2015-01-15

An Analysis of the Predictive Capability of C5.0 and Chaid Decision Trees and Bayes Net in the Classification of fatal Traffic Accidents in the UK

Aiden O'Connor
Technological University Dublin

Recommended Citation

O'Connor, A (2015) *An analysis of the predictive capability of C5.0 and Chaid decision trees and Bayes net in the classification of fatal traffic accidents in the UK*. DIT 2015

“An analysis of the predictive capability of C5.0 and Chaid decision trees and Bayes net in the classification of fatal traffic accidents in the UK”

Aiden O’Connor

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

M.Sc. in Computing (Data Analytics)

January 2015

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: ***Aiden O'Connor***

Date: ***27th January 2015***

ABSTRACT

Road traffic accidents are a significant cause of deaths worldwide and there is a global focus on understanding accident contributory factors and implementing prevention strategies. Although accident statistics are steadily improving, effective prevention must be persistent, evidence based and properly resourced. This research aimed to extract fatal traffic accident prediction from UK STATS19 accident data using C5.0 and Chaid decision trees and Bayes net classification models. Data was grouped as either fatal or non-fatal. The class imbalance due to fatal accident infrequency was considered and data transformation and sampling techniques were applied to increase prediction likelihood. Chaid was used for supervised discretisation and proved effective in identifying homogeneous subgroups. SPSS Modeler was used for data preparation and model build. Model performance was evaluated using accuracy, recall, precision and ROC curves.

The experiment design and data preparation approach successfully predicted fatal accidents with high recall results, however, significant misclassification of non-fatals as fatals led to poor accuracy and precision performance. Boosting was subsequently tested and achieved some accuracy improvement. Serious accidents were grouped as non-fatal in the initial data analysis, however, are likely to hold similar characteristics to fatal and the models therefore struggled to classify correctly as non-fatal. Changing the experiment design to select fatal, serious and slight as targets may improve the models accuracy. Overall, the models succeeded in classifying fatal traffic accidents correctly and this was the original objective of the research.

Interpretation of business rules, by ranking rules and summarising in a standard format, proved effective for understanding and comparison of key predictors. When comparing both C5.0 and Bayes net models, the contributory factors identified were consistent, with road surface and urban/rural identified as the strongest predictors for both models. The experiment demonstrated that classification techniques can be used to predict infrequent events once sampling techniques are applied.

Key words: *Predictive analytics, fatal traffic accidents, classification techniques, imbalanced datasets.*

ACKNOWLEDGEMENTS

I would like to thank my supervisor Pierpaolo Dondio for his support and guidance in completing my dissertation.

I would also like to thank my wife Marie for her support and patience throughout the dissertation process and the M.Sc.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 INTRODUCTION

1.1 Introduction

Road traffic accidents are the eighth leading cause of deaths worldwide with over one million people dying on the roads each year and trends suggest that by 2030 road traffic deaths will have risen to the fifth leading cause (The World Health Organisation, 2013, p. vii). Organisations across the globe are focussed on road traffic accident analysis and understanding and proven strategies exist which can help to reduce road traffic deaths (The World Health Organisation, 2013, p. 1). Research has sought to identify factors which contribute to traffic accidents and use those factors for more effective prediction and as a guide to road safety planning and traffic accident prevention (Lord & Mannering, 2010). Speed, age, alcohol consumption and driving fatigue are some of the factors commonly associated with fatal road traffic accidents (The World Health Organization, 2004). The ability to accurately identify the key factors which contribute most to fatalities could help focus road safety planning efforts. Extracting actionable insights from historical information is a key aim of using predictive analytics. Predictive analytics techniques can be used to extract prediction from data by identifying patterns which may otherwise have gone unnoticed. Classification techniques are commonly used to identify key underlying relationships between data features and identify the key predicting features. Fatal traffic accidents are infrequent and are considered to be random events which increases the prediction difficulty. Sampling techniques can be applied to extract patterns from data with infrequent events.

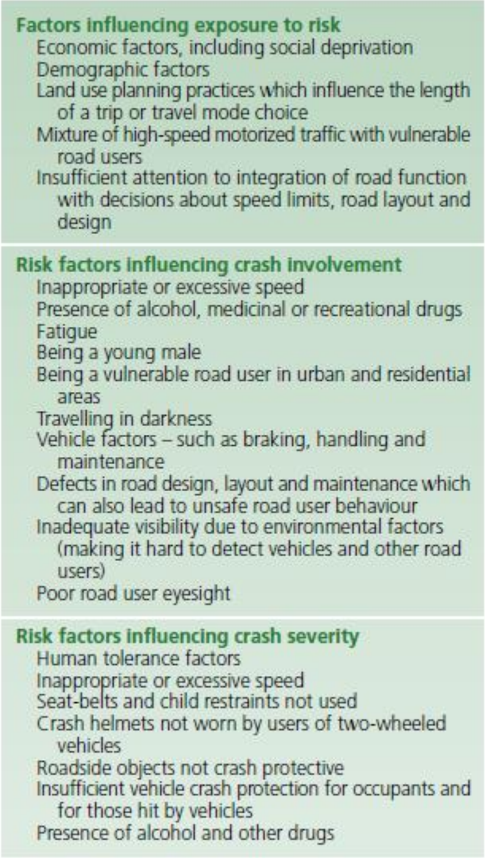
This research project investigates the use of three classification techniques, C5.0 and Chaid decision trees and Bayes net to predict fatal traffic accidents. An outline of road traffic accident literature provides background to the experiment and an understanding of the key data characteristics. Literature review for data mining and predictive analytics, including the three algorithms selected for the experiment, relevant model evaluation techniques and current traffic accident prediction research are discussed. The experiment design follows a standard methodology and focuses on understanding and preparing the data, building the models and model performance assessment and evaluation. The experiment implementation is described and model results are presented, assessed and evaluated, with key findings outlined. The conclusion summarises the experiment execution, the findings, the limitations and future work which could enhance the findings.

1.2 Background

This chapter outlines the key components of traffic safety and the role road safety systems play to prevent accidents. Information sharing, the role of data and predictive analytics techniques are discussed to understand how they can help prevent road safety situations from occurring (Nyce, 2007). A commonly accepted definition for road accidents is the following:

"a rare, random, multi-factor event which is always preceded by a situation in which one or more road users have failed to cope with their environment" (Baguley, 2001).

From this definition, it can be concluded that road accidents are rare events in time. In fact, road accidents have the characteristics of random events (David & Branche, 2004) which mean that they cannot be easily predicted. Accidents have many contributory factors such as driver behaviour, vehicle condition, road or environmental conditions as outlined in Fig. 1.1 (The World Health Organization, 2004, p. 71).



Factors influencing exposure to risk
- Economic factors, including social deprivation
- Demographic factors
- Land use planning practices which influence the length of a trip or travel mode choice
- Mixture of high-speed motorized traffic with vulnerable road users
- Insufficient attention to integration of road function with decisions about speed limits, road layout and design

Risk factors influencing crash involvement
- Inappropriate or excessive speed
- Presence of alcohol, medicinal or recreational drugs
- Fatigue
- Being a young male
- Being a vulnerable road user in urban and residential areas
- Travelling in darkness
- Vehicle factors – such as braking, handling and maintenance
- Defects in road design, layout and maintenance which can also lead to unsafe road user behaviour
- Inadequate visibility due to environmental factors (making it hard to detect vehicles and other road users)
- Poor road user eyesight

Risk factors influencing crash severity
- Human tolerance factors
- Inappropriate or excessive speed
- Seat-belts and child restraints not used
- Crash helmets not worn by users of two-wheeled vehicles
- Roadside objects not crash protective
- Insufficient vehicle crash protection for occupants and for those hit by vehicles
- Presence of alcohol and other drugs

Figure 1. 1 The main risk factors for road traffic accidents

Source: (The World Health Organization, 2004, p. 71)

To improve road safety, factors which cause road safety issues in particular countries or regions should be identified (Hermans, et al., 2009). Although individual accidents cannot be predicted, by identifying and predicting the causes of accidents, appropriate counter measures can be put in place to target the contributory factors. It is important that road safety policies are not anecdotal and instead based on robust analysis and interpretation of data and consideration must be given to the application of local solutions based on local knowledge (The World Health Organization, 2004, p. 25). The five E's of road safety improvement are education, enforcement, engineering, encouragement and evaluation as described in Fig. 1.2 (Abugessaisa, 2008). Following evaluation, policy makers can focus efforts on preventing accidents by targeting safety awareness campaigns at high risk groups, deploying limited policing resources to high risk areas and allocating funding to infrastructure improvements.

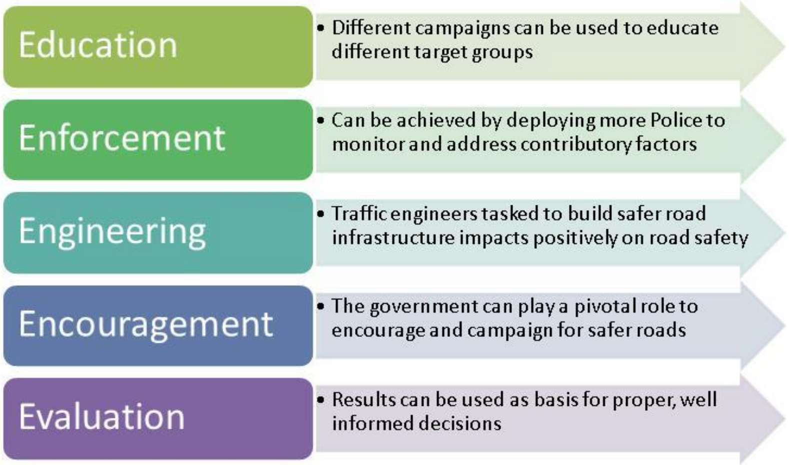

Education
• Different campaigns can be used to educate different target groups
Enforcement
• Can be achieved by deploying more Police to monitor and address contributory factors
Engineering
• Traffic engineers tasked to build safer road infrastructure impacts positively on road safety
Encouragement
• The government can play a pivotal role to encourage and campaign for safer roads
Evaluation
• Results can be used as basis for proper, well informed decisions

Figure 1. 2 The 5 E's of Road Safety Improvement

Source: (Abugessaisa, 2008)

Fig. 1.3 outlines Baguley's general framework of road safety improvement achieved by either accident prevention or reducing the cause of accidents, with an accident database at the centre of planning and evaluation. In order for road safety efforts to be effective they should be based on evidence, sustainable, properly resourced and the cost considered (The World Health Organization, 2004, p. 12).

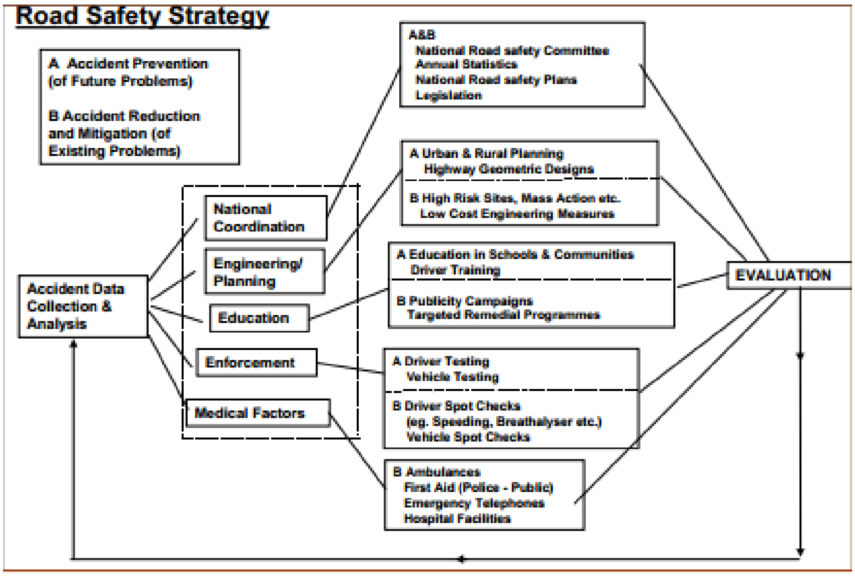

Road Safety Strategy
A Accident Prevention (of Future Problems)
B Accident Reduction and Mitigation (of Existing Problems)
A&B
National Road safety Committee
Annual Statistics
National Road safety Plans
Legislation
A Urban & Rural Planning
Highway Geometric Designs
B High Risk Sites, Mass Action etc.
Low Cost Engineering Measures
National Coordination
Engineering/ Planning
A Education in Schools & Communities
Driver Training
B Publicity Campaigns
Targeted Remedial Programmes
EVALUATION
Accident Data Collection & Analysis
Education
Enforcement
Medical Factors
A Driver Testing
Vehicle Testing
B Driver Spot Checks
(eg. Speeding, Breathalyser etc.)
Vehicle Spot Checks
B Ambulances
First Aid (Police - Public)
Emergency Telephones
Hospital Facilities

Figure 1. 3 A framework for road safety improvement

Source: (Baguley, 2001)

In order to prevent and reduce the causes of accidents, sharing accident data from various sources is vital. Road safety data is collected by different agencies, e.g.

hospitals and police. Road safety information is conducive to sharing due to the features it holds (Mitchell, 2002). In the UK, police complete a standard STATS19 accident report form for each road traffic accident reported to them. This data is available on the UK open government website for the period 1979 to 2013.[1] STATS19 data can answer questions such as where, when and what type of injury occurred, the consequence of a collision as well as the environmental conditions.

To improve road safety, a prerequisite is that information is available about accidents, fatalities, injuries and roads (Abugessaisa, 2008, p. 9). Many countries experience problems defining the accident information, collecting the information, maintaining quality and ensuring completeness (Abugessaisa, 2008, p. 9). In the UK, fatalities are known to be recorded accurately, however, under reporting of non-fatalities is a significant issue (The International Transport Forum, 2013). Two notable characteristics of road safety data is that their sources vary and they suffer from under reporting issues (Abugessaisa, 2008, p. 10). Extracting data from different data sources, verification of data and harmonising into a consistent format are time consuming tasks. Time spent on data integration is time which cannot be spent analysing road safety situations and thus helping to prevent road safety issues. To help address data integration issues, well defined methods should be adopted by road safety experts. STRADA, the Swedish traffic accident data management system, is used by police and hospitals to coordinate accident reports and aims to make road traffic accident details reliable and consistent and harmonise data. By bringing together data, the volume of data available on road traffic injuries and accidents increased and the number of unrecorded incidents reduced (Abugessaisa, 2008, pp. 30-33). The availability, quality, reliability and accuracy of relevant data would seem to be paramount to a predictive road safety strategy. (Nyce, 2007, p. 2) stated:

> *"the validity of any predictive model depends on the quality and quantity of the data available to develop it".*

Data on road traffic accidents is not consistently collected and harmonised into databases in many countries (The International Transport Forum, 2013). Similarly, data on accidents caused by environmental, technical and other factors is not well captured. Information available in databases such as city event calendars and weather conditions can provide additional awareness around the events that lead to road safety issues. The more data that is available, the more opportunity there is to identify factors which might influence road safety issues. A prominent issue is the lack of available government policy to facilitate the sharing of data between government agencies. This is a significant predictive modelling issue as data which might improve the accuracy of

[1] Department of Transport UK, 2014. *Road Safety Data*, http://data.gov.uk/dataset/road-accidents-safety-data, [Accessed 26 10 2014].

a predictive model is not available arising from data sharing issues.[2] This impacts road safety policy makers capability to improve road safety polices and accurately monitor performance. Recent initiatives to publicly share road safety and weather data would seem to be a move in the right direction.[3] By releasing high quality and diverse data to the public, crowd sourcing could use predictive analytics to help improve road safety. Policy makers use road safety performance indicators to measure road safety effectiveness. They provide a method to characterise the safety quality of road safety components (Abugessaisa, 2008, p. 22).

In IRTAD countries, between 2000 and 2010, fatal road traffic accidents have reduced substantially mainly due to improved safety features in cars and sustained anti drink driving campaigns (The International Transport Forum, 2013). Road traffic accidents relating to vulnerable road users such as pedestrians and cyclist have reduced but the reduction was smaller than that recorded for vehicle occupants. It may be concluded that significant improvements have been achieved, however, there is no time for complacency as the World Health Organisation (WHO) estimate that approximately 1.24 million people will lose their lives each year as a result of road traffic accidents with vulnerable road users making up half of those who die (The World Health Organisation, 2013).

A key requirement in the data mining and predictive analytics process is an understanding of the data. Inaccurate or missing data impact on the quality of the prediction which can be achieved. Given that fatal road traffic accidents are low frequency events, when considered in the context of all data recorded for road traffic accidents, this would be a key consideration when creating a predictive model as it could pose problems in terms of acquiring an adequate sample size to make the data amenable to prediction.

Data mining tools and techniques can be used to predict future events and trends which allow proactive and knowledge driven decisions. A part of the data mining process includes using machine learning techniques to find patterns and relationships in data (Miner, et al., 2009, p. 17). Examples of modelling techniques include decision trees and Bayes net. Many modelling techniques produce a propensity score which is a

[2] Travis, A., 2012, '*Government revives plan for greater data-sharing between agencies',The Guardian, 24 May.* Available at: http://www.theguardian.com/politics/2012/apr/23/government-plan-share-personal-data, [Accessed 01 12 2014].

[3] UK Government, 2014. *UK open data portal.* http://data.gov.uk/, [Accessed 1 12 2014].

number in the range zero to one which indicates the likelihood of the event modelled occurring given a set of predictors. The score is ordered descending from highest to lowest with higher scores indicating that the event is more likely to occur. When approaching a data mining experiment, adoption of a methodology provides structure and best practice to the process. CRISP-DM is an industry accepted methodology which outlines six phases to a data mining project.

Predictive analytics relates to a broad field which applies statistical and analytical techniques to build predictive models to identify future events or behaviours (Nyce, 2007, p. 1). Predictive analytics is defined as a

> *"set of business intelligence (BI) technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behaviour and events. Unlike other BI technologies, predictive analytics is forward-looking, using past events to anticipate the future"* (Eckerson, 2007, p. 5).

The main component of a predictive analytics technique is the predictor. These are variables that can be measured to predict future behaviour. Predictive analytics tools include mathematical algorithms as well as machine learning and statistical methods. These are very effective in terms of overcoming manual searching of data. Examples of modelling techniques include clustering (McCue, 2007, p. 51), supervised learning (Chong, et al., 2005) and time series analysis (Monfared, et al., 2013).

This research focuses on three classification techniques, C5.0 and Chaid decision trees and Bayes net. C5.0 is a supervised learner developed by Ross Quinlan to build decision trees using the concept of information gain. It works by splitting the data based on the field that provides the most information gain. Each subsample defined by the split is split again based on the next most important field. This process continues until the subsamples cannot be split any more. Finally, the lowest splits in the decision tree, which provide the least information, are removed. Similar to C5.0, Chaid is a supervised learning algorithm used for classification. Chaid (Miner, et al., 2009, p. 246) stands for Chi-squared automatic interaction detection. It was proposed by Kass in 1980. The splitting mechanism is specific to Chaid. Chaid uses the Chi-squared statistical test for proportion to determine a split. Chaid uses multiway splits (Miner, et al., 2009, p. 246) to construct trees and has a stopping mechanism which determines when a sub tree is complete (Miner, et al., 2009, p. 792). A Bayesian network (Bayes net) is a probabilistic graphical modelling technique used to represent knowledge about an uncertain domain such as traffic accidents (Simoncic, 2004). Bayes net can be used to classify a target variable such as fatal traffic accident. The network represents a set of random variables and their conditional dependencies. In the network, nodes represent random variables and edges represent the conditional dependencies among random variables.

1.3 Research Problem

(Tesema, et al., 2005) stated that we are drowning in data, yet knowledge of the factors which contribute to road traffic accidents fatalities seem to be limited. The STATS19 data is used extensively for road traffic accident statistics reporting in the UK and records accident related features in a consistent and relatively complete fashion. Predictive analytics would seem to be suitable for sifting through the data to identify useful patterns which could help predict road fatality risk. Fatal traffic accidents are considered infrequent events which adds to the prediction difficulty.

The research problem addressed in this dissertation is whether three selected classification techniques, C5.0 and Chaid decision trees and Bayes net, can predict fatal road traffic accidents based on a STATS19 UK Road Safety dataset and whether key contributory factors to fatal road traffic accidents can be extracted from the models.

1.4 Research Aims and Objectives

This research aims to:

- Apply three predictive modelling techniques, C5.0 and Chaid decision trees and Bayes net, to build predictive models to classify fatal road traffic accidents.
- Evaluate the technical and non-technical performance of the best prediction models.
- Identify the key contributory factors of fatal traffic accidents from the predictive models.

The objectives of the research are to:

- Review academic literature for road traffic accidents
- Review literature for data mining, predictive analytics and evaluation techniques and current research specific to traffic accident prediction
- Understand the data and prepare the data for modelling
- Design the experiment to build and evaluate models for C5.0 and Chaid decision trees and Bayes net to predict fatal traffic accidents
- Conduct the experiment to classify fatal traffic accidents and assess and evaluate the models
- Extract key findings from the experiment and the key contributory factors identified from the models
- Suggest future work which could expand or enhance the experiment findings.

Given fatal traffic accidents are a significant cause of preventable death, it is hoped that this research will make a contribution to the existing body of knowledge.

1.5 Research Methodology

An overall design for the experiment will be prepared, following a recognised methodology to ensure a reliable and repeatable process is adopted. The design will include defining and preparing training data, test data and validation data, building models using selected modelling techniques, assessing model performance as well as model evaluation. IBM SPSS Modeler, a leading commercial predictive modelling tool, will be used to build the predictive models.

