIOD DATA

Table 3 summarises the misclassification rates for the non-aggregated and aggregated Court datasets respectively. A misclassification rate of 0.14% (i.e. 0.0014) could be achieved simply by classifying all offenders into a single “no-risk” category. Due to the lack of a test subset, the change in misclassification rates between the training and validation subsets are of interest here.

Table 3. Misclassification rates for the Court datasets.

Model	Non-aggregated Period data			Aggregated Period data		
	Misclassification rate			Misclassification rate		
	Training	Validation	Test	Training	Validation	Test
Chi square	0.0013926177	0.0013864818	NA	0.0013926177	0.0013864818	NA
Entropy	0.0013926177	0.0013864818	NA	0.0013926177	0.0013864818	NA
Gini	0.0013926177	0.0013864818	NA	0.0013926177	0.0013864818	NA

Table 3 suggests that, for the Period data, aggregated and non-aggregated data both provide suitable predictive models. The results further suggest that the three decision tree models perform equally well with this data. Examination of the models showed that, while the decision trees tended to be simpler than those for the Court data, with much fewer nodes, there were simply no useful nodes. Compared to the nodes produced from the Court data, the nodes based on the Period data that contained large numbers of offenders also contained a greater mixture of “risk” and “no risk” offenders.

In essence, the very low proportion of "risk" offenders in the Period data appears to have been swamped by the volume and complexity of patterns of traffic offending across both "risk" and "no-risk" offenders. This problem occurred for both the non-aggregated and aggregated Period datasets. With data this complex, it is likely that more than 0.14% of the entire dataset (approximately 450 offenders) should be "risk" offenders in order to produce a more predictive model.

4. DISCUSSION

This research compared the relative usefulness of three types of decision trees (Chi-square, Entropy reduction, and Gini reduction) for predicting "risk" offenders, using two types of data: (historical Court traffic offences and Period traffic offences). The major finding was that none of the decision tree models was successful in accurately identifying characteristics of "risk" offenders who committed a traffic offence "causing death". While the overall misclassification rates for the models were acceptably low, the lack of individual nodes containing predominantly "risk" offenders means that the decision tree models are simply not useful in practice. In other words, no rules were produced that successfully differentiated "risk" offenders from "non-risk" offenders.

While the modelling work was not very successful in practice, some useful conclusions can be drawn. First, on a general methodological note, because the target offenders represented a very small proportion of offenders overall (1.44% for Court data and 0.14% for the Period data), two special adjustments were made: the use of a profit matrix where the "risk" offenders were heavily weighted, particularly with regard to the cost of a false negative decision (cell B of Table 1); and the Model Assessment Measure within each decision tree node was set to “Average loss in top 10%”. This combined adjustment methodology should be followed in every similar decision tree analysis where the target group is extremely small, where the values on the predictor variables have a large overlap with the non-target group. Failure to follow this methodology where the target group is extremely small is likely to cause the decision tree to produce only one node (i.e. the tree does not “grow”).

Second, when dealing with non-aggregated offending data, serious offences – defined as those that reach Court – appear to be somewhat more useful in identifying offending patterns that are more likely to lead to a subsequent "risk" offence. Two of the non-aggregate Court data models contained useful nodes for targeting offenders who appear more likely to proceed to committing a Risk offence. Given the large social cost associated with a motor vehicle fatality ($2,626,800 using the June 2002 estimate), mistargeting of even a relatively large number of “not risk” offenders may still produce a cost-effective result from reallocation of resources into prevention (e.g. selection of offenders into defensive driving courses).

Third, aggregation of offence data using predefined Police offence categories (A-series offences, B-series offences, etc) worsened the predictive power of the decision trees. This outcome occurred for both the Court and the Period data. While the Police offence categories have face validity, this does not appear to translate into construct or predictive validity. This result does not mean that offences cannot be grouped prior to decision tree analysis, but rather that a mathematical grouping of the individual offences must occur instead, using a statistical dimension reduction technique such as principal components analysis. Due to the time constraints on this research, such a classification was not undertaken. This type of classification work should be seriously considered prior to any similar (i.e. aggregated offence) analyses being performed in the future.

Finally, the Period data produced the least useful models, regardless of whether or not the data was aggregated. Given that the Court data represented – in most instances – each offender’s entire offending history as it related to Court, the lack of useful models from the Period data may simply demonstrate that a longer time series is necessary when including all minor and moderate, as well as serious, traffic offending in modelling work. Therefore, until longer time series’ are analysed using this methodology, it is premature to form any conclusions on the usefulness (or otherwise) of performing decision tree analyses on “all offending” data.

REFERENCES

Kalyoncuoglu, S.F. & Tigdemir, M. (2004). An alternative approach for modelling and simulation of traffic data: artificial neural networks. Simulation Modelling Practice and Theory, 12, 351-362.

Land Transport Safety Authority (2002). The social cost of road crashes and injuries: June 2002 update. Wellington: Land Transport Safety Authority.

Sohn, S.Y., & Shin, H. (2001). Technical note: Pattern recognition for road traffic accident severity in Korea. Ergonomics, 44(1), 107-117.