The Cross Industry Standard Process for Data Mining (CRISP-DM) will be used to guide the modelling process. CRISP-DM, as outlined in Fig. 1.4, is a standard process used to implement a predictive analytics and data mining solution (McCue, 2007, p. 49). An adaptation of this methodology, further discussed in chapter 3, was applied for the experiment to align to the specific experiment requirements.

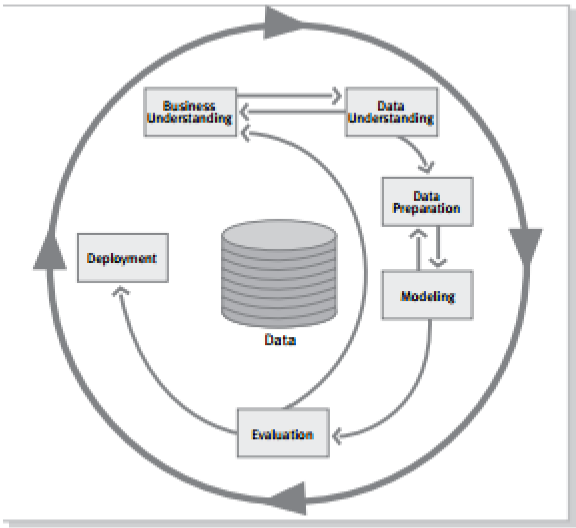


Figure 1. 4 CRISP-DM Process
Source: (Chapman P., et al, 2000)

Business understanding was derived mainly from literature review. Data understanding was based on data exploration and analysis using SPSS Modeler and Toad for Oracle database. The dataset selected for the experiment was the STATS19 UK Road Safety dataset, which is publicly available and is often used for academic research and the most commonly used source of UK road safety statistics. The literature suggests that the STATS19 data is well recorded for fatal accidents[4], it was hoped that the data

[4] Department of Transport UK, 2014, *STATS19 Road Accident dataset.* http://www.adls.ac.uk/department-for-transport/stats19-road-accident-dataset/?detail, [Accessed 03 10 2014].

collected could be used to identify contributory factors for fatal accidents. Data preparation involved data selection, reduction and construction. Modelling stage included selection and design of the modelling techniques, building models in line with design and assessment of models. Evaluation was the final stage of the experiment where unseen data was scored against the models.

1.6 Scope and Limitations

The scope of the dissertation was to design, build and implement three classification models which can effectively predict fatal road traffic accidents and to identify key fatal accident contributory factors identified by the models. The data source was the UK STATS19 road safety dataset which records all reported road traffic accidents. The UK dataset was selected due to the reported quality and completeness of the data and also the volume of fatal road traffic accidents in the UK should be sufficient to extract meaningful prediction.

The focus of the research was on prediction of fatal accidents. Modelling and sampling techniques were selected and applied to improve the likelihood of fatal accident prediction. As fatal traffic accidents were infrequent events in the STATS19 dataset, sampling techniques were used to improve fatal accident recall. However, as a result precision and accuracy performance were expected to be negatively impacted. Data preparation focussed on fatal accidents only which may also impact on accuracy performance if non-fatal data is misclassified where features for fatal and non-fatal are similar. This limitation means some performance metrics were expected to be low. This research was completed without the assistance of a subject matter expert so data exploration and preparation was completed using SPSS Modeler and Toad for Oracle and using Chaid for data reduction. This limitation meant that business understanding was not applied to the research and an informed review of data preparation prior to model build may have provided more meaningful data groups and therefore data relationships. The original experiment limits model builds to twenty seven, for each classification model, three modelling techniques and three sampling techniques were selected. Five additional models were built post evaluation to test if accuracy performance could be improved. The limitation of model numbers was necessary in order to ensure the research was completed within the time and project size constraints.

1.7 Organisation of the Dissertation

Following this brief introduction chapter, the remaining chapters describe in more detail the literature review completed for road traffic accidents and data mining and prediction techniques, the experiment design, implementation, evaluation and research conclusions.

Chapter two provides a description of the relevant literature on road traffic accidents. The chapter provides an overview of road traffic accident environment, the role of road safety data, UK accident statistics and contributory factors and the UK road traffic accident data. The chapter also introduces relevant data mining and predictive analytics research literature. An overview of data mining and the key data considerations is provided, together with relevant data sampling techniques. Predictive analytics, specifically focussed on classification techniques, is outlined and assessment and evaluation techniques relevant to this research are discussed. Current traffic accident prediction research is briefly discussed.

Chapter three proposes the research experiment design, including the implementation methodology, the key requirements for data understanding and data preparation, the model build and approach to model evaluation.

Chapter four describes the experiment implementation stages in line with the experiment design outlined in chapter 3. This chapter discusses each stage in more detail and provides details about the models built.

Chapter five outlines the experiment evaluation including the assessment of the model performance for training and test data and the technical and non-technical evaluation against validation data. Two suggested model improvements are evaluated and the key findings are summarised.

Chapter six provides a summary of the research completed, contributions to the body of knowledge and the experiment evaluation and limitations. Future work considerations are suggested.

1.8 Conclusion

The introduction provides an outline for the research experiment. The background to road traffic accidents and predictive analytics and classification techniques are introduced. The research problem to be addressed as part of this research and the main aims and objectives are presented. The planned research methodology is outlined, together with the scope of the research and the key limitations identified. Finally the structure of the research is summarised by chapter to provide an outline of the subsequent research.

2 LITERATURE REVIEW

2.1 Introduction

This chapter discusses current literature relating to road traffic accidents, data mining and prediction techniques relevant to this research. An overview of road traffic accidents is provided to frame an understanding of the relevant subject matter. The overview is followed by a brief discussion of the role of road safety data, road traffic accident data as well as information sharing. Relevant road traffic accident statistics for the UK are presented together with a description of contributing factors as identified in current literature.

Data mining is a broad term used to describe a variety of statistical and machine learning techniques used to extract knowledge from data. Data mining techniques can be applied to interrogate data and identify underlying trends allowing for the development of models aimed at predicting future events. The literature review focuses particularly on techniques which are of relevance to the research experiment and applicable techniques which could be applied to the prediction of fatal accidents. An overview of CRISP-DM is provided as this is a widely used methodology for data mining and predictive analytics.

Predictive analytics uses data mining and machine learning techniques to predict future events or behaviours. Classification techniques extract prediction by applying machine learning techniques and identifying relationships in data and grouping into classes. An overview of classification techniques including the algorithms selected for the experiment, C5.0 and Chaid decision trees and Bayes net, is provided.

Relevant evaluation techniques are discussed which are used to evaluate the performance of models in this experiment. Techniques include the confusion matrix, receiver operator curve (ROC), area under the curve as well as model interpretability. Four academic papers are briefly discussed outlining current research in the field of traffic accident prediction. The results achieved by (Wah, et al., 2012) in predicting fatal traffic accidents led to consideration of this research problem.

2.2 Road Traffic Accident Overview

It is estimated that more than a million people die from injuries sustained on the world's roads annually and road fatalities are ranked eighth as the cause of deaths globally (The World Health Organisation, 2013, p. 1). The consequential socio-economic impact of dealing with road traffic accidents is estimated to run into the billion's with young people aged 15 – 29 representing the largest proportion of casualties (The World Health Organisation, 2013, p. 1). The costs of road traffic injuries are staggering (David & Branche, 2004) and include, but are not limited to, ambulance, hospital care, earnings lost as well as lifestyle disruption and emotional costs. Traffic accident injury is considered largely predictable and preventable (The World Health Organization, 2004, p. 25). For example, remedial and inexpensive interventions can be undertaken such as removing overgrown hedging which obscure stop signs (The Irish Road Safety Authority, 2013, p. 36). In recent years, countries in the developed world have reduced road traffic accidents by adopting road safety strategies and enforcing legislation to address some key risks such as speed, drink driving and seat belt wearing, however, it is noted that encouraging a safe road culture requires persistent effort (The World Health Organisation, 2013, p. 12). Fig. 2.1 presents the worldwide increase in comprehensive legislation enforced to target the key five road risk factors.

Road traffic deaths each year have not increased, however, the volume of approx. 1.24 million remains *"unacceptably high"* (The World Health Organisation, 2013, p. 4) and more action is needed to further reduce road traffic accidents. Although many useful strategies exist to address road safety behaviour they could be more widely implemented (David & Branche, 2004). 2012 was an important milestone for the OECD-IRTAD with many countries recording their lowest fatality rate on record (The International Transport Forum, 2013, p. 9). In order to achieve the 2020 targets set by the UN, to halve the fatality rate worldwide, improved road safety strategies will need to be adopted by those countries trailing behind the trend.

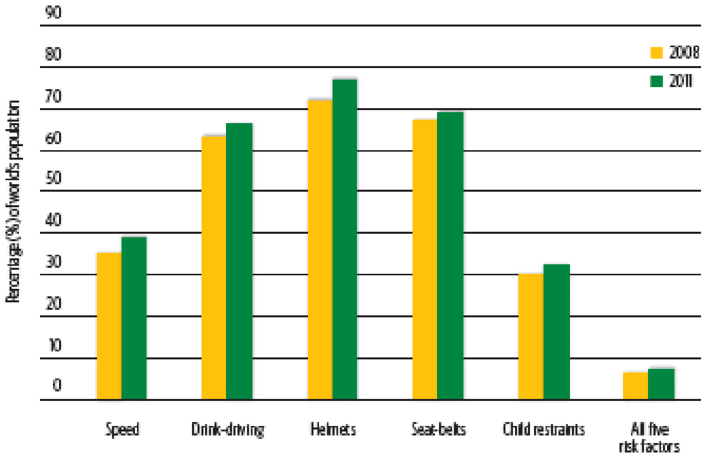


Figure 2. 1 World population covered by legislation for five key road risk factors

Source: (The World Health Organisation, 2013, p. 12)

Fig. 2.2 displays the average fatality by road user from 2007 – 2011. Pedestrians and cyclists represent a neglected group of road users which comprise 27% of road traffic fatalities worldwide (The World Health Organisation, 2013, p. v). Recently, at the Rio+20 UN conference on sustainable development, a link was established between road safety and sustainable development. Road safety policy must now focus on increasing the safety of non-motorised road users by protecting them from high speed traffic which is in line with a sustainable transport policy (The World Health Organisation, 2013, p. v).

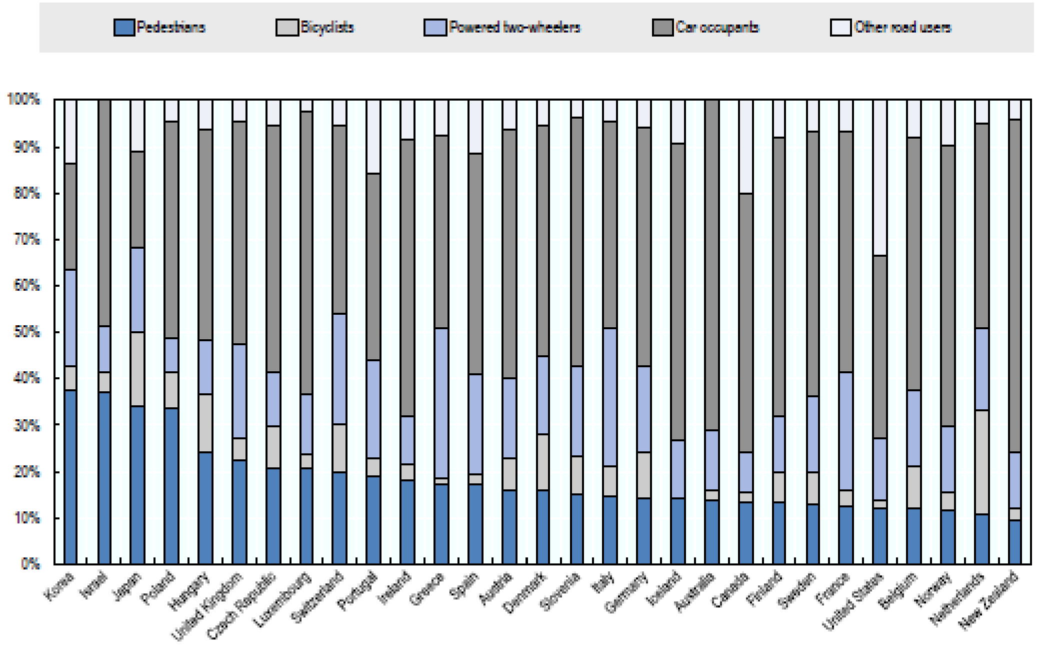
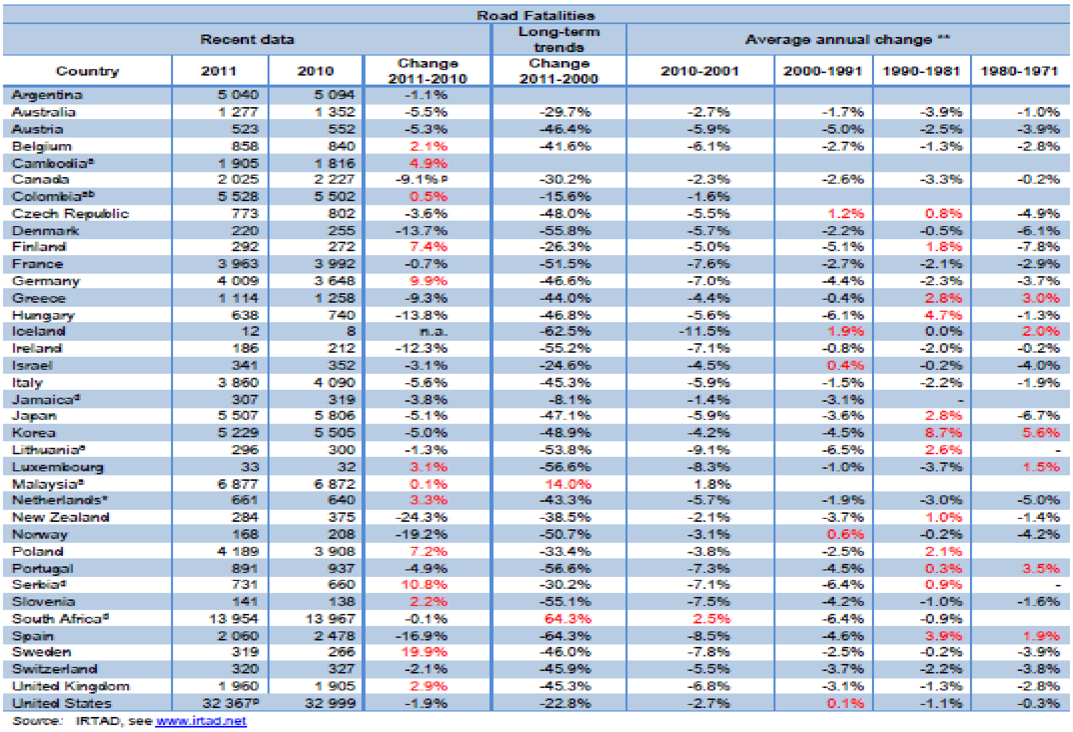


Figure 2. 2 Fatalities by road user class (average 2007 - 2011)

Source: (The International Transport Forum, 2013, p. 8)

Figure 2.3 shows that the average annual reduction in fatalities was higher in the last decade than any of the previous three decades for most IRTAD countries. This evidence supports the assertion that the implementation of road safety strategies has produced good results over the long term and the effective implementation of road safety policies would seem to be crucial. The safety of vehicle occupants has improved substantially over the last decade through public campaigns such as drink driving awareness and speed reduction programs (The Irish Road Safety Authority, 2013). National television networks have been used to graphically illustrative the consequences of bad driving practice. According to UTV News, some cohorts have strongly objected to the graphic nature of these accounts [5] but others maintain that this is what is required to deliver a compelling message to young people who are the largest casualty group in the OECD (The World Health Organisation, 2013, p. vii).

[5] UTV News, 2014, '*Shocking NI road safety ad goes viral',* Available at: http://www.u.tv/News/Shocking-NI-road-safety-ad-goes-viral/81cf1549-f38a-4d28-a274-0060a6b2c43c, [Accessed 23 09 2014].



Road Fatalities									
Recent data				Long-term trends	Average annual change **				
Country	2011	2010	Change 2011-2010	Change 2011-2000	2010-2001	2000-1991	1990-1981	1980-1971
Argentina	5 040	5 094	-1.1%					
Australia	1 277	1 352	-5.5%	-29.7%	-2.7%	-1.7%	-3.9%	-1.0%
Austria	523	552	-5.3%	-46.4%	-5.9%	-5.0%	-2.5%	-3.9%
Belgium	858	840	2.1%	-41.6%	-6.1%	-2.7%	-1.3%	-2.8%
Cambodia[a]	1 905	1 816	4.9%					
Canada	2 025	2 227	-9.1% p	-30.2%	-2.3%	-2.6%	-3.3%	-0.2%
Colombia[ab]	5 528	5 502	0.5%	-15.6%	-1.6%			
Czech Republic	773	802	-3.6%	-48.0%	-5.5%	1.2%	0.8%	-4.9%
Denmark	220	255	-13.7%	-55.8%	-5.7%	-2.2%	-0.5%	-6.1%
Finland	292	272	7.4%	-26.3%	-5.0%	-5.1%	1.8%	-7.8%
France	3 963	3 992	-0.7%	-51.5%	-7.6%	-2.7%	-2.1%	-2.9%
Germany	4 009	3 648	9.9%	-46.6%	-7.0%	-4.4%	-2.3%	-3.7%
Greece	1 114	1 258	-9.3%	-44.0%	-4.4%	-0.4%	2.8%	3.0%
Hungary	638	740	-13.8%	-46.8%	-5.6%	-6.1%	4.7%	-1.3%
Iceland	12	8	n.a.	-62.5%	-11.5%	1.9%	0.0%	2.0%
Ireland	186	212	-12.3%	-55.2%	-7.1%	-0.8%	-2.0%	-0.2%
Israel	341	352	-3.1%	-24.6%	-4.5%	0.4%	-0.2%	-4.0%
Italy	3 860	4 090	-5.6%	-45.3%	-5.9%	-1.5%	-2.2%	-1.9%
Jamaica[d]	307	319	-3.8%	-8.1%	-1.4%	-3.1%	-	
Japan	5 507	5 806	-5.1%	-47.1%	-5.9%	-3.6%	2.8%	-6.7%
Korea	5 229	5 505	-5.0%	-48.9%	-4.2%	-4.5%	8.7%	5.6%
Lithuania[a]	296	300	-1.3%	-53.8%	-9.1%	-6.5%	2.6%	-
Luxembourg	33	32	3.1%	-56.6%	-8.3%	-1.0%	-3.7%	1.5%
Malaysia[a]	6 877	6 872	0.1%	14.0%	1.8%			
Netherlands*	661	640	3.3%	-43.3%	-5.7%	-1.9%	-3.0%	-5.0%
New Zealand	284	375	-24.3%	-38.5%	-2.1%	-3.7%	1.0%	-1.4%
Norway	168	208	-19.2%	-50.7%	-3.1%	0.6%	-0.2%	-4.2%
Poland	4 189	3 908	7.2%	-33.4%	-3.8%	-2.5%	2.1%	
Portugal	891	937	-4.9%	-56.6%	-7.3%	-4.5%	0.3%	3.5%
Serbia[d]	731	660	10.8%	-30.2%	-7.1%	-6.4%	0.9%	-
Slovenia	141	138	2.2%	-55.1%	-7.5%	-4.2%	-1.0%	-1.6%
South Africa[d]	13 954	13 967	-0.1%	64.3%	2.5%	-6.4%	-0.9%	
Spain	2 060	2 478	-16.9%	-64.3%	-8.5%	-4.6%	3.9%	1.9%
Sweden	319	266	19.9%	-46.0%	-7.8%	-2.5%	-0.2%	-3.9%
Switzerland	320	327	-2.1%	-45.9%	-5.5%	-3.7%	-2.2%	-3.8%
United Kingdom	1 960	1 905	2.9%	-45.3%	-6.8%	-3.1%	-1.3%	-2.8%
United States	32 367[p]	32 999	-1.9%	-22.8%	-2.7%	0.1%	-1.1%	-0.3%

Source: IRTAD, see www.irtad.net

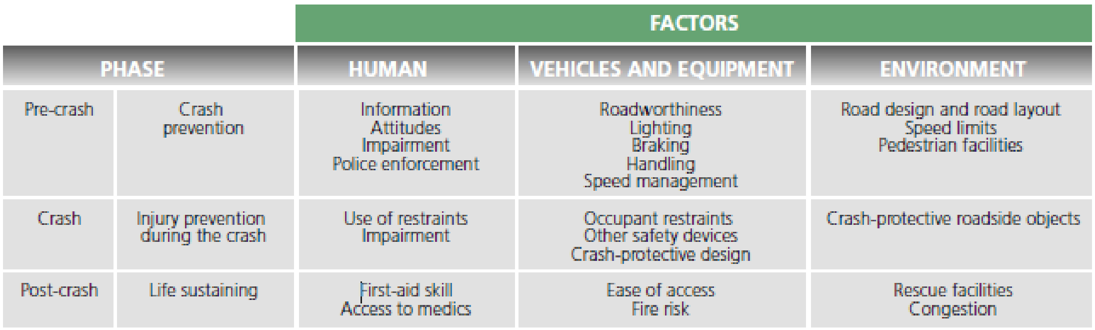
Figure 2. 3 Trends in road fatalities

Source: (The International Transport Forum, 2013, p. 12)

Approaches such as legislative enforcement and police checkpoints have proven to be effective (The Police Chief, 2005). Road users are now encouraged to consider and reflect on their road usage behaviour by employing sophisticated road safety advertisements and mass media campaigns (The Irish Road Safety Authority, 2013).

Attitudes to road safety and road user behaviour differ greatly worldwide. Cultural considerations need to be factored into road safety planning and actions to prevent road traffic accidents should be tested locally (The World Health Organization, 2004, p. 162).

The WHO recommends that road safety agencies should be appointed in each country and should be given decision making authority to co-ordinate road safety efforts and resources (The World Health Organisation, 2013, p. 27). For example semi-state bodies such as the Irish Road Safety Authority play a key role educating the public about road safety. Unless urgent action is taken, current research suggests that by 2030, road traffic accidents will become the fifth leading cause of death world wide (The World Health Organisation, 2013, p. vii). A systems approach has been proposed as a necessary tool to effectively prevent road traffic injury. Haddon's matrix, as outlined in Fig. 2.4, has been useful in the development of strategies and techniques (The World Health Organization, 2004, p. 12).

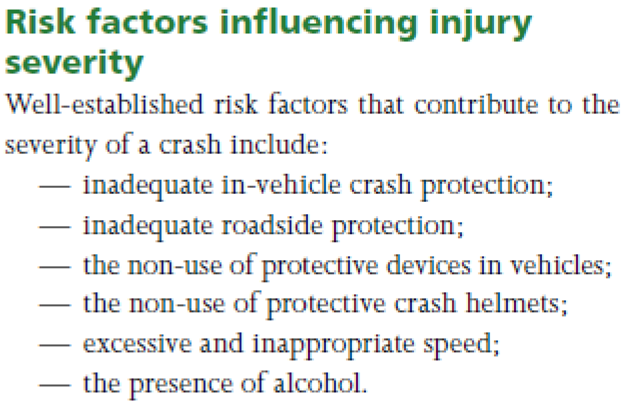


		FACTORS		
PHASE		HUMAN	VEHICLES AND EQUIPMENT	ENVIRONMENT
Pre-crash	Crash prevention	Information Attitudes Impairment Police enforcement	Roadworthiness Lighting Braking Handling Speed management	Road design and road layout Speed limits Pedestrian facilities
Crash	Injury prevention during the crash	Use of restraints Impairment	Occupant restraints Other safety devices Crash-protective design	Crash-protective roadside objects
Post-crash	Life sustaining	First-aid skill Access to medics	Ease of access Fire risk	Rescue facilities Congestion

Figure 2. 4 Haddon Matrix - A Systems Approach

Source: (The World Health Organization, 2004, p. 13)

The WHO and World Bank intensified work on road traffic injury prevention and prepared a detailed joint report which sought to describe patterns and impacts at a global and local level, review key risk factors and discuss intervention strategies (The World Health Organization, 2004, p. xx). Setting road safety targets has become an important part of national road safety strategies in many high-income countries. Governments are recommended to set interim targets to encourage public and political support for long term strategies but collection of data is key (The World Health Organisation, 2013, p. 27). There is strong scientific evidence available which supports the claim that adopting intervention's, such as creating, adopting and enforcing legislation relating to key risk factors such as drink-driving, speed and wearing of seat belts, leads to a reduction in road traffic injury (The World Health Organisation, 2013, p. v). If action is taken, many lives can be saved and the evidence would seem to suggest that improvements can be achieved by taking simple measures (The Irish Road Safety Authority, 2013, p. 36). The WHO and World Bank joint report identified that there are well established risk factors which influence the severity of a road traffic accidents as summarised in Fig 2.5 (The World Health Organization, 2004, p. 88).



Risk factors influencing injury severity

Well-established risk factors that contribute to the severity of a crash include:

— inadequate in-vehicle crash protection;
— inadequate roadside protection;
— the non-use of protective devices in vehicles;
— the non-use of protective crash helmets;
— excessive and inappropriate speed;
— the presence of alcohol.

Figure 2. 5 Risk factors influencing road traffic injury severity

Source: (The World Health Organization, 2004, p. 88)

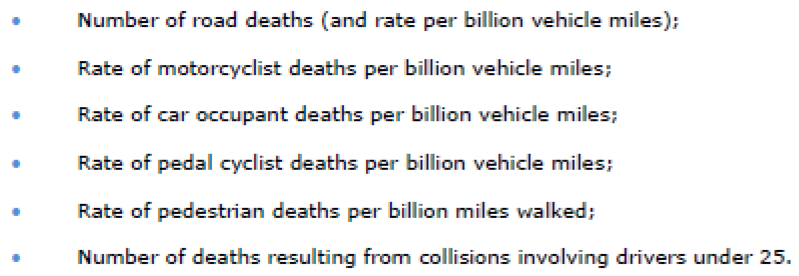
2.3 The Role of Road Safety Data

In order to improve road safety and reduce fatalities, high quality, reliable and consistent information relating to accident circumstances as well as vehicle and casualty details should be made available to road safety professionals.

"Only by systematic and data-led management of the leading road injury problems will significant reductions in exposure to crash risk and in the severity of crashes be achieved" (The World Health Organization, 2004, p. 8).

The availability of traffic accident data will enable road safety professionals to accurately assess the current situation and propose appropriate counter measures to reduce the likelihood of road safety situations. Data driven decisions taken, following analysis, is a function of data quality. The higher the data quality, the more targeted the corrective actions can be (Abugessaisa, 2008, p. 11). (Abugessaisa, 2008, p. 10) noted that road safety data has two notable characteristics. The first is that not all traffic accidents are reported. There may be a record of an injury at a hospital or insurance claim at an insurance company but no official road traffic accident record with the police. Secondly the consistency and accuracy of road traffic accident data sources vary. Data consumers find themselves needing to analyse different sources to materialise a consistent and accurate view of events. It would seem that there are data quality issues with road safety data (Abugessaisa, 2008, p. 11) and the data owners in many jurisdictions may need to be educated on the important role of road safety data. Consideration should be given to how data is gathered, organised and analysed. According to (Baguley, 2001, p. 8), studies of hospital records have shown that road accidents are considerably under reported, although the level of reporting tends to be higher for more severe injuries. However, in the UK all fatal accidents are reported by the police (The International Transport Forum, 2013, p. 429). By involving all the key participants responsible for road safety and implementing safety measures systematically, road deaths and serious injuries can be avoided (The World Health Organization, 2004, p. 19). The participants include but are not limited to drivers, vehicle designers and manufacturers.

According to (Hermans, et al., 2009, p. 178), performance indicators representing road safety risk factors can be used to quantify road safety performance. Accident or injury safety performance indicators can be used to measure if actions are effective (The World Health Organization, 2004, p. 19). Indicators are needed by road safety planners as basic accident counts do not evaluate accidents in terms of costs which are critically important to society e.g. social cost. By evaluating accidents in terms of critical factors, performance indicators can be used to help legislators and road safety professionals identify sectors in road safety which are performing well and those which require attention. In the UK a new strategic framework was launched in May 2011 identifying six key road accident performance indicators as outlined in Fig 2.6.



- Number of road deaths (and rate per billion vehicle miles);
- Rate of motorcyclist deaths per billion vehicle miles;
- Rate of car occupant deaths per billion vehicle miles;
- Rate of pedal cyclist deaths per billion vehicle miles;
- Rate of pedestrian deaths per billion miles walked;
- Number of deaths resulting from collisions involving drivers under 25.

Figure 2. 6 UK six road traffic accident performance indicators
Source: (The International Transport Forum, 2014, p. 501)

2.4 UK Accident Statistics and Contributory Factors

According to (The International Transport Forum, 2014, p. 491), between 2000 – 2012, a fatality reduction rate of 50% was recorded in the UK and as in Fig. 2.7 the trend for road traffic deaths has been steadily falling since 2006 (The World Health Organisation, 2013, p. 225).

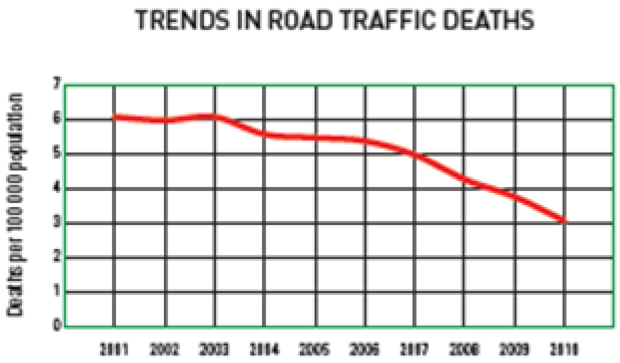


Figure 2. 7 Trends in UK road traffic accident deaths
Source: (The World Health Organisation, 2013, p. 225)

(The International Transport Forum, 2014, p. 490) states in 2013, 13% of the total 183,670 road casualties in the UK were killed or serious injury (KSI) casualties as displayed in Fig. 2.8. Although traffic flow increased in the period, there was a 2% decrease in the killed group. The reduction in accidents or fatal accidents was noted on all road types in 2013 when compared to 2012.

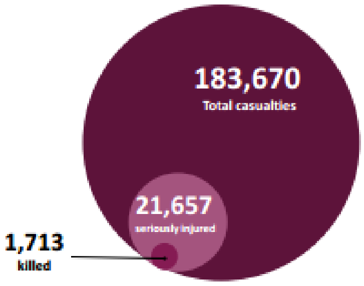


Figure 2. 8 KSI as a proportion of total casualties
Source: (Department of Transport UK, 2013)

Drivers in 4-wheeled and light vehicles are the highest proportion of road deaths in the UK in 2010 followed by pedestrians and motorbike riders as displayed in Fig. 2.9.

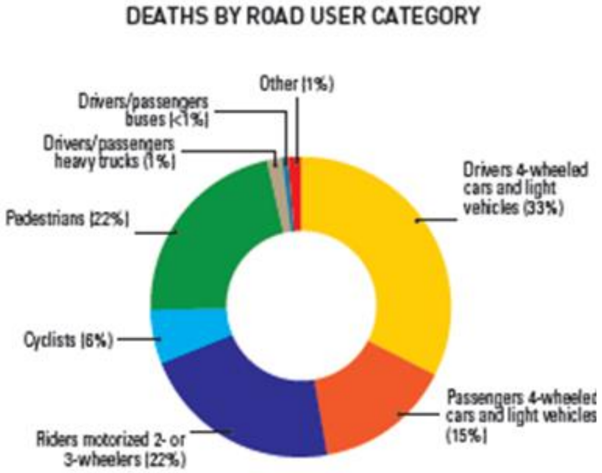


Figure 2. 9 Deaths by road user in 2010 for UK
Source: (The World Health Organisation, 2013, p. 225)

In 2013, the (Department of Transport UK, 2013) reported that most fatalities were car occupants and occurred on non-built up roads while most serious injuries occurred on built up roads as shown in Fig. 2.10.

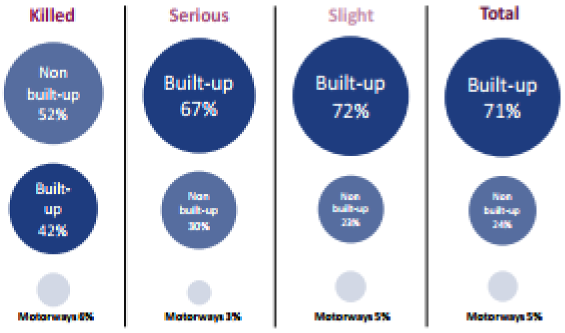


Figure 2. 10 Proportion of casualties types by motorway
Source: (Department of Transport UK, 2013)

At the end of 2013, there were 35 million vehicles licensed for driving in the UK (Grove, 2014) as outlined in Fig 2.11, this number has increased year on year for the last 10 years. Even with the increase in licensed vehicles, the fatality rate has reduced significantly over the last decade and in 2013, road safety incidents decreased again, with fatalities at their lowest levels since records began (Department of Transport UK, 2013).

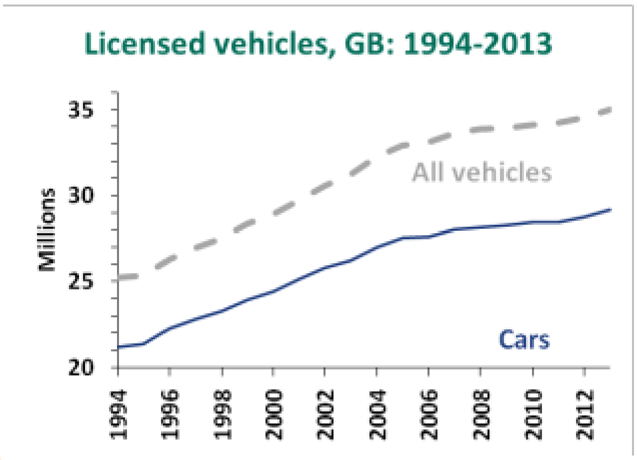


Figure 2. 11 Licensed vehicles in the UK
Source: (Grove, 2014)

Fig 2.12 describes the number of fatalities reported for the period 2005 – 2013. The 2012 – 2013 fatality count was 39% below the 2005 – 2009 average which would supports the claim that recent road safety strategies were effective (Department of Transport UK, 2013).

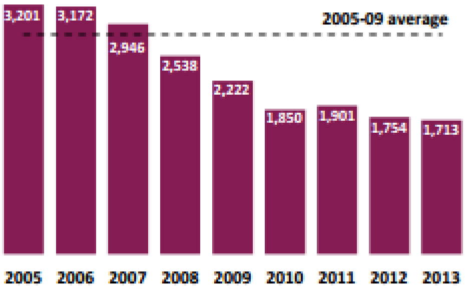


Figure 2. 12 Reduction in road traffic accident fatalities in recent years in the UK
Source: (Department of Transport UK, 2013)

The largest ever reduction in fatalities in the UK was observed in 2010 due to sustained periods of adverse snow and ice weather conditions (The International Transport Forum, 2013, p. 429). Environmental factors impact on road traffic accidents. (The International Transport Forum, 2014, p. 490) noted that in the first quarter of 2013 the weather was notably colder when compared with 2012, this was likely to have contributed to reduced casualties for pedal and motor cyclists and car occupants. There are various factors which might have contributed to this reduction including but not limited to improved vehicle safety, road engineering, hospital care and road safety education (The International Transport Forum, 2014, p. 491).

In the event of a road traffic accident, a number of characteristics are known to increase the risk of traffic accidents which include demographic, behavioural, environmental and vehicle (The World Health Organization, 2004). Demographic characteristics include age and address as well as the occupation of the driver. Behavioural characteristics include drug or alcohol taking while driving, seat belt usage, speed and fatigue. Environmental characteristics include road and visibility conditions as well as weather conditions. Vehicle characteristics include car class, age and engine size.

Contributory factors for road safety accidents are wide ranging. The factors identified are different depending on the particular characteristics being considered whether demographic, behavioural, environmental or vehicle. For example a behavioural contributory factor may be speed whereas as environmental factor may be road type. (The International Transport Forum, 2014) road safety annual report presents the key statistics relating to road safety accidents in the UK for 2013. Fig. 2.13 to Fig. 2.17 present the most recent statistics.

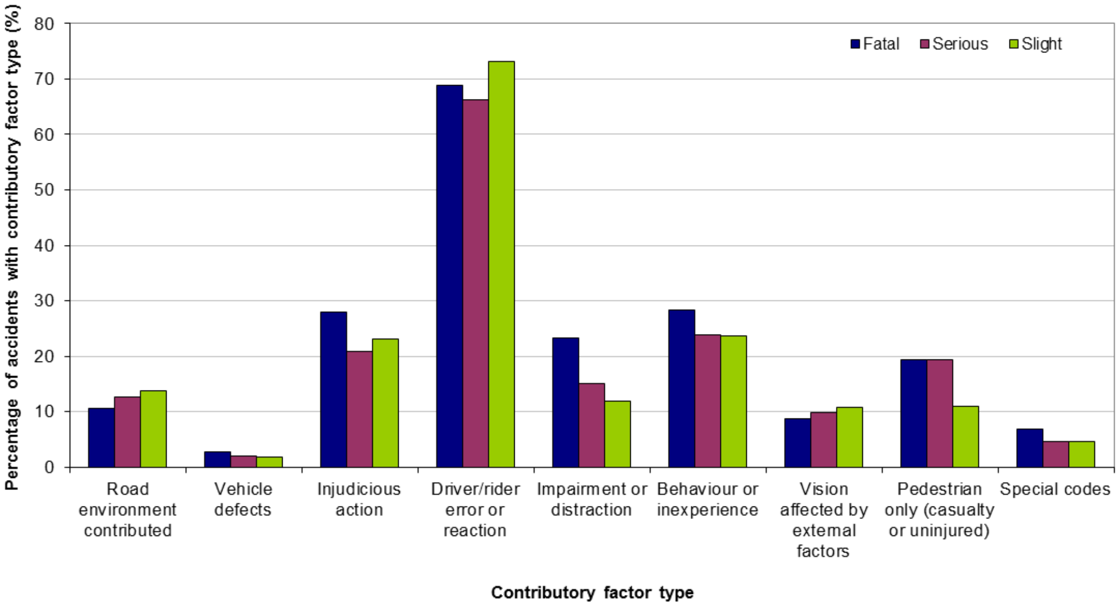
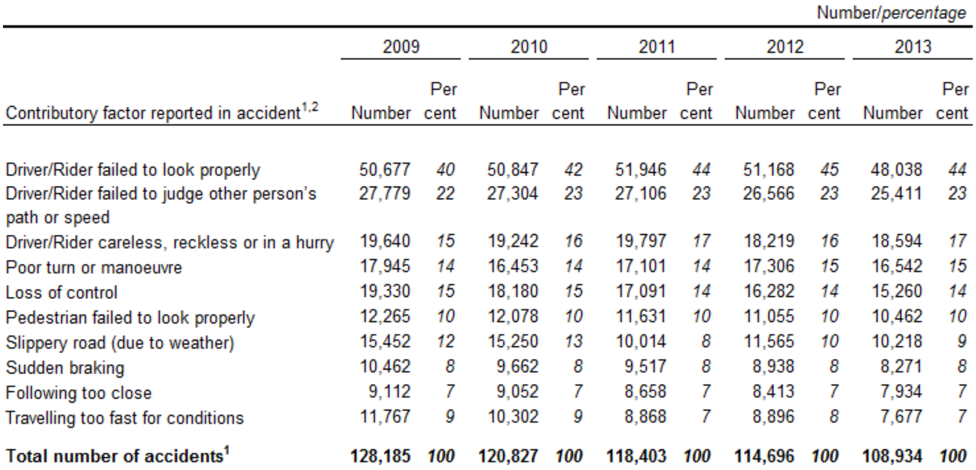


Figure 2. 13 UK 2013 contributory factor and severity[6]

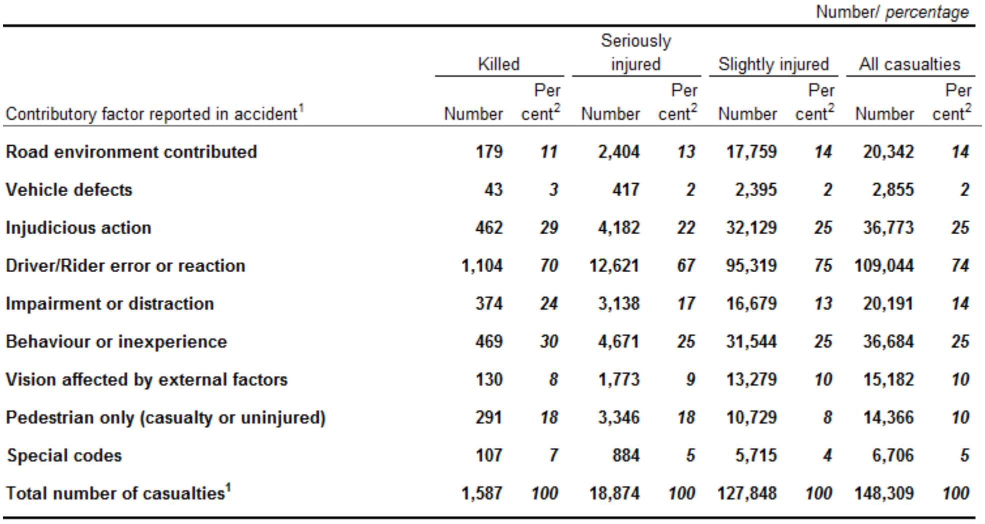
[6] Department of Transport UK, 2014. *Contributory factors for reported road accidents*, https://www.gov.uk/government/statistical-data-sets/ras50-contributory-factors, [Accessed 01 12 2014].



Number/*percentage*

Contributory factor reported in accident[1,2]	2009 Number	2009 Per cent	2010 Number	2010 Per cent	2011 Number	2011 Per cent	2012 Number	2012 Per cent	2013 Number	2013 Per cent
Driver/Rider failed to look properly	50,677	*40*	50,847	*42*	51,946	*44*	51,168	*45*	48,038	*44*
Driver/Rider failed to judge other person's path or speed	27,779	*22*	27,304	*23*	27,106	*23*	26,566	*23*	25,411	*23*
Driver/Rider careless, reckless or in a hurry	19,640	*15*	19,242	*16*	19,797	*17*	18,219	*16*	18,594	*17*
Poor turn or manoeuvre	17,945	*14*	16,453	*14*	17,101	*14*	17,306	*15*	16,542	*15*
Loss of control	19,330	*15*	18,180	*15*	17,091	*14*	16,282	*14*	15,260	*14*
Pedestrian failed to look properly	12,265	*10*	12,078	*10*	11,631	*10*	11,055	*10*	10,462	*10*
Slippery road (due to weather)	15,452	*12*	15,250	*13*	10,014	*8*	11,565	*10*	10,218	*9*
Sudden braking	10,462	*8*	9,662	*8*	9,517	*8*	8,938	*8*	8,271	*8*
Following too close	9,112	*7*	9,052	*7*	8,658	*7*	8,413	*7*	7,934	*7*
Travelling too fast for conditions	11,767	*9*	10,302	*9*	8,868	*7*	8,896	*8*	7,677	*7*
Total number of accidents[1]	**128,185**	***100***	**120,827**	***100***	**118,403**	***100***	**114,696**	***100***	**108,934**	***100***

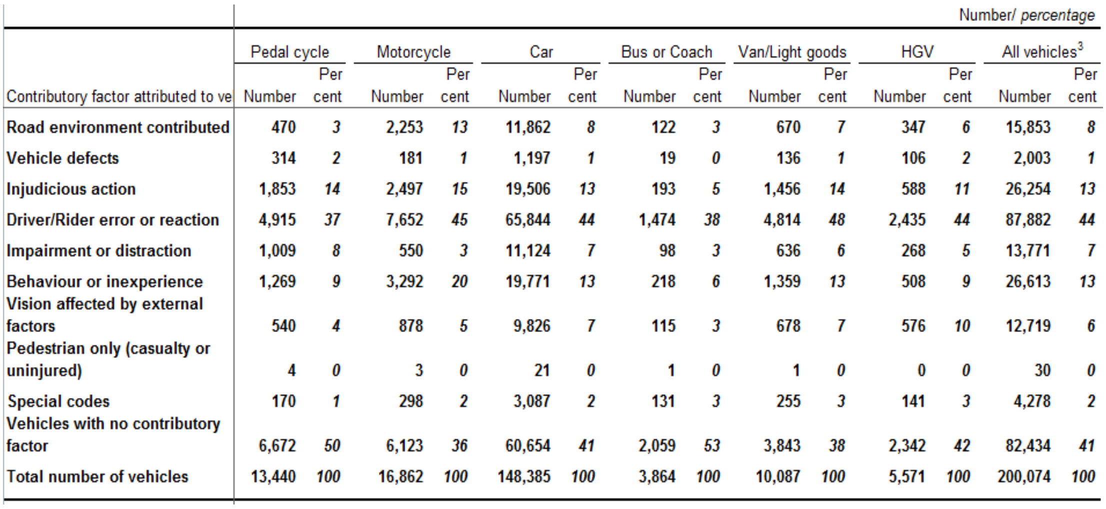
Figure 2. 14 UK 2009-2013 contributory factors for reported accidents[6]



Number/ *percentage*

Contributory factor reported in accident[1]	Killed Number	Killed Per cent[2]	Seriously injured Number	Seriously injured Per cent[2]	Slightly injured Number	Slightly injured Per cent[2]	All casualties Number	All casualties Per cent[2]
Road environment contributed	**179**	***11***	**2,404**	***13***	**17,759**	***14***	**20,342**	***14***
Vehicle defects	**43**	***3***	**417**	***2***	**2,395**	***2***	**2,855**	***2***
Injudicious action	**462**	***29***	**4,182**	***22***	**32,129**	***25***	**36,773**	***25***
Driver/Rider error or reaction	**1,104**	***70***	**12,621**	***67***	**95,319**	***75***	**109,044**	***74***
Impairment or distraction	**374**	***24***	**3,138**	***17***	**16,679**	***13***	**20,191**	***14***
Behaviour or inexperience	**469**	***30***	**4,671**	***25***	**31,544**	***25***	**36,684**	***25***
Vision affected by external factors	**130**	***8***	**1,773**	***9***	**13,279**	***10***	**15,182**	***10***
Pedestrian only (casualty or uninjured)	**291**	***18***	**3,346**	***18***	**10,729**	***8***	**14,366**	***10***
Special codes	**107**	***7***	**884**	***5***	**5,715**	***4***	**6,706**	***5***
Total number of casualties[1]	**1,587**	***100***	**18,874**	***100***	**127,848**	***100***	**148,309**	***100***

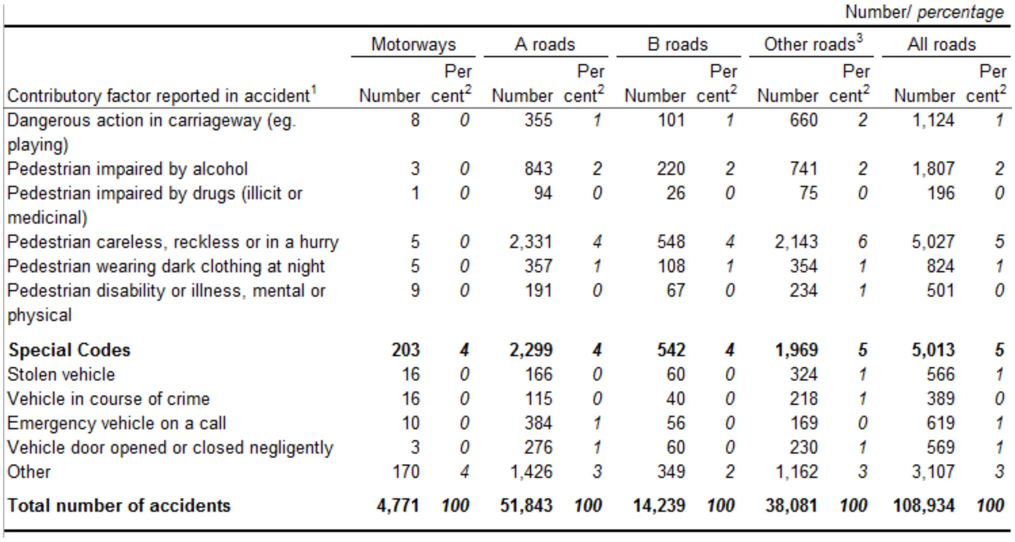
Figure 2. 15 UK 2013 casualties by contributory factor and severity[6]



Number/ *percentage*

Contributory factor attributed to ve	Pedal cycle Number	Pedal cycle Per cent	Motorcycle Number	Motorcycle Per cent	Car Number	Car Per cent	Bus or Coach Number	Bus or Coach Per cent	Van/Light goods Number	Van/Light goods Per cent	HGV Number	HGV Per cent	All vehicles[3] Number	All vehicles[3] Per cent
Road environment contributed	**470**	***3***	**2,253**	***13***	**11,862**	***8***	**122**	***3***	**670**	***7***	**347**	***6***	**15,853**	***8***
Vehicle defects	**314**	***2***	**181**	***1***	**1,197**	***1***	**19**	***0***	**136**	***1***	**106**	***2***	**2,003**	***1***
Injudicious action	**1,853**	***14***	**2,497**	***15***	**19,506**	***13***	**193**	***5***	**1,456**	***14***	**588**	***11***	**26,254**	***13***
Driver/Rider error or reaction	**4,915**	***37***	**7,652**	***45***	**65,844**	***44***	**1,474**	***38***	**4,814**	***48***	**2,435**	***44***	**87,882**	***44***
Impairment or distraction	**1,009**	***8***	**550**	***3***	**11,124**	***7***	**98**	***3***	**636**	***6***	**268**	***5***	**13,771**	***7***
Behaviour or inexperience	**1,269**	***9***	**3,292**	***20***	**19,771**	***13***	**218**	***6***	**1,359**	***13***	**508**	***9***	**26,613**	***13***
Vision affected by external factors	**540**	***4***	**878**	***5***	**9,826**	***7***	**115**	***3***	**678**	***7***	**576**	***10***	**12,719**	***6***
Pedestrian only (casualty or uninjured)	**4**	***0***	**3**	***0***	**21**	***0***	**1**	***0***	**1**	***0***	**0**	***0***	**30**	***0***
Special codes	**170**	***1***	**298**	***2***	**3,087**	***2***	**131**	***3***	**255**	***3***	**141**	***3***	**4,278**	***2***
Vehicles with no contributory factor	**6,672**	***50***	**6,123**	***36***	**60,654**	***41***	**2,059**	***53***	**3,843**	***38***	**2,342**	***42***	**82,434**	***41***
Total number of vehicles	**13,440**	***100***	**16,862**	***100***	**148,385**	***100***	**3,864**	***100***	**10,087**	***100***	**5,571**	***100***	**200,074**	***100***

Figure 2. 16 UK 2013 contributory factor by vehicle type[6]



Number/ *percentage*

	Motorways		A roads		B roads		Other roads[3]		All roads	
Contributory factor reported in accident[1]	Number	Per cent[2]	Number	Per cent[2]	Number	Per cent[2]	Number	Per cent[2]	Number	Per cent[2]
Dangerous action in carriageway (eg. playing)	8	*0*	355	*1*	101	*1*	660	*2*	1,124	*1*
Pedestrian impaired by alcohol	3	*0*	843	*2*	220	*2*	741	*2*	1,807	*2*
Pedestrian impaired by drugs (illicit or medicinal)	1	*0*	94	*0*	26	*0*	75	*0*	196	*0*
Pedestrian careless, reckless or in a hurry	5	*0*	2,331	*4*	548	*4*	2,143	*6*	5,027	*5*
Pedestrian wearing dark clothing at night	5	*0*	357	*1*	108	*1*	354	*1*	824	*1*
Pedestrian disability or illness, mental or physical	9	*0*	191	*0*	67	*0*	234	*1*	501	*0*
Special Codes	**203**	***4***	**2,299**	***4***	**542**	***4***	**1,969**	***5***	**5,013**	***5***
Stolen vehicle	16	*0*	166	*0*	60	*0*	324	*1*	566	*1*
Vehicle in course of crime	16	*0*	115	*0*	40	*0*	218	*1*	389	*0*
Emergency vehicle on a call	10	*0*	384	*1*	56	*0*	169	*0*	619	*1*
Vehicle door opened or closed negligently	3	*0*	276	*1*	60	*0*	230	*1*	569	*1*
Other	170	*4*	1,426	*3*	349	*2*	1,162	*3*	3,107	*3*
Total number of accidents	**4,771**	***100***	**51,843**	***100***	**14,239**	***100***	**38,081**	***100***	**108,934**	***100***

Figure 2. 17 UK 2013 contributory factor by road[6]

2.5 UK Road Traffic Accident Data

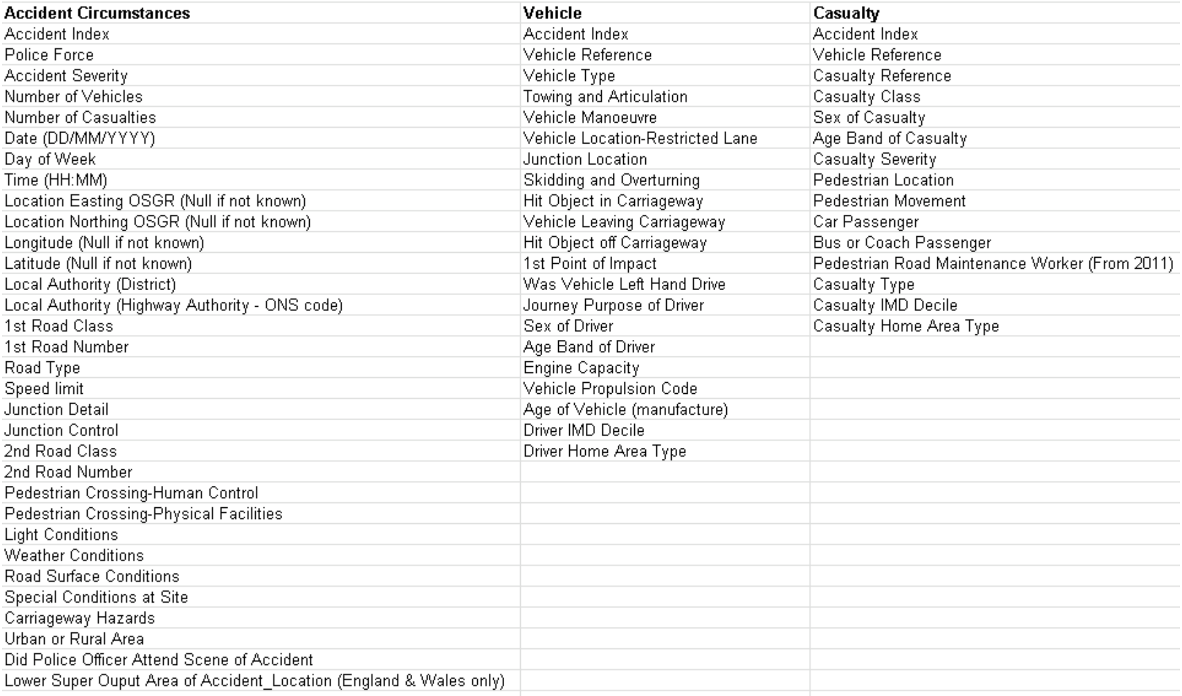
To build an accident prediction model, a core set of data is required. From this data, exploratory analysis can be conducted followed by model design. As previously mentioned road safety information tends to be managed by multiple agencies and is amenable to sharing (Mitchell, 2002). In the UK, the two main sources of road safety information are STATS19, the national road accident reporting system which includes police information, and the hospital episode statistics (HES) (The International Transport Forum, 2013, p. 428). Each agency manages different information of interest and uses it for specific purposes. All personal injury accidents which are reported to the police are recorded on a standard form called the STATS19 form. The STATS19 Road Safety dataset is published annually by the UK Department of Transport. The Department publishes the STATS19 dataset on the UK open data website which is licensed under the open government license.[7] Under this license, an individual is free to copy, publish, distribute and adapt the STATS19 dataset. The dataset is supported by the "road accident safety data guide" which is a data dictionary which describes the structure of the STATS19 data.[8]

The dataset contains 7.5 million observations providing details about the circumstances of personal injury road accidents, vehicles involved and casualty details recorded since 1979. Each observation is classified by accident severity. The accident severity classifications are "fatal", "serious" and "slight". The dataset is divided into three categories being accident, vehicle and casualty. Accident features include date, time, speed limit, road type as well as weather, light and road surface conditions and

[7] Department of Transport UK, 2014. *Road Safety Data*, http://data.gov.uk/dataset/road-accidents-safety-data, [Accessed 26 10 2014].

junction detail. Vehicle features include but are not limited to vehicle type and manoeuvre, driver sex and age as well as engine capacity. Casualty features include casualty type, sex and age band of casualty. All features recorded in each category are described in Table 2.1.

Table 2. 1 STATS19 features[8]



Accident Circumstances	Vehicle	Casualty
Accident Index	Accident Index	Accident Index
Police Force	Vehicle Reference	Vehicle Reference
Accident Severity	Vehicle Type	Casualty Reference
Number of Vehicles	Towing and Articulation	Casualty Class
Number of Casualties	Vehicle Manoeuvre	Sex of Casualty
Date (DD/MM/YYYY)	Vehicle Location-Restricted Lane	Age Band of Casualty
Day of Week	Junction Location	Casualty Severity
Time (HH:MM)	Skidding and Overturning	Pedestrian Location
Location Easting OSGR (Null if not known)	Hit Object in Carriageway	Pedestrian Movement
Location Northing OSGR (Null if not known)	Vehicle Leaving Carriageway	Car Passenger
Longitude (Null if not known)	Hit Object off Carriageway	Bus or Coach Passenger
Latitude (Null if not known)	1st Point of Impact	Pedestrian Road Maintenance Worker (From 2011)
Local Authority (District)	Was Vehicle Left Hand Drive	Casualty Type
Local Authority (Highway Authority - ONS code)	Journey Purpose of Driver	Casualty IMD Decile
1st Road Class	Sex of Driver	Casualty Home Area Type
1st Road Number	Age Band of Driver	
Road Type	Engine Capacity	
Speed limit	Vehicle Propulsion Code	
Junction Detail	Age of Vehicle (manufacture)	
Junction Control	Driver IMD Decile	
2nd Road Class	Driver Home Area Type	
2nd Road Number		
Pedestrian Crossing-Human Control		
Pedestrian Crossing-Physical Facilities		
Light Conditions		
Weather Conditions		
Road Surface Conditions		
Special Conditions at Site		
Carriageway Hazards		
Urban or Rural Area		
Did Police Officer Attend Scene of Accident		
Lower Super Ouput Area of Accident_Location (England & Wales only)		

2.6 Data Mining Overview

A key consideration for data mining is the type of data in the dataset, including the volume, structure, frequency and specific characteristics. Fatal traffic accidents are rare or infrequent events and therefore pose additional challenges for accurate prediction. Sampling techniques can be applied to help identify patterns which would otherwise be unseen. Data mining analyses data in order to identify underlying relationships and patterns and the knowledge extracted can be used to develop predictive models (Nyce, 2007, p. 9). By rationalising the trends and relationships in data, knowledge is discovered (Han, et al., 2011, p. 17). Data mining, also referred to as knowledge discovery, is defined as;

> "the *nontrivial extraction of implicit, previously unknown, and potentially useful information from data*" (Frawley, et al., 1992).

[8] Department of Transport UK, 2013. *Road Accident Safety Data Guide*, http://data.dft.gov.uk/road-accidents-safety-data/Road-Accident-Safety-Data-Guide.xls, [Accessed 25 1 2014].

The stages in data mining and knowledge discovery are outlined in Fig. 2.18 and displays the relationship between data mining and knowledge discovery. The key stages in the data mining process are briefly discussed below.

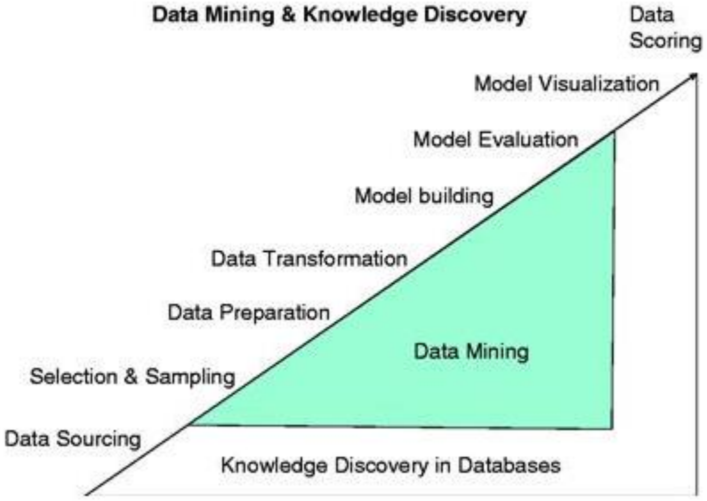


Figure 2. 18 Data mining & knowledge discovery process
Source: (Miner, et al., 2009, p. 17)

2.6.1 Data understanding and selection

A key stage in the data mining process is the selection of data and often described as data understanding stage. Data needs to be of good quality and clean as the quality of predictive models is only as good as the data used to create them (Eckerson, 2007, p. 12). An understanding of the data characteristics, content and structure should be gained as the nature of the data can affect the selection of appropriate mining and prediction techniques to apply (McCue, 2007, p. 50).

Data quality considerations include accuracy, completeness and consistency (Han, et al., 2011, p. 79). Data quality and volume are vital to ensure the reliability of a predictive model and therefore prior to choosing a dataset an assessment of the data quality should be completed. Data volume is a consideration as a dataset used for predictive modelling must be large enough to be split into training, test and validation data in order to evaluate the model. Training data is used to build a model, test data estimates model accuracy and validation data, validates the model accuracy (Miner, et al., 2009, p. 70). Similarly, enough test and validation data should be available to validate model accuracy. The validation dataset is required as it is not sufficient to report model performance on the basis of a dataset which was used to create the model and the validation data should be kept separate from data included in model building (Miner, et al., 2009, p. 70). The larger the volume of training data available, the more accurate the resulting predictive model is likely to be. The data used for this research is the STATS19 traffic accident dataset.

In data mining, data is structured as continuous data or categorical data. Continuous data relates to numbers such as the number of accidents while categorical data relates to data grouping or categorisation such as road type. Data is typically described in a data description and can include field data type, size as well as descriptive statistics such as mean, standard deviations and data groupings for categorical fields. (Miner, et al., 2009, p. 40) This research will focus on categorical data only.

Uncommon or infrequent data relates to the trends and pattern in data which do not occur very often. Some infrequent patterns in data can contain useful prediction information. However, these patterns can appear so infrequently, data mining techniques can have difficulty capturing this information (He & Garcia, 2009, p. 1265). Where infrequent events exist the dataset may also suffer from class imbalance, where the minority class is limited within the dataset. Data sampling techniques can be applied to data which can make uncommon patterns more prominent in datasets (He & Garcia, 2009, p. 1266).

2.6.2 Data preparation and transformation

Data preparation involves getting the data ready for modelling stages and involves selecting the data relevant to the experiment, transforming and reshaping the data so it is in a suitable format for analytical modelling (Miner, et al., 2009, p. 40). Data preparation can present many challenges and can be a time consuming stage of predictive modelling (Zhang, et al., 2003, p. 377). By creating a smaller dataset through selection of relevant data only and applying data reduction techniques, such as sampling, significant data mining efficiencies can be achieved (Zhang, et al., 2003, p. 377). Techniques for data transformation reduce the size of the dataset but attempts to minimise the loss of information contained in the data (Han, et al., 2011, p. 111).

A sampling technique consists of building a representative sample of a dataset under the:

> *'hypothesis that a classifier trained from that sample will not perform significantly worse than a classifier trained on the entire'* (Aounallah, et al., 2004) dataset.

Data sampling techniques are used in data mining to select a representative sample of the data population which estimates the characteristics of the data population under consideration (The SAS Institute, 1998, pp. 16-17). In the context of this research experiment, these techniques will be used to rebalance the traffic accident data so fatal traffic accidents are more prevalent. An additional feature of infrequent events is that their occurrence is often limited in datasets, with features being outweighed by more frequent events. The dataset is then considered imbalanced which poses a problem when extracting relationships in the data, however, sampling techniques can be applied to make the data amenable to prediction. When a class imbalance problem is identified, experimentation with sampling techniques may help improve prediction performance.

Sampling techniques require specialised skill and it can take a significant timeframe to identify the best sample. Where an extreme imbalance exists in a dataset, most algorithms will not perform well and will likely assign the minority imbalance as negative (Ling & Li, 1998, p. 74). For most imbalanced datasets the application of sampling techniques assists in improving classifier accuracy (He & Garcia, 2009, p. 1266).

Undersampling is a technique used in data mining to adjust the class distribution of a dataset in favour of the minority class (He & Garcia, 2009, p. 1266). With undersampling, the majority class is reduced or under sampled (Han, et al., 2011, p. 320) and randomly eliminates data from the majority class until both classes match (Japkowicz, 2000, p. 13). For example, in the case of cancer diagnosis, patients given the all clear are the majority class and patients diagnosed with cancer are the minority class (He & Garcia, 2009). With undersampling, the volume of patients in the all clear class would be reduced to bring them in line with patients diagnosed with cancer. By undersampling the majority class, trends and patterns may be removed from the data that might lead to a worse prediction for the majority class. In SPSS Modeler, undersampling is referred to as majority reduction.

Oversampling is a technique used in data mining to adjust the class distribution of a dataset in favour of the majority class (He & Garcia, 2009, p. 1266). With oversampling, the minority class is increased or over sampled (Han, et al., 2011, p. 320) until the size meets that of the majority class (Japkowicz, 2000, p. 13). For example in cancer diagnosis patients given the all clear are the majority class and patients diagnosed with cancer are the minority class (He & Garcia, 2009). With oversampling, the volume of patients in the cancer class would be increased to bring them in line with patients who were given the all clear. With oversampling, there is a risk of overfitting the minority class to a model (He & Garcia, 2009, p. 1267). Using a validation dataset to test a model trained from oversampled data will provide additional evidence that the model classifies accurately and overfitting has not occurred. In SPSS Modeler, oversampling is referred to as minority boosting.

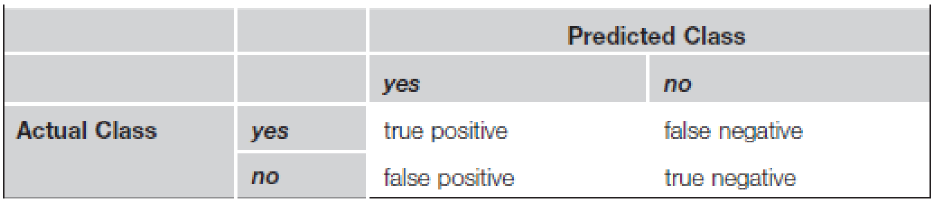
Class imbalance occurs when the class of interest is rare or infrequent i.e. the majority class far outweighs the rare class (Han, et al., 2011, p. 305). In the case of fatal traffic accidents, the non-fatal accident class far outweighs the fatal accident class. Class imbalanced datasets, when used as training data, can lead to poor predictions for the minority class as the minority class is not prevalent in the dataset.

2.6.3 Model building and evaluation

Model building involves applying a predictive modelling technique to the data created at the preparation stage. When selecting the appropriate modelling techniques and tools, consideration should be given to the appropriateness and availability of modelling tools and the intended use of the model results (McCue, 2007, p. 118). For example, accuracy may sometimes be compromised to produce a model which can be easily understood and actioned. Neural networks provide high degrees of accuracy but it can be difficult to understand the basis of the result, whereas, for decision trees rules can be extracted which can then be interpreted

Evaluation is a key stage in the data mining process and helps to assess the predictive capability of the model and identify the model which performs best (Souza, et al., 2002, p. 1). Specific focus is given to techniques used to evaluate classification models which will be constructed as part of this experiment research such as the confusion matrix, receiver operator curve (ROC) and the area under the curve (AUC).

A confusion matrix is designed to show correct and incorrect predictions (Han, et al., 2011, p. 304). The terminology used to describe correct and incorrect predictions are true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). It is common for classification models to be evaluated using confusion matrix results. Fig. 2.19 is an example of a typical confusion matrix.



		Predicted Class	
		yes	***no***
Actual Class	***yes***	true positive	false negative
	no	false positive	true negative

Figure 2. 19 Confusion matrix example

Source: (Witten, et al., 2011, p. 164)

Commonly used evaluation measures which can be calculated from confusion matrix results, include accuracy, which measures the percentage of data correctly classified, precision, which measures the percentage of data which are correctly labelled as positive and recall, which measures the percentage of the positive targets labelled correctly (Han, et al., 2011, pp. 305-307).

- Accuracy – The proportion of TPs and TNs which were classified correctly. This is also called the accuracy rate.
- Recall/Sensitivity - The proportion of TPs which were classified correctly. This is also called the true positive rate.
- Precision – The proportion of TP's which were classified as fatals which were actually fatal.

The ROC curve provides a method to compare classification models (Han, et al., 2011, p. 312). The x-axis represents the false positive rate while the y-axis represents the true positive rate (Han, et al., 2011, p. 312). In this experiment research, the true positive rate is the proportion of fatal accidents which are classified correctly. Fig. 2.20 presents a sample ROC curve and represents the trade-off between the true positive rate, also known as recall, and the false positive rate. The further the ROC curve is from the diagonal line the more accurate the model is.

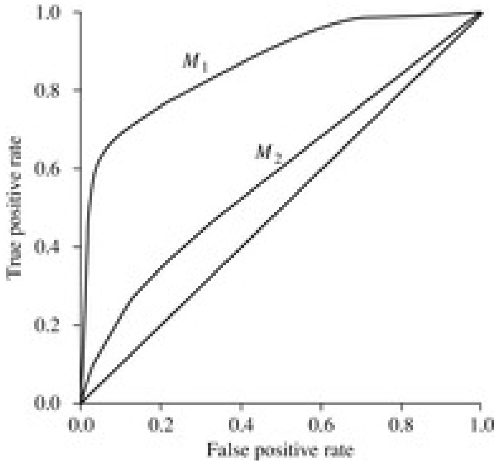


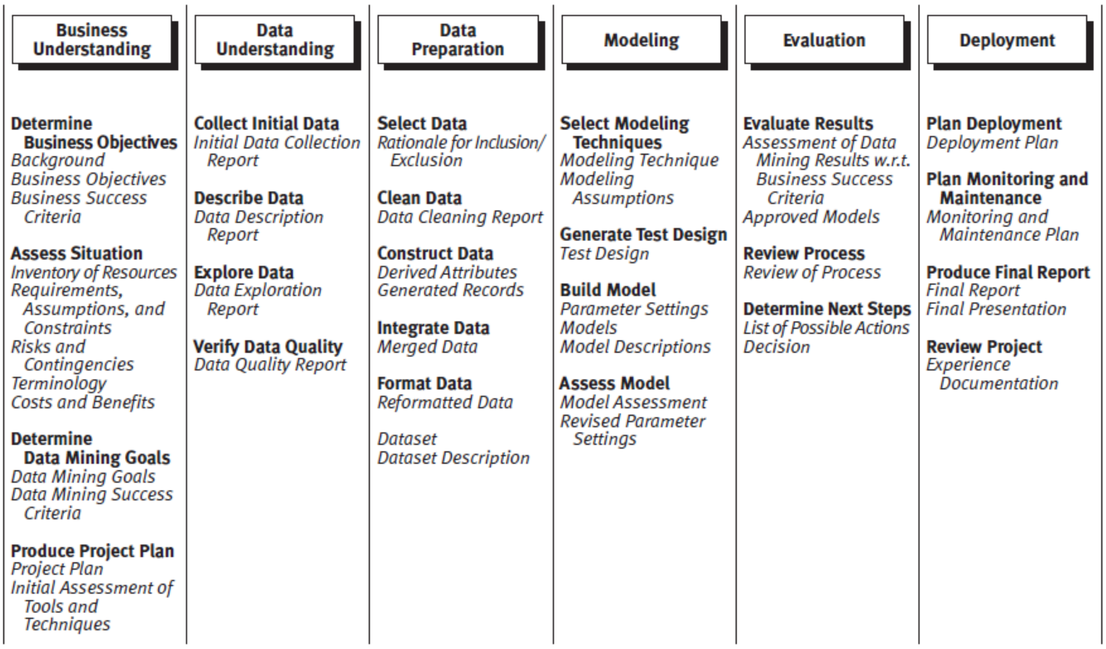
Figure 2. 20 Sample ROC curve
Source: (Han, et al., 2011, p. 312)

Area under the curve (AUC) is a metric used to assess model accuracy (Han, et al., 2011, p. 312). AUC is measured on a scale ranging from 0.5 to 1. AUC refers to the area under the ROC curve. The larger the AUC (Witten, et al., 2011, p. 177), the more accurate a model is. A large area indicates an AUC which is close to 1.

Interpretability of models is a non-technical evaluation method. Decision trees are popular because business rules can be explained in English and can be easily understood by users (Berry & Linoff, 2004, p. 165). A balance must be found between interpretability and accuracy (McCue, 2007, p. 118).

2.6.4 Data mining methodology

The CRISP-DM process is based on the industry experience of data mining practitioners rather than academic researchers and is a best practice model for data mining (McCue, 2007, p. 50). CRISP-DM is designed to encourage best practice, aims to aid faster and better results from data mining (Shearer, 2000, p. 13) and provides a structured approach to planning and implementing a data mining project. (Beshah, et al., 2013) in recent research on road accident data, followed the CRISP-DM approach to conform to an industry standard. A similar methodology was adopted for this research, with adaptation to meet the specific requirements of this research experiment. The image in Fig. 2.21 depicts the standard CRISP-DM process and the recommended phases and tasks.



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 2. 21 CRISP-DM process phases and tasks
Source: (Chapman P., et al, 2000, p. 12)

2.7 Predictive Analytics

Predictive analytics is a technique used to predict future outcomes and trends using quantitative techniques to derive insights from data. The main component of predictive analytics is the predictor. Statistical, data mining and machine learning techniques are utilised. A variety of algorithms can be used to analyse historical information to make predictions about future events or behaviour (McCue, 2007, p. 117). The focus of this dissertation research is to apply prediction techniques to extract knowledge from traffic accident data. Prediction techniques such as decision trees and Bayes net enable organisations to rationalise the relationships in data through building prediction models which are used to score new data. Often the techniques identify relationships which would otherwise go unnoticed and the insights provided can be used for more focussed plans and decision making.

Predictive analytics is widely used in the business environment. In marketing it helps marketers to understand purchasing patterns to create new sales and reduce churn to the competition (Berry & Linoff, 2004, pp. 115-116). In public safety predictive analytics is used to support analytics applications designed to keep the public safe (McCue, 2007, p. 52) for example predictive policing.

Predictive techniques or modelling algorithms in general are considered as supervised or unsupervised learning techniques. The supervised techniques include classification and regression models and aim to identify rules within the data which can be applied to predict a defined outcome (McCue, 2007, p. 119). Unsupervised techniques group data with similar attributes but interpretation can be challenging.

This research focuses on the application of three classification techniques. A classification technique is described by (Wahed, et al., 2012) as:

> *"a machine learning technique used to predict the correlation between data samples and classes".*

Classification involves considering the features of a case and aligning it to one of a predefined set of classes (Berry & Linoff, 2004, p. 9). For example, in telecommunications, "churn" and "stay" are two predefined classes, with customers who leave grouped into the "churn" class and customers who stay grouped into the "stay" class. When model building is complete, new data can be scored against the model to extract prediction. There are many classification techniques, however, this experiment applies two types of decision trees, C5.0 and Chaid, and Bayes net.

A decision tree is a classification technique which separates data using class labelled cases (Han, et al., 2011, p. 274). When viewed graphically, the splits resemble an inverted tree or decision tree. C5.0 and Chaid trees are presented in the same format but use different split mechanisms. Decision trees are based on a set of rules extracted from the data and most commonly the rules are presented in a tree like format. A target or output is set for the decision tree, presented as the top node on the tree, the data is split into homogeneous subgroups, linked to that target, and the subgroups are represented as branches on the tree. The process is iterative and once the decision tree identifies the strongest predictor relating to the target it records it as a node on the branch and moves to the next level down on the tree to identify the next strongest predictor and so on until terminal nodes for each branch are created. The splitting generally ceases when separation is no longer meaningful in relation to the target (Konstantinos & Chorianopoulos, 2009, pp. 110-117). Once the decision tree is built new data can be scored against the model and a class prediction can be extracted. Decision trees are popular because business rules can be explained in English which subject matter experts can understand (Berry & Linoff, 2004, p. 165). A key characteristic of decision trees is the interpretability of rules. Typically, there is a trade-off between model interpretability and model accuracy. Examples of the decision tree format and the format of rule extraction in English are presented in Fig. 2.22 and Fig. 2.23.

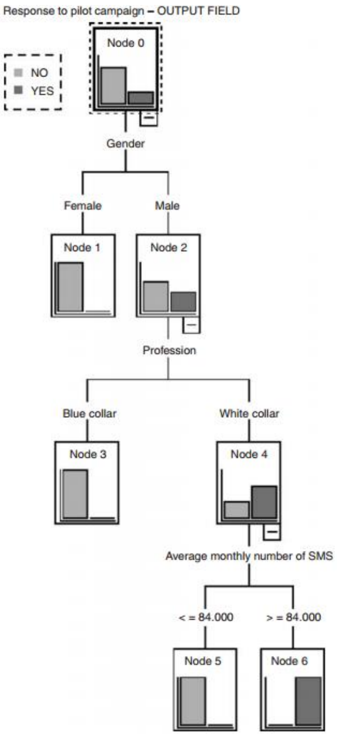
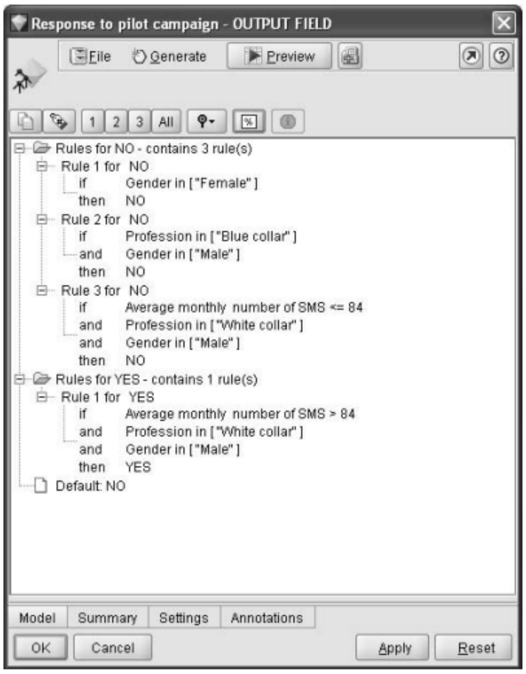


Figure 2. 22 Sample decision tree from IBM SPSS Modeler

Source: (Konstantinos & Chorianopoulos, 2009, p. 116)



```
Rules for NO - contains 3 rule(s)
   Rule 1 for  NO
      if      Gender in ["Female"]
      then    NO
   Rule 2 for  NO
      if      Profession in ["Blue collar"]
      and     Gender in ["Male"]
      then    NO
   Rule 3 for  NO
      if      Average monthly number of SMS <= 84
      and     Profession in ["White collar"]
      and     Gender in ["Male"]
      then    NO
Rules for YES - contains 1 rule(s)
   Rule 1 for  YES
      if      Average monthly number of SMS > 84
      and     Profession in ["White collar"]
      and     Gender in ["Male"]
      then    YES
Default: NO
```

Figure 2. 23 Sample decision tree ruleset from IBM SPSS Modeler

Source: (Konstantinos & Chorianopoulos, 2009, p. 117)

Chaid stands for Chi-squared interaction detection and was proposed by Kass in 1980. Chaid applies the decision tree process, which, applies the Chi-squared statistical test for proportion to determine a split and uses multiway splits to construct trees (Miner, et al., 2009, p. 246). Chaid has a stopping mechanism which determines when a sub tree is complete (Miner, et al., 2009, p. 792). The technique is popular for market segmentation, however, Chaid decision trees can be large and this can restrict the user understanding (Miner, et al., 2009, p. 247).

The C5.0 decision tree is a popular modelling technique which is based on entropy and information gain.[9] It is a descendant of the C4.5 decision tree technique. As with Chaid, the decision tree process is followed but the splitting mechanism is specific to C5.0. C5.0 uses the highest information gain to determine a split. When tree construction is complete, the lowest level splits which contribute least are removed or pruned[.9] C5.0 models are robust to missing data and large numbers of input features. C5.0 can only be used with categorical outcomes, and unlike Chaid, cannot construct trees based on numeric outcomes.

A Bayesian belief network, more commonly referred to as Bayes net is a probabilistic graphical modelling technique (Han, et al., 2011, p. 323). Bayes net displays a dataset in a graphical model. Each variable in the graphical model is called a node. Each node has a conditional dependency or importance. There may be a strong causal relationship between variables in Bayes net but this doesn’t mean there is a cause and effect relationship (Han, et al., 2011, p. 324). This experiment research will use Bayes net to represent the probabilistic relationship between fatal traffic accidents and the characteristics of fatal traffic accidents.

2.8 Traffic Accident Prediction Research

This research experiment investigates the area of fatal traffic accident classification and identifying current research in the field of traffic accident prediction helped to define the experiment.

Decision Tree Models for Count Data

(Wah, et al., 2012) conducted an experiment on a Malaysian motor cycle road accident dataset and used categorical features to model motor cycle accident occurrences and compared and contrasted the classification performance of CART decision trees, poisson regression and negative binomial regression to predict death or serious injury accidents. Their research produced 78% prediction accuracy using CART which was marginally better than the other two models. Their research did not present findings for precision or recall. They noted that the rules extracted from CART were easy to interpret. They found that the most significant contributors to high frequency serious

[9] IBM, 2014. *SPSS Modeler C5.0 Node.* http://www-01.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm ,[Accessed 10 11 2014].

injury or fatal accidents were when an accident occurred on a straight road or on a bend or at a junction. Most serious injury or fatal accidents occurred on straight roads. Finally, their research found that serious injury or fatal accidents occurred most often when the weather was clear and the road surface conditions were dry.

The prediction results achieved by (Wah, et al., 2012), led to this research giving consideration to experimenting with the STATS19 dataset to predict fatal traffic accidents using a number of classification techniques including two decision trees and Bayes net.

Bayesian Network model of two car accidents

(Simoncic, 2004) describes using the Bayes net technique to model two car traffic accident data using factors which influence accident outcomes including "fatality". Bayes net models use probability and graph theory to model the behaviour of complex situations such as traffic accident behaviour. A model was presented which captured the relationships between different accident factors. Some of the factors considered include road, traffic, speed and time. (Simoncic, 2004) states that some of the factors are interrelated such as traffic and time.

Their research found that Bayes net can be *"fruitfully"* (Simoncic, 2004) used to model traffic accidents. The results are presented as probabilities rather than accuracy, precision or recall. The main advantage of this modelling technique was its ability to find relationships between factors that relate to fatal or serious injury outcomes. (Simoncic, 2004) found that the model results were encouraging and mentioned that adding additional features is one approach which might provide an improvement. Finally, (Simoncic, 2004) found that by modelling more data, the reliability of the model improved.

Traffic Accident Analysis Using Machine Learning Paradigms

(Chong, et al., 2005) research summarises the performance of neural networks, support vector machines, decision trees and a concurrent hybrid approach to model driver injury severity resulting from traffic accidents. Their research found that the hybrid approach produced the best classification accuracy for the fatal injury class at 90% accuracy but recall or precision performance was not discussed. A hybrid approach is a technique which combines learning models into one model which can exploit the best features of each model to provide a better prediction result. They combined a decision tree and a neural network technique to create a hybrid model.

Their research did not present findings for precision or recall. They mentioned that *"fatality has the highest cost to society economically and socially"*, therefore predicting fatal accidents accurately is beneficial to society. (Chong, et al., 2005) mentioned that speed is recognised as an important factor which contributes to injury severity. In their research, they could not use speed to predict as it was unknown in 68% of cases. According to (Chong, et al., 2005), if speed could have being used, it would likely have improved the performance of the models.

Analysis of factors associated with traffic injury severity on rural roads in Iran

(Kashani, et al., 2012) research considers crash fatality and injury rates on two lane, two way and freeway roads in Iran which has one of the highest injury and fatality rates in the world. Using the classification and regression tree (CART) technique, their research found that the factors which influence injury severity most were seat belt use, cause of crash and collision type. Their research found that seat belt use was the most important factor for two lane and two way rural roads. Seat belt use is less important on freeways as police enforce the use of seat belts on freeways. Cause of crash was the next most important factors with speeding and inappropriate overtaking being the biggest causes on two way and two lane rural roads. In order to reduce accident severity, (Kashani, et al., 2012) suggested improved policing and road design and stopping pedestrians and animals from crossing freeways.

2.9 Conclusion

This chapter focused on the current literature in the domain of road safety and specifically fatal road traffic accidents. The role of road safety data in understanding road traffic accidents was briefly discussed. A cross section of the factors which significantly affect the rate of severe and fatal traffic accidents were discussed and road traffic accident data and statistics in the UK were outlined. The current literature shows that the factors which contribute to fatal road traffic accidents are multifaceted and difficult to quantify and can be considered to be "*neither simple or linear*" (The International Transport Forum, 2014, p. 491) in nature.

In addition the chapter considered data mining and prediction techniques in the context of fatal traffic accidents. The various stages of data mining including data understanding, selection and preparation were presented. Considerations for sampling techniques and model build were discussed and the CRISP-DM best practice data mining methodology was outlined. Predictive analytics classification techniques and assessment and evaluation techniques were discussed. Finally current research papers in the area of traffic accidents prediction were summarised.

3 EXPERIMENT DESIGN

3.1 Introduction

This chapter outlines the experiment design which was conducted as part of this dissertation research. The nature of the data used for the experiment is presented together with the data preparation employed in order to construct the data for predictive modelling.

Predictive modelling classification techniques were used during the modelling phase of the experiment and C5.0 and Chaid decision trees and Bayes net were selected. Each technique trained a model on baseline or normal, semi reduced and reduced data and sampling techniques were applied. In addition the techniques used to assess model performance and evaluate models are discussed.

3.2 Implementation Methodology

The methodology adopted to implement the research and experiment was based on the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology as described in chapter 2. This research methodology was adapted to align with the specific requirements of this experiment design and the adapted methodology guided the focus and phases of the experiment. An overview of the key stages in the adapted experiment methodology is outlined in Fig. 3.1.

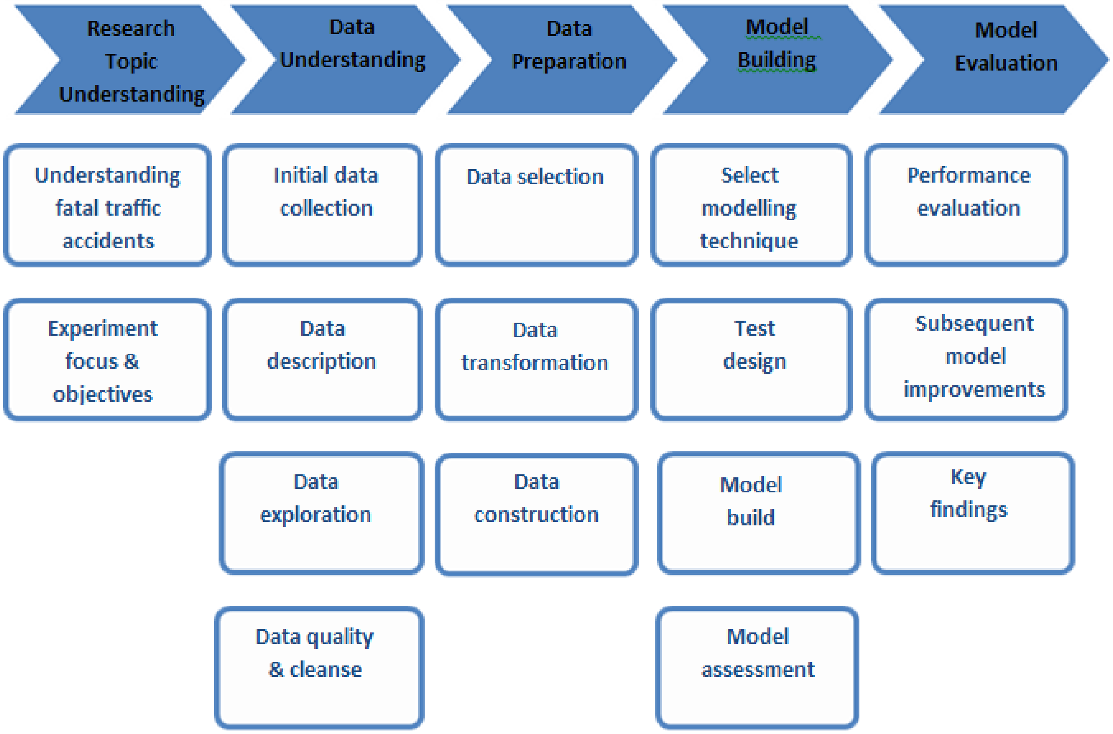


Figure 3. 1 Adapted Experiment Methodology

The initial stage was to develop an understanding of fatal traffic accident background, key data characteristics and relevant industry wide significant factors worth considering. The focus and the objective of the research experiment was defined. Understanding the data was divided into four stages including an explanation of the initial data collection, a description of the data, an outline of the planned exploration techniques and the methods to be used to verify data quality and cleanse if required. Data preparation was divided into two stages being the selection and transformation of data and an outline for construction of the data. The approach to building each model included a description of the modelling techniques employed, the key features of the test design, the stages in model build and the steps in assessment of the model performance. The final stage in this research methodology was to outline the steps in evaluation of performance of the model, consideration of model improvements and presentation of the key findings.

3.2.1 Database & analytics software

To execute this methodology two market leading database and analytics tools were used, "Toad for Oracle database" and IBM SPSS Modeler. Oracle is a well-established database vendor which sells database technology to large companies and governments worldwide. Toad is an Oracle database development and administration tool which provides capabilities to administer and design Oracle databases. Oracle database and Toad were selected to host and manage the data for the experiment as they are widely used and well regarded. The Oracle PL/SQL programming language was used to construct the data for the experiment. SPSS Modeler, IBM's flagship predictive modelling product, was used to manipulate data and build predictive models. It is mature, intuitive and easy to use. In addition it has strong ETL, data exploration, data preparation and predictive modelling capabilities. It provides a range of modelling techniques including decision trees, statistical models and text analytics.

For this experiment, SPSS Modeler was used for data understanding and preparation, model design and build as well as results analysis and presentation. Three modelling techniques were used to build predictive models for the experiment. These techniques were C5.0, Chaid and Bayes net and technique descriptions are outlined in chapter 2.

3.2.2 Focus of experiment and objectives

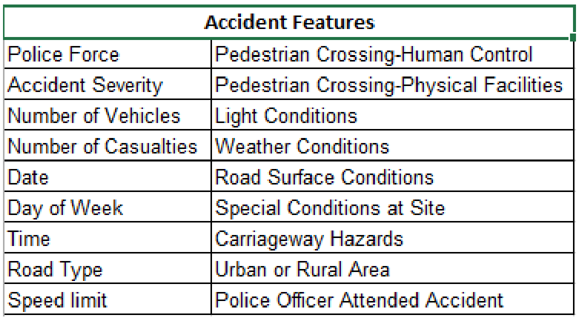
As identified in chapter 2, fatal traffic accidents are a significant worldwide issue. Road safety agencies across the world, are focussing on researching and identifying causes and potential improvements in an effort to develop enhanced road safety plans to focus efforts and reduce the number of fatalities. Current developments in road accident research, considers the potential for analysing existing road safety data and using predictive analytics techniques to identify key contributory factors which would allow more focused actions. The focus of this research was to build predictive models, based on an extensive UK road traffic accident dataset STATS19, to classify fatal road traffic accidents using C5.0, Chaid and Bayes net classification techniques and to evaluate model performance. The models which provide the best results using each technique were analysed and evaluated using accuracy, recall, precision and ROC curve performance metrics. In addition the aim was to interpret models to identify the predictive factors which are most likely to contribute to fatal traffic accidents.

(Wah, et al., 2012) conducted a similar experiment on a Malaysian motor cycle road accident dataset and used categorical features to model motor cycle accident occurrences and compared and contrasted the classification performance of different modelling techniques to predict death or serious injury accidents. They converted the frequency of motor cycle accidents which involved death or serious injury into categorical dependent variables of zero, low and high. The factors they considered were collision type, road geometry, time, weather, road surface conditions and time of day. Their research produced 78% prediction accuracy using decision trees (Wah, et al., 2012). Following review of the paper and considering the good prediction results achieved, the question arose whether classification techniques applied to UK traffic accident data could hold similar prediction characteristics?

As mentioned, the data source for the experiment was the UK's STATS19 dataset of reported personal injury road accidents and is the only national source of detailed road accident information in the UK.[10] The STATS19 dataset contains three separate datasets, vehicle, casualty and accident. The vehicle dataset contains information relating to vehicles which were involved in accidents while the casualty dataset contains information relating to the people who were the casualties of accidents. The accident dataset contains information which directly relates to traffic accidents and is the focus of this experiment. The majority of accident features in STATS19 are categorical and are listed in Table 3.1. This research only modelled categorical features which is similar to research by (Wah, et al., 2012). However, each categorical feature in the dataset can have many values, for example, Police Force includes the name of each individual police constabulary who recorded an accident. When a categorical feature has many values, this is described as a high order nominal.

[10] Department of Transport UK, 2014, *STATS19 Road Accident dataset.* http://www.adls.ac.uk/department-for-transport/stats19-road-accident-dataset/?detail [Accessed 03 10 2014].

Table 3. 1 Traffic accident features



Accident Features	
Police Force	Pedestrian Crossing-Human Control
Accident Severity	Pedestrian Crossing-Physical Facilities
Number of Vehicles	Light Conditions
Number of Casualties	Weather Conditions
Date	Road Surface Conditions
Day of Week	Special Conditions at Site
Time	Carriageway Hazards
Road Type	Urban or Rural Area
Speed limit	Police Officer Attended Accident

Part of this research experiment was to identify the features which have the most predictive power. The predictive features in their basic form may not have much predictive power, however, grouping categorical features can increase the predictive power. In data mining terminology, this grouping is called data transformation. To gain the most predictive power, the best groupings must be identified. In this research experiment, the Chaid decision tree technique was used to identify the groups with the most predictive information. By modelling the accident dataset, it was hoped that the resulting models would accurately identify the combinations of predictive factors that contribute to fatal traffic accidents. While it was important to identify the factors that contribute to fatal traffic accidents, model interpretability was also important. For example, from a decision tree, it should be possible to extract and interpret rules which explain the factors that contribute to fatal traffic accidents. In data mining terminology, a hold out or validation dataset is the term used to describe data which a model has not previously seen. A further aim of this experiment was to measure the performance of predicted results as compared to actual results on hold out data. Accident dynamics do not vary dramatically, especially in the short term, and this assessment evaluates how well a model is likely to behave when new data is scored.

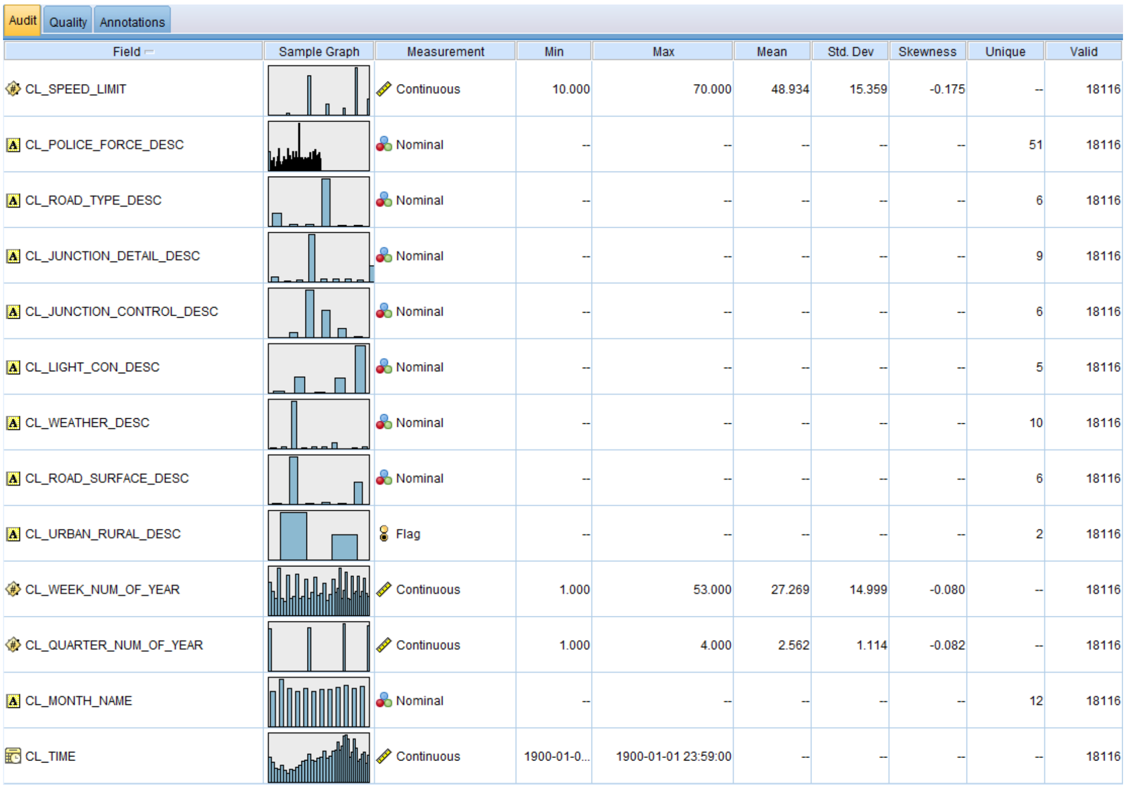
3.3 Data Understanding

Extracting actionable insights from historical information is a key aim of using predictive analytics. Other issues must also be considered and addressed to ensure successful results. These issues are discussed in the chapters that follow.

3.3.1 Initial data collection

Due to the nature of fatal accidents, recording of fatal events is largely complete given the involvement of police, hospitals, death certification and legal reporting requirements. In the UK, all personal injury accidents which are reported to the police are recorded on a standard STATS19 form. The STATS19 dataset contains detailed data relating to reported road accidents in the UK including accident circumstances, vehicle type and related casualties. Although fatalities are accurately reported, the dataset is considered incomplete in relation to non-fatal accidents. Despite this limitation, the STATS19 dataset is considered "*the most detailed, complete and reliable single source of information on road casualties covering the whole of Great*

Britain, in particular for monitoring trends over time".[11] As mentioned in chapter 2, the STATS19 dataset is divided into three categories accident, vehicle and casualty. For the purpose of this research experiment the accident dataset was selected as it contained a wider range of features. As described in Table. 2.1 accident features include date, time, speed limit, road type as well as weather, light and road surface conditions and junction detail. The features in the accident dataset are mainly environmental characteristics. This research was based on accident records in the recent past and data from 2005 to 2012 was selected for this experiment. Older data was not considered as driver behaviour and road volumes adjust over time. An Oracle database was created and accident data was loaded from a comma delimited file to a staging area. SPSS Modeler managed table creation and loading the data into the staging area. The data audit node was used to initially understand the accident data and the results are presented in Fig. 3.2. For continuous or numeric data, min was the minimum value and max was the maximum value. For example, speed limit has a minimum of 10 and a maximum of 70. For nominal or categorical data, the unique field described the number of unique values each nominal can have. For example, possible values for road surface are dry and snow.

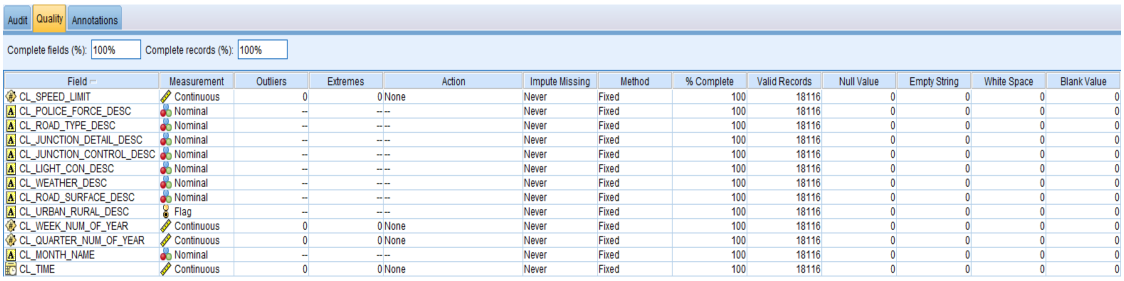


Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
CL_SPEED_LIMIT		Continuous	10.000	70.000	48.934	15.359	-0.175	--	18116
CL_POLICE_FORCE_DESC		Nominal	--	--	--	--	--	51	18116
CL_ROAD_TYPE_DESC		Nominal	--	--	--	--	--	6	18116
CL_JUNCTION_DETAIL_DESC		Nominal	--	--	--	--	--	9	18116
CL_JUNCTION_CONTROL_DESC		Nominal	--	--	--	--	--	6	18116
CL_LIGHT_CON_DESC		Nominal	--	--	--	--	--	5	18116
CL_WEATHER_DESC		Nominal	--	--	--	--	--	10	18116
CL_ROAD_SURFACE_DESC		Nominal	--	--	--	--	--	6	18116
CL_URBAN_RURAL_DESC		Flag	--	--	--	--	--	2	18116
CL_WEEK_NUM_OF_YEAR		Continuous	1.000	53.000	27.269	14.999	-0.080	--	18116
CL_QUARTER_NUM_OF_YEAR		Continuous	1.000	4.000	2.562	1.114	-0.082	--	18116
CL_MONTH_NAME		Nominal	--	--	--	--	--	12	18116
CL_TIME		Continuous	1900-01-0...	1900-01-01 23:59:00	--	--	--	--	18116

Figure 3. 2 SPSS Modeler audit of initial data

[11] Department of Transport UK, 2013. *Road accidents and safety statistics.* https://www.gov.uk/government/collections/road-accidents-and-safety-statistics, [Accessed 03 11 2014].

Fig. 3.3 presents the initial data quality assessment. There were a total of 18,115 fatal traffic accident records. The data audit node was used to understand the distribution of the data, which is described in more detail in chapter 4. This initial data understanding was completed prior to implementation to ensure that an effective experiment design was outlined to meet the data requirements.



Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
CL_SPEED_LIMIT	Continuous	0	0	None	Never	Fixed	100	18116	0	0	0	0
CL_POLICE_FORCE_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_ROAD_TYPE_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_JUNCTION_DETAIL_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_JUNCTION_CONTROL_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_LIGHT_CON_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_WEATHER_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_ROAD_SURFACE_DESC	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_URBAN_RURAL_DESC	Flag	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_WEEK_NUM_OF_YEAR	Continuous	0	0	None	Never	Fixed	100	18116	0	0	0	0
CL_QUARTER_NUM_OF_YEAR	Continuous	0	0	None	Never	Fixed	100	18116	0	0	0	0
CL_MONTH_NAME	Nominal	--	--	--	Never	Fixed	100	18116	0	0	0	0
CL_TIME	Continuous	0	0	None	Never	Fixed	100	18116	0	0	0	0

Figure 3. 3 SPSS Modeler data quality check of initial data

3.3.2 Data description

Each observation in STATS19 accident dataset was classified by accident severity. The accident severity classifications are fatal, serious and slight. For the initial data understanding, fatal cases were reviewed as these are the focus of this research experiment. Using SPSS Modeler, the data auditing capabilities were used to profile the characteristics of the data. The output from the SPSS Modeler data audit node, presented in table format, was used as the initial data description for the data preparation stage. The descriptive statistics reviewed and the data quality of each feature was considered. Generally, the higher the quality of the data being used, the more accurate the predictions are likely to be (Guillet & Hamilton, 2007, p. 120).

3.3.3 Data exploration

This task initially ran querying and reporting techniques in Toad query designer to answer questions which may help to gain a better understanding of the data. The data audit node in SPSS Modeler was used to expand this understanding so that data types and the distribution of key fields was better understood. The data was then searched for patterns and interesting relationships between features which might provide additional data understanding.

3.3.4 Data quality and cleanse

Data quality and volume are vital to ensure the reliability of a predictive model. If the source data is of good quality and there is sufficient volume, a model's reliability will likely increase. At the data quality stage, fatal cases were assessed for completeness and correctness. Any missing, incorrect data or quality issues were identified and these issues were addressed. All cleansing, if any, was applied to fatal accidents only given the focus of the experiment.

The STATS19 dataset is a well-known data set which is often used for academic research and the most commonly used source of UK road safety statistics. Fatal traffic accidents, which are the focus of this research, are known to be accurately recorded (The International Transport Forum, 2014, p. 491). For this reason, it is expected that the data quality will be high and therefore for this research experiment the data quality requirements are reduced.

3.4 Data Preparation

Many of the features considered for prediction are high order nominals and a good prediction result was difficult to achieve without reducing category levels to higher level groupings. A simple analogy will help to explain this concept. For example traffic accidents can be assigned classifications such as slight or serious and these can then be grouped into non-fatal accidents. In this context, non-fatal accidents are a higher level grouping. The chapters that follow outline the processes followed to select categorical features, which were reduced and used in the modelling phase to build prediction models, as well as building the base datasets from which prediction models were trained in the model building phase.

3.4.1 Data selection and transformation

Each observation in STATS19 dataset was classified by accident severity and the severity classifications are fatal, serious and slight. For the initial data analysis, serious and slight were grouped and named non-fatal and fatal continued to be named fatal. In SPSS Modeler, accident outcome is used to describe the non-fatal and fatal groups. All accident data from 2005 to 2012 was extracted from STATS19 was loaded to Oracle and used as the basis for the data transformation.

Data transformation relates to the process of transforming original data sources into formats appropriate for data mining (Han, et al., 2011, p. 113). The technique selected for this experiment was supervised discretisation, which used class label information to identify split-points in the data. These spilt-points were presented on a decision tree as branches, representing homogeneous subgroups with respect the target field (Konstantinos & Chorianopoulos, 2009). To identify the features which had the best predictive information, the Chaid decision tree data mining technique was used. Chaid was configured with a target of accident outcome in order to identify the features in STATS19 data with the most predictive information. Chaid is based on the Chi-squared statistic of proportion and splits data into groups which can be used for categorical feature level transformation. Fig. 3.4 outlines the process for discretisation applied for this experiment.

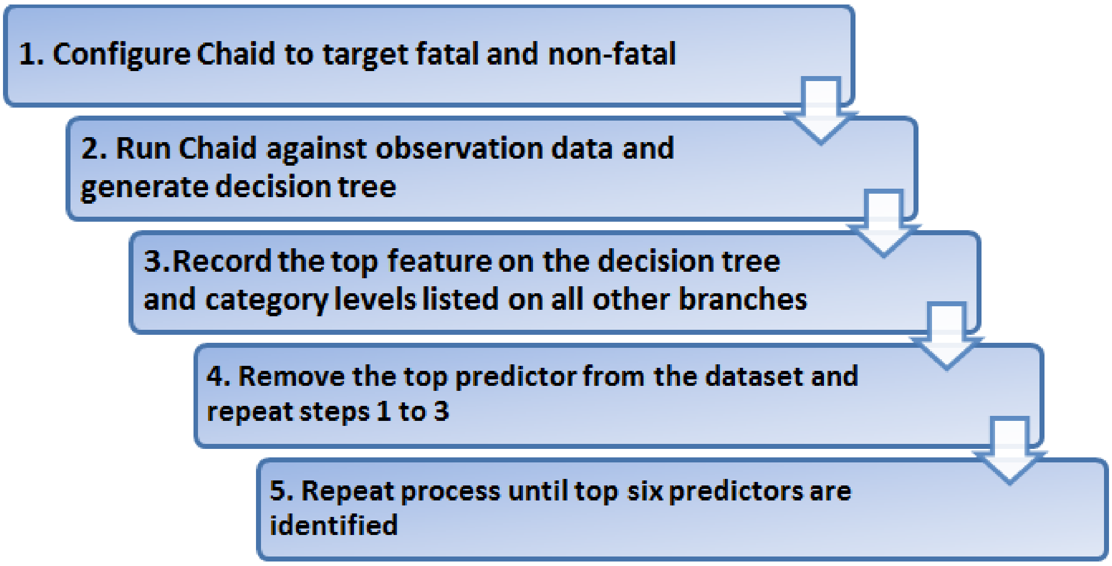

1. Configure Chaid to target fatal and non-fatal
2. Run Chaid against observation data and generate decision tree
3.Record the top feature on the decision tree and category levels listed on all other branches
4. Remove the top predictor from the dataset and repeat steps 1 to 3
5. Repeat process until top six predictors are identified

Figure 3. 4 Process for discretisation

The technique identified the feature at the top of the decision tree as having the most predictive information and the category levels of the various branches as having a relationship with the target. By removing the top feature and repeating the process, the technique was reapplied to search for the next most important feature and related split information. On each run, the feature with the strongest relationship to the accident outcome appeared at the top of the tree. The top features and the category levels recorded at all six stages were used to create datasets at the data construction stage. The process was limited to six repetitions for this experiment as further repetitions affected the processing capability of the available hardware.

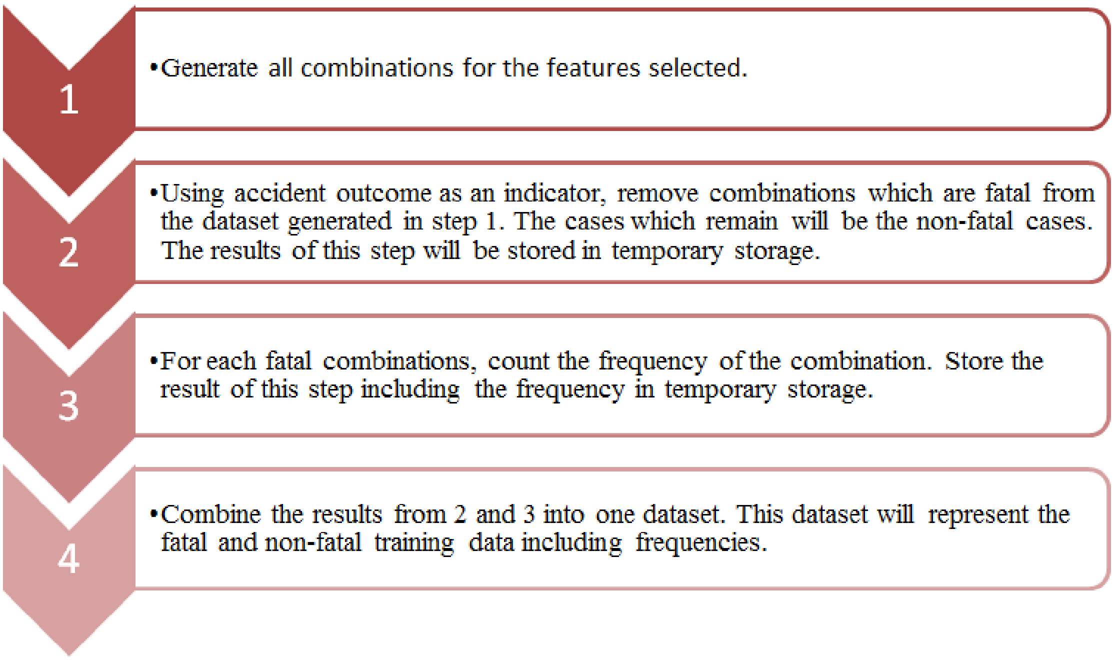
3.4.2 Data construction

As previously mentioned, the target for this experiment was accident outcome. The basis for the target was the combinations of factors and the frequency of these combinations. All fatal accidents had a positive frequency while non-fatal accidents had a frequency of 0. Four datasets were created as part of data construction as outlined in Table 3.2. Three were training data and the fourth was validation data. The training datasets were used during the modelling phase of this experiment and the validation data used for independent testing during evaluation.

Table 3. 2 Data Construction Datasets

Type of data	Data set name	Description
Training	Reduced	All categorical features are reduced
Training	Semi reduced	Some categorical features are reduced
Training	Normal	No categorical features are reduced
Hold out	Hold out dataset	This is the 2013 data which will be unseen by the model

Fig. 3.5 outlines the process followed to construct training data for the experiment models.



1 •Generate all combinations for the features selected.

2 •Using accident outcome as an indicator, remove combinations which are fatal from the dataset generated in step 1. The cases which remain will be the non-fatal cases. The results of this step will be stored in temporary storage.

3 •For each fatal combinations, count the frequency of the combination. Store the result of this step including the frequency in temporary storage.

4 •Combine the results from 2 and 3 into one dataset. This dataset will represent the fatal and non-fatal training data including frequencies.

Figure 3. 5 Training set construction process

The output of step 4 was the input to the modelling phase of this research. Each of the training sets was complimented by sampling techniques in the modelling phase. The combination of data creation and sampling techniques was designed to help identify the optimal combination to give the balance between prediction accuracy and the interpretability of model results.

As this research used the accident dataset only, there was no data integration process required. Formatting took place in SPSS Modeler and Oracle, however, most formatting took place prior to the construction of the training and validation datasets. Some formatting and data manipulation was required at the modelling stage. For example, the accident outcome was defined and sampling technique were applied at the modelling stage.

3.5 Model Building

A number of modelling techniques were used for this research experiment. Models were built on training data and model quality was estimated on test data. Following testing, models were validated against validation data.

3.5.1 Select modelling technique

One of the objectives of this experiment was to classify fatal traffic accidents. This type of modelling problem is called classification. Three classification techniques were selected to model the STATS19 data. These techniques come from the decision tree and Bayesian family of classifiers. The modelling techniques were C5.0 decision tree, Chaid decision tree and Bayes net. The modelling techniques were modelled on three datasets. These datasets were created in the data preparation phase. These datasets were:

a) Reduced dataset: training data where all features were grouped.
b) Semi reduced dataset: training data where some features were grouped.
c) Normal dataset: training data where no features were grouped.

Fatal traffic accidents are generally described as low frequency and these low frequency classes are often imbalanced. Sampling techniques were employed to try to address the class imbalance in the dataset. The objective of these techniques was to improve classification accuracy by rebalancing the dataset in favour of the rare class or fatal accidents. The sampling techniques selected for this experiment research were:

1. Undersampling
2. Oversampling

In addition, no sampling, where a model was trained with training data but no sampling technique was selected, was considered. In the next phase of model building, a test design was generated.

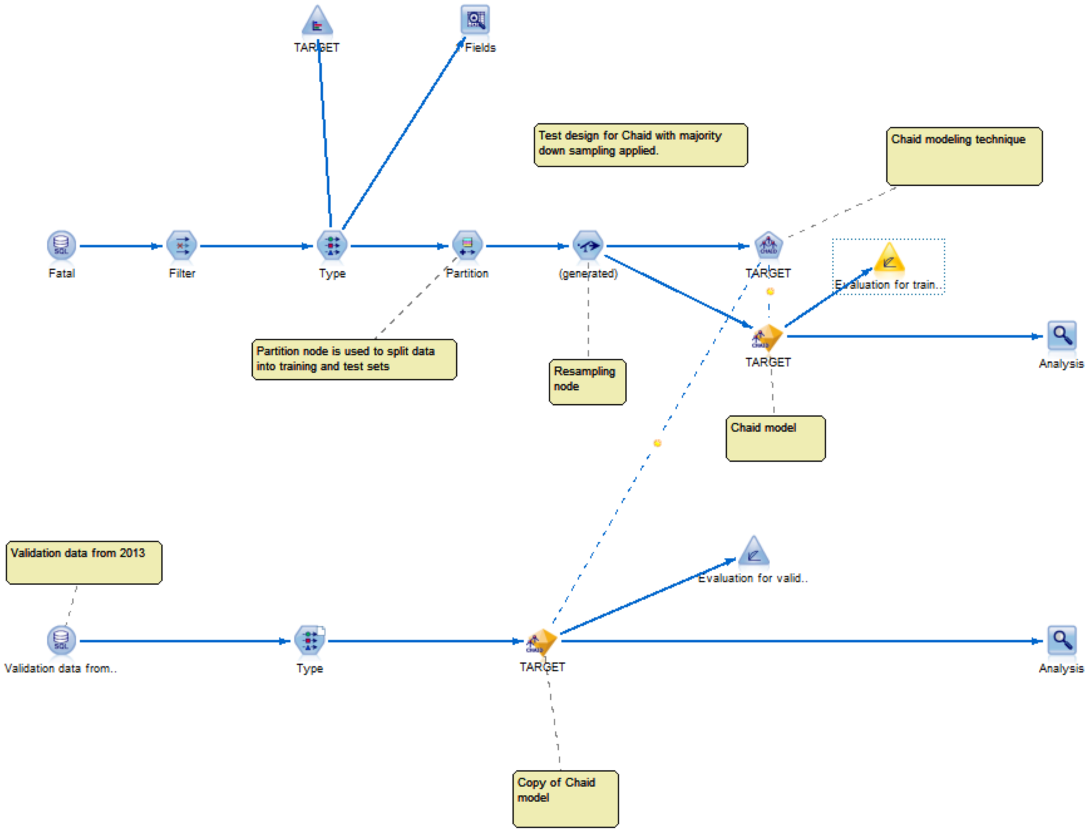
3.5.2 Test design

A test design or workflow was constructed in SPSS Modeler. This design was used to estimate model quality. A workflow is a series of interconnected nodes. The nodes in SPSS Modeler combine data and modelling techniques to generate an initial model. Fig. 3.6 was the test design workflow which used undersampling and the Chaid modelling technique. This workflow was adapted to work with other modelling and sampling techniques.

Typically models are measured in terms of overall accuracy e.g. the percentage of fatal and non-fatal classified accurately, however, this research was focused on fatal accident classification. Recall was the measure used to test fatal accident classification and was the key measure used for this research experiment. Model performance was presented using the confusion matrix and the matrix presented event counts for:

1. True positives (TP) – fatality occurred and was predicted
2. False positives (FP) – fatality did not occur and was predicted
3. True negatives (TN) – fatality did not occur and was predicted not to occur
4. False negatives (FN) – fatality occurred and was predicted not to occur
5. Total positives (P) – total fatal
6. Total negatives (N) – total on-fatal

In addition, recall was used to estimate quality by presenting the percentage of actual fatals classified as fatals. Confusion matrices were generated for each model using the SPSS Modeler analysis node. Two confusion matrices were generated for each model, one for training and one for test data. Recall was extrapolated from the confusion matrices. As part of the modelling process, the modelling data was split into two datasets. These datasets were described as training and testing data.



Test design for Chaid with majority down sampling applied.

Chaid modeling technique

Partition node is used to split data into training and test sets

Resampling node

Chaid model

Validation data from 2013

Copy of Chaid model

Figure 3. 6 Test design using resampling and Chaid modelling technique

SPSS Modeler has a capability called partitioning which automates the splitting of training and test data. The aim of partitioning is to provide a mechanism to test model quality. Models are built using training data and quality is estimated using test data. On the partition node, the training to test ratio must be selected. For this research experiment, the ratio was 80% training and 20% test. This means that the model was built from 80% of the available data and tested on the remainder.

The above design outlines a modelling framework which was used for all of the selected modelling techniques and allowed testing and performance metrics to be extracted for assessment. In the next chapter, the modelling techniques selected, their parameter settings and expected behaviour are discussed.

3.5.3 Model build

As previously discussed three modelling techniques were selected C5.0 and Chaid decision trees and Bayes net. When building each model, SPSS Modeler provided the option to calibrate model settings. Standard parameter settings were set for each modelling technique. Each technique was used to build models using training data and sampling combinations.

3.5.4 Model assessment

The objective of this chapter was to outline how the technical performance of models produced as part of this experiment was to be assessed. The SPSS Modeler analysis node was used to generate a confusion matrix. This node produced the confusion matrix in a table format which was easy to interpret and extract data from. The performance measures extracted from the confusion matrix were then recorded in a spreadsheet.

Accuracy

For each model under assessment, accuracy was considered the proportion of fatal and non-fatal which were classified correctly. Accuracy is calculated as follows:

Accuracy = (Count of TP + Count of TN) / (Count of P + Count of N).

For infrequent events there is often an adverse link between focus on recall at the expense of accuracy. Accuracy tends not to be the key concern when attempting to predict infrequent events as the focus is on predicting when infrequent event will occur rather than when it won't occur (Weiss & Hirsh, 2000). For this experiment, the focus was on predicting fatal accidents therefore recall was the priority as it measures the proportion of fatals classified as fatals. Data preparation and sampling techniques focussed on the fatal accident prediction rather than fatal and non-fatal prediction. It was therefore expected that accuracy rates may be low for this experiment, especially where sampling techniques had been applied. While overall accuracy of the models was assessed, the performance measure which was of most interest was recall.

Recall

Recall is the percentage of fatal accidents classified as fatals (Han, et al., 2011, p. 368). It is calculated as follows:

Recall = (Count of TP)/((Count of FN)+(Count of TP))

This was a key measure for this experiment which focussed on fatal accident prediction, recall was considered more important than accuracy as it identifies the true positive rate or proportion of fatal accidents correctly classified.

Precision

Precision is a measure of exactness being the percentage of fatals classified as fatals which are actually fatals or true positives (Han, et al., 2011, p. 368). Precision is calculated using the following formula:

Precision = (Count of TP)/((Count of TP)+(Count of FP))

Precision performance for prediction of infrequent events can be low when the focus is on prediction of the infrequent event. Sampling techniques to improve the recall for fatal accidents are likely to negatively affect the precision result as there "*tends to be an inverse relationship between precision and recall*" (Han, et al., 2011, p. 368).

Accuracy, recall and precision were extrapolated from the confusion matrix and were the key metrics from which comparisons were based. Performance was reviewed using the confusion matrix data. It was expected that the model estimates extrapolated from the confusion matrices for the training and test should be similar. The test estimate should underperform the training estimate as the modelling technique had never seen the test data. A large difference in performance can indicate a model which has under or overfitted the data. Overfitting occurs when a model fits the data too well and fails to generalise to new data. Underfitting occurs when a model fails to capture the underlying trend in the data and therefore fails to predict. Both under and over fitting lead to poor predictions.

To assist with the evaluation of models, the SPSS Modeler evaluation node was used to assess performance visually. The evaluation node generated an ROC curve and this ROC curve was used to visually assess the difference in classification accuracy between the training and test data.

3.6 Model Evaluation

In the model evaluation phase, the models and the processes followed to create the models as well as their practical use was reviewed.

3.6.1 Evaluation results

At this phase, the models were tested for usefulness towards achieving the goals of this research experiment. The main goals were to predict fatal traffic accidents and identify the factors which were most likely to predict fatal traffic accidents. As previously mentioned, recall was considered the most important measure given the focus on identifying fatal accidents. However, this focus on fatal was expected to impact on the performance results for accuracy and precision. To more thoroughly examine the performance of the models, the models were tested using 2013 STATS19 accident dataset or validation data. Two types of evaluation were conducted i.e. a technical and non-technical evaluation.

The technical evaluation used the same performance measures used at model assessment, except the performance results related to validation data. This data was separate from the previous model training and test data and related to a recent period in time. This made the evaluation more robust as it represented new data which the models have not previously seen.

The non-technical evaluation related to the interpretability of the model. In the case of decision trees, it should be possible to extract business rules from the selected model and use the rules to understand the relationships between the factors which contribute to fatal traffic accidents.

The final part of the results evaluation was to explicitly state whether the research goal of identifying the factors which were most likely to accurately predict fatal traffic accidents was achieved. Also consideration was given to other useful factors identified from each model.

3.6.2 Subsequent model improvements

At this stage the modelling work completed to date was reviewed and any flaws in workmanship were identified. The review process was outlined, findings highlighted and any issues identified which required immediate attention were addressed.

3.6.3 Key findings

This was the final step in the evaluation phase. The model was assessed to see whether the objective of predicting fatal accidents was achieved. A review of the model results was completed to identify the factors most likely to contribute to fatal traffic accidents or consideration was given to repeat some of the steps in the experiment methodology with a view to improving the quality of the models.

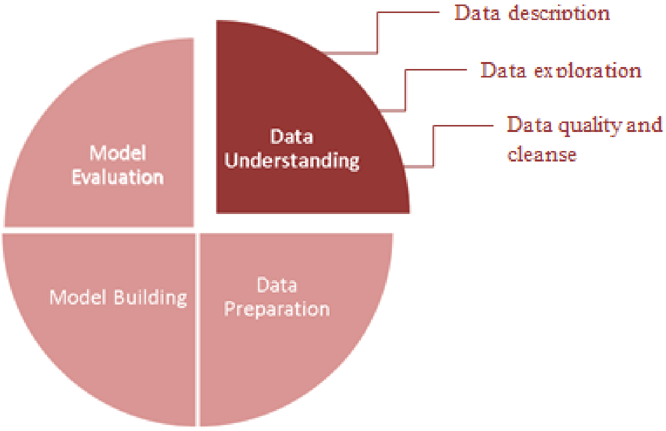
3.7 Conclusion

Experiment design was a key element of this dissertation as it outlined the methodology followed when implementing the experiment. Understanding the selected data and preparation of the data prior to model build helped to improve the quality and understanding of the information extracted from the models. Building the models was expected to be an iterative process in order to identify the optimum performing model. A consistent approach to model evaluation assisted in analysis and comparison of model results and performance.

4 EXPERIMENT IMPLEMENTATION

4.1 Introduction

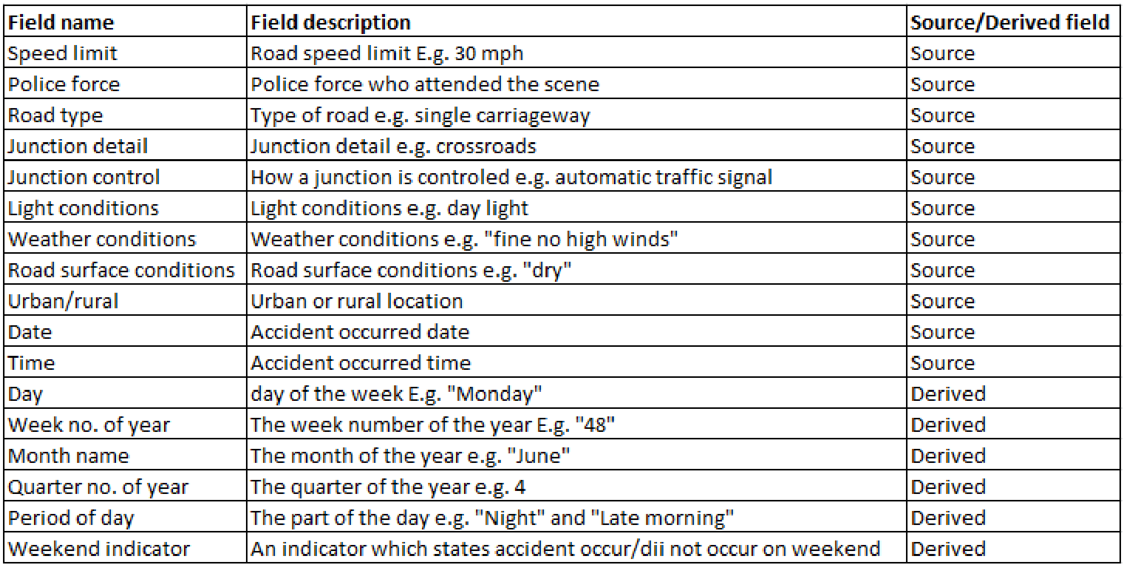
This research activity was focused on finding patterns in traffic accident characteristics which are specific to fatal traffic accidents. Initially, the descriptive statistics relating to fatal traffic accidents were created and explored to assist with data understanding prior to data construction. After data construction, predictive modelling techniques were selected and model designs were constructed. The modelling techniques selected for this experiment were from the decision tree and Bayesian families. C5.0 and Chaid decision trees and Bayes net were selected. These techniques were used to build classifiers to classify fatal traffic accidents. Following model build, twenty seven models were assessed using technical criteria and three were selected for further evaluation. The best three models were evaluated using technical and non-technical criteria. These models were used to better understand the factors which contribute to fatal traffic accidents.



4.2 Data Understanding

In data understanding, fatal accidents from the STATS19 data were explored to understand and discover patterns which relate to fatal traffic accidents. Descriptive statistics were used to summarise and describe data. Descriptive statistics were constructed using the SPSS Modeler and the audit node. These statistics were used to understand frequency counts, data groups and the distribution of data in each field reviewed. All descriptive statistics described in this chapter relate to an observation period from the year 2005 to 2012 inclusive. Data from this observation period was also used to build predictive models in the model build chapter. Table 4.1 presents the main fields which were explored.

Table 4. 1 STATS19 fields explored

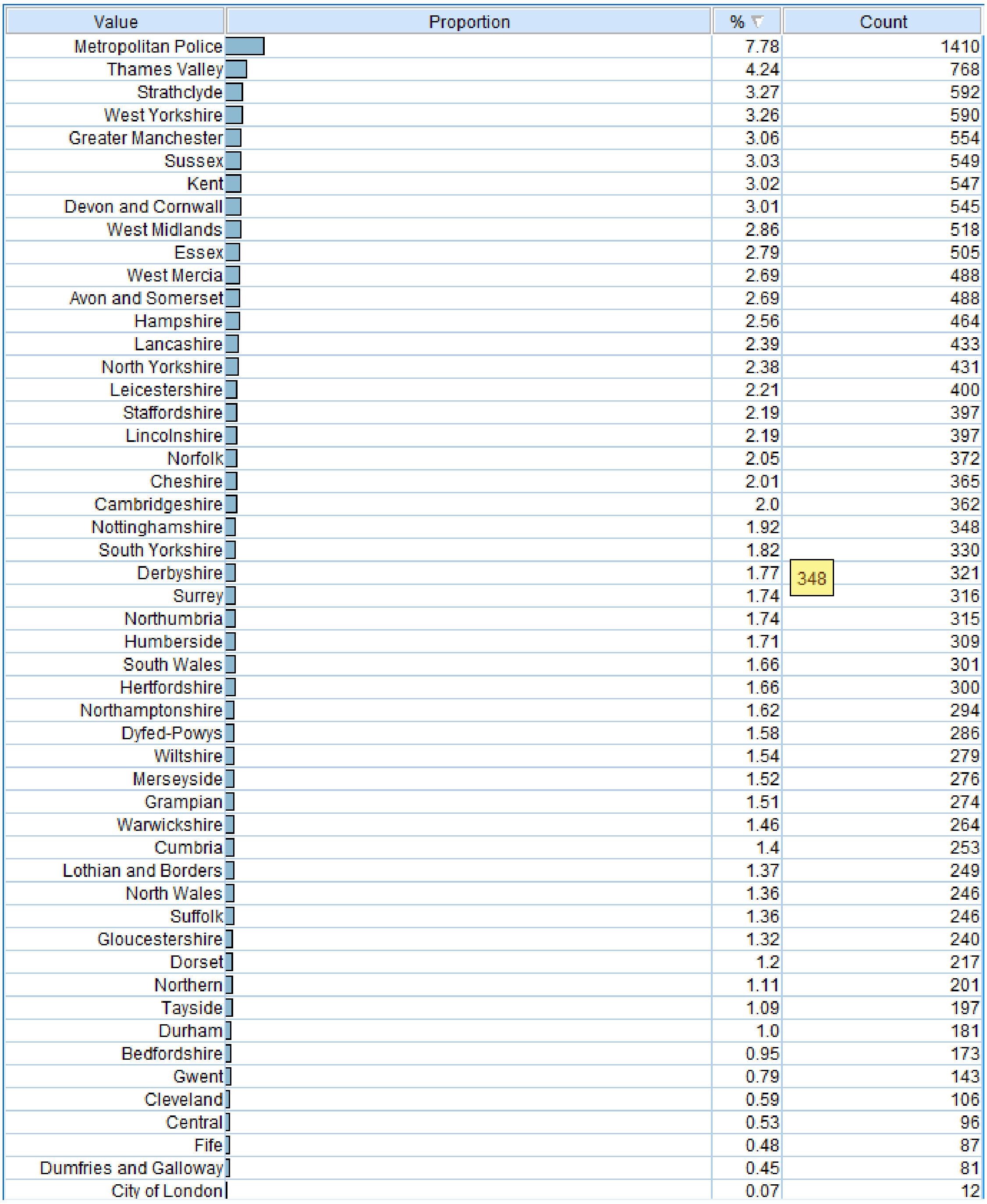


Field name	Field description	Source/Derived field
Speed limit	Road speed limit E.g. 30 mph	Source
Police force	Police force who attended the scene	Source
Road type	Type of road e.g. single carriageway	Source
Junction detail	Junction detail e.g. crossroads	Source
Junction control	How a junction is controled e.g. automatic traffic signal	Source
Light conditions	Light conditions e.g. day light	Source
Weather conditions	Weather conditions e.g. "fine no high winds"	Source
Road surface conditions	Road surface conditions e.g. "dry"	Source
Urban/rural	Urban or rural location	Source
Date	Accident occurred date	Source
Time	Accident occurred time	Source
Day	day of the week E.g. "Monday"	Derived
Week no. of year	The week number of the year E.g. "48"	Derived
Month name	The month of the year e.g. "June"	Derived
Quarter no. of year	The quarter of the year e.g. 4	Derived
Period of day	The part of the day e.g. "Night" and "Late morning"	Derived
Weekend indicator	An indicator which states accident occur/dii not occur on weekend	Derived

In order to develop an understanding of the data contained in the fatal dataset, it was necessary to load the data described in Table 4.1 into SPSS Modeler. Using SPSS Modeler and the data audit node, key fields were reviewed and important attributes, characteristics and prevalent features in the data were noted. The following chapter outlines the key summarised statistics by key field as produced by the data audit node.

Police force

Police force describes the police force which attended the scene of the accident recorded. In Fig. 4.1, the police with the top five largest proportions of accidents are Metropolitan Police, Thames Valley, Strathclyde, West Yorkshire and Greater Manchester. The fatality frequency ranges from 12 to 1410. The average fatality frequency is 355. Of the fifty one police forces included in the database, forty two recorded fatalities between 1% and 3.3%. The three main outliers were the City of London, Thames Valley and the Metropolitan Police with 0.07%, 4.2% and 7.8% respectively. The database does not provide any detail in relation to the population size covered by the various police forces, therefore, fatality percentage cannot on its own confirm a high risk area.



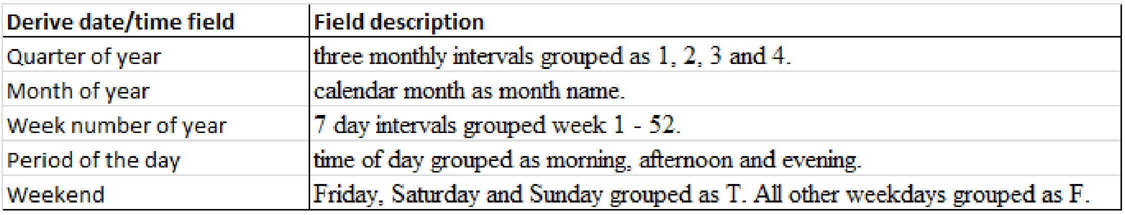
Value	Proportion	%	Count
Metropolitan Police		7.78	1410
Thames Valley		4.24	768
Strathclyde		3.27	592
West Yorkshire		3.26	590
Greater Manchester		3.06	554
Sussex		3.03	549
Kent		3.02	547
Devon and Cornwall		3.01	545
West Midlands		2.86	518
Essex		2.79	505
West Mercia		2.69	488
Avon and Somerset		2.69	488
Hampshire		2.56	464
Lancashire		2.39	433
North Yorkshire		2.38	431
Leicestershire		2.21	400
Staffordshire		2.19	397
Lincolnshire		2.19	397
Norfolk		2.05	372
Cheshire		2.01	365
Cambridgeshire		2.0	362
Nottinghamshire		1.92	348
South Yorkshire		1.82	330
Derbyshire		1.77	321
Surrey		1.74	316
Northumbria		1.74	315
Humberside		1.71	309
South Wales		1.66	301
Hertfordshire		1.66	300
Northamptonshire		1.62	294
Dyfed-Powys		1.58	286
Wiltshire		1.54	279
Merseyside		1.52	276
Grampian		1.51	274
Warwickshire		1.46	264
Cumbria		1.4	253
Lothian and Borders		1.37	249
North Wales		1.36	246
Suffolk		1.36	246
Gloucestershire		1.32	240
Dorset		1.2	217
Northern		1.11	201
Tayside		1.09	197
Durham		1.0	181
Bedfordshire		0.95	173
Gwent		0.79	143
Cleveland		0.59	106
Central		0.53	96
Fife		0.48	87
Dumfries and Galloway		0.45	81
City of London		0.07	12

Figure 4. 1 Fatal accident proportion by policing region

Date and time

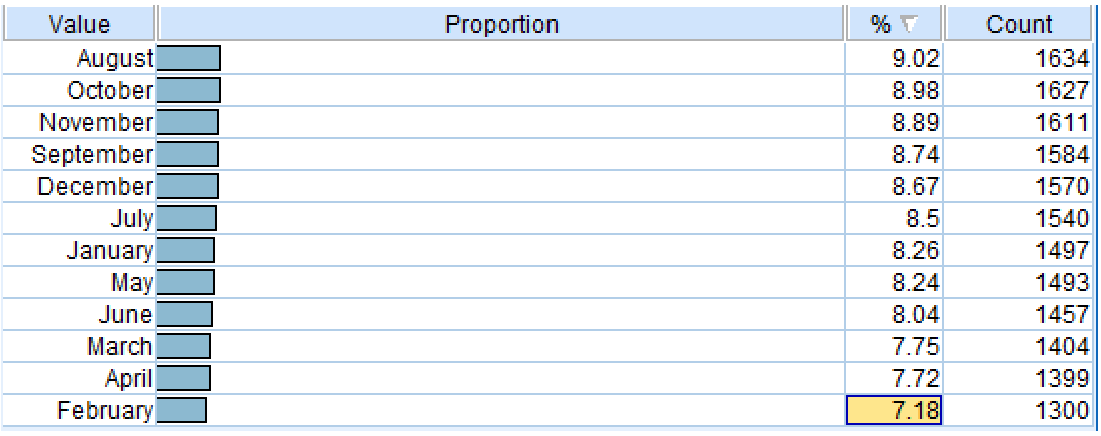
The fatal accident data records contain date and time information. These fields are granular and no useful information was found in these fields. (Wah, et al., 2012) in recent research, considered accident events in terms of the day an event occurred and time of event. They grouped day and time so they were considered as part of the week such as weekday and day period such as rush hour. To explore the relationship between fatal accident and date and time in the STATS19 data, the original date and time fields were grouped into derived fields. The new derived fields are described in Table 4.2.

Table 4. 2 Derived date/time fields



Derive date/time field	Field description
Quarter of year	three monthly intervals grouped as 1, 2, 3 and 4.
Month of year	calendar month as month name.
Week number of year	7 day intervals grouped week 1 - 52.
Period of the day	time of day grouped as morning, afternoon and evening.
Weekend	Friday, Saturday and Sunday grouped as T. All other weekdays grouped as F.

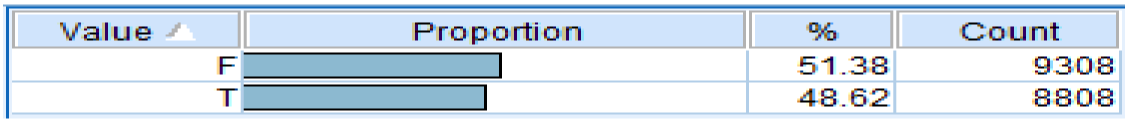
In Fig. 4.2 the traffic accident proportions for each month are presented. The average fatality frequency per month is 1,510. The majority of months were within 10% of this average. Fatality frequencies in February, March and April were over 8% lower than the average whereas August, October and November are over 8% above the average. August accounted for the highest number of fatalities at 1,634.



Value	Proportion	%	Count
August		9.02	1634
October		8.98	1627
November		8.89	1611
September		8.74	1584
December		8.67	1570
July		8.5	1540
January		8.26	1497
May		8.24	1493
June		8.04	1457
March		7.75	1404
April		7.72	1399
February		7.18	1300

Figure 4. 2 Proportion of fatalities per month

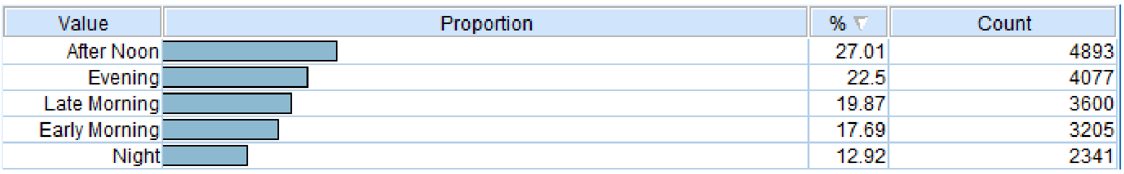
In Fig. 4.3 "T" refers to accidents which occurred on weekends (Friday, Saturday and Sunday) and "F" refers to accidents which occurred on week days. This analysis indicated that the risk of fatality is higher on weekdays, however, proportionally weekdays includes four days and weekends include three days.



Value	Proportion	%	Count
F		51.38	9308
T		48.62	8808

Figure 4. 3 Proportion of fatality on weekend days

In Fig. 4.4 fatal traffic accidents which occur at different times of the day are shown. Approx. 50% of accidents occurred in the afternoon and evening with the highest risk of fatality in the afternoon.

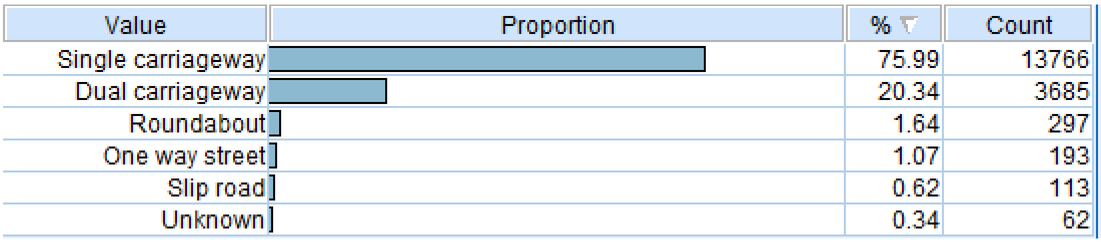


Value	Proportion	%	Count
After Noon		27.01	4893
Evening		22.5	4077
Late Morning		19.87	3600
Early Morning		17.69	3205
Night		12.92	2341

Figure 4. 4 Proportion of fatalities by time of day

Road type

Fig. 4.5 presents fatal accidents which occurred on different road types. The majority of fatal accidents occurred on single carriageways and dual carriageways. Three times as many fatal accidents occur on single carriageways when compared to dual carriageways. The proportions presented in Fig. 4.5 indicate that there was a significantly higher risk of fatality on carriageways when compared to all other road types.

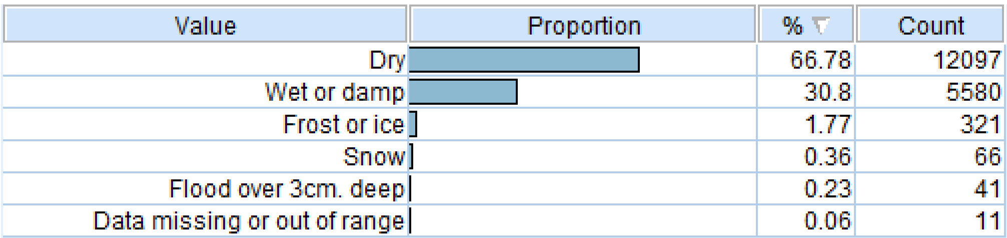


Value	Proportion	%	Count
Single carriageway		75.99	13766
Dual carriageway		20.34	3685
Roundabout		1.64	297
One way street		1.07	193
Slip road		0.62	113
Unknown		0.34	62

Figure 4. 5 Proportion of fatalities by road type

Road surface

Fig. 4.6 presents fatal accidents which occurred on different road surfaces. “Dry” account for 67%. “Wet or damp” account for 31%. There were at least twice as many “Dry” accidents as “Wet or damp” accidents.

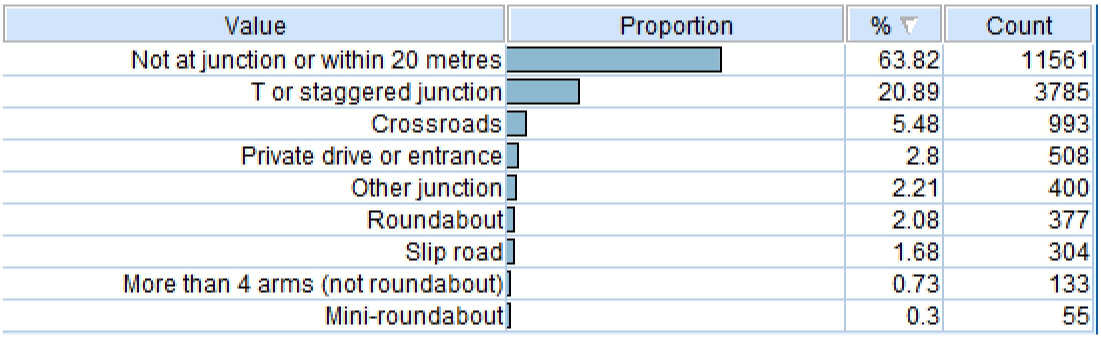


Value	Proportion	%	Count
Dry		66.78	12097
Wet or damp		30.8	5580
Frost or ice		1.77	321
Snow		0.36	66
Flood over 3cm. deep		0.23	41
Data missing or out of range		0.06	11

Figure 4. 6 Proportion of fatalities by road surface

Junction detail

Fig. 4.7 outlines fatalities which occurred relative to junctions. 64% of fatalities did not occur at or near a junction. T or staggered junctions accounted for 21% of fatalities and therefore were considered to be the highest risk junction type. Of the remaining junction types, cross roads were the most significant at 5% with the remaining junction types at less than 3% each.

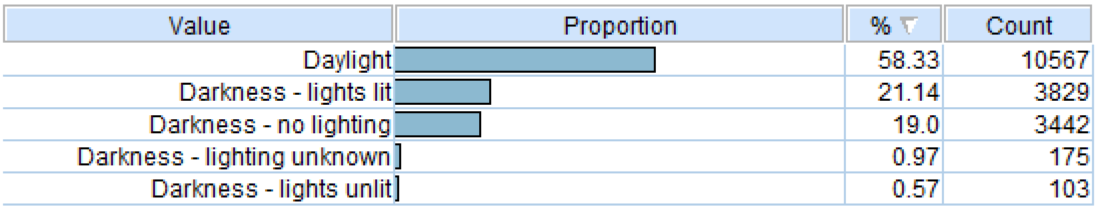


Value	Proportion	%	Count
Not at junction or within 20 metres		63.82	11561
T or staggered junction		20.89	3785
Crossroads		5.48	993
Private drive or entrance		2.8	508
Other junction		2.21	400
Roundabout		2.08	377
Slip road		1.68	304
More than 4 arms (not roundabout)		0.73	133
Mini-roundabout		0.3	55

Figure 4. 7 Proportion of fatalities by junction detail type

Light conditions

As per Fig. 4.8, 58% of fatal traffic accidents occurred in daylight and 42% occurred in darkness. Of fatal accidents which occurred in darkness, over half occurred when lights were lit. This analysis indicates that fatality risk in "Daylight" was more than twice as likely as during the "Darkness - lights lit".

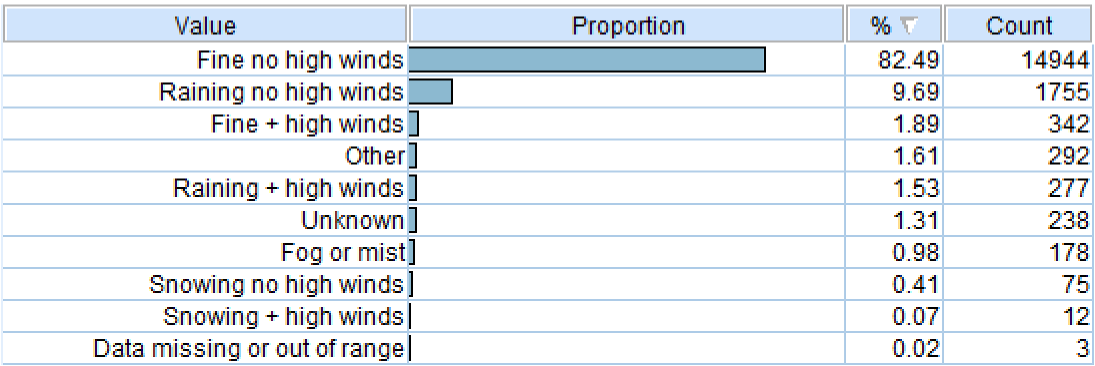


Value	Proportion	%	Count
Daylight		58.33	10567
Darkness - lights lit		21.14	3829
Darkness - no lighting		19.0	3442
Darkness - lighting unknown		0.97	175
Darkness - lights unlit		0.57	103

Figure 4. 8 Light conditions

Weather conditions

82% of fatal traffic accidents occurred in fine conditions with no high winds as displayed in Fig. 4.9. 10% occurred in rain conditions with no high winds. Other weather conditions ranged from 0.02% to 1.9% of fatal accidents.

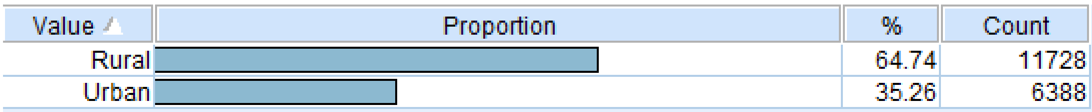


Value	Proportion	%	Count
Fine no high winds		82.49	14944
Raining no high winds		9.69	1755
Fine + high winds		1.89	342
Other		1.61	292
Raining + high winds		1.53	277
Unknown		1.31	238
Fog or mist		0.98	178
Snowing no high winds		0.41	75
Snowing + high winds		0.07	12
Data missing or out of range		0.02	3

Figure 4. 9 Wind conditions

Urban or rural conditions

65% of fatalities occurred in rural areas while 35% occurred in urban areas as outlined in Fig. 4.10. This analysis indicated that fatality risk in rural areas is over 1.8 times more likely.



Value	Proportion	%	Count
Rural		64.74	11728
Urban		35.26	6388

Figure 4. 10 Urban or rural conditions

4.3 *Data Exploration*

The previous chapters considered each field individually. In the exploratory analysis, the SPSS Modeller web analysis node describes the relationships between fields. Aggregations are counts of the number of occurrences of a relationship between two fields. Those with the strongest relationship are presented as thicker lines. Aggregations were constructed in Structured Query Language (SQL), a standard language which helps to query data, and were used to validate findings. Fig. 4.11 displays the strongest data relationships identified in the STATS19 data.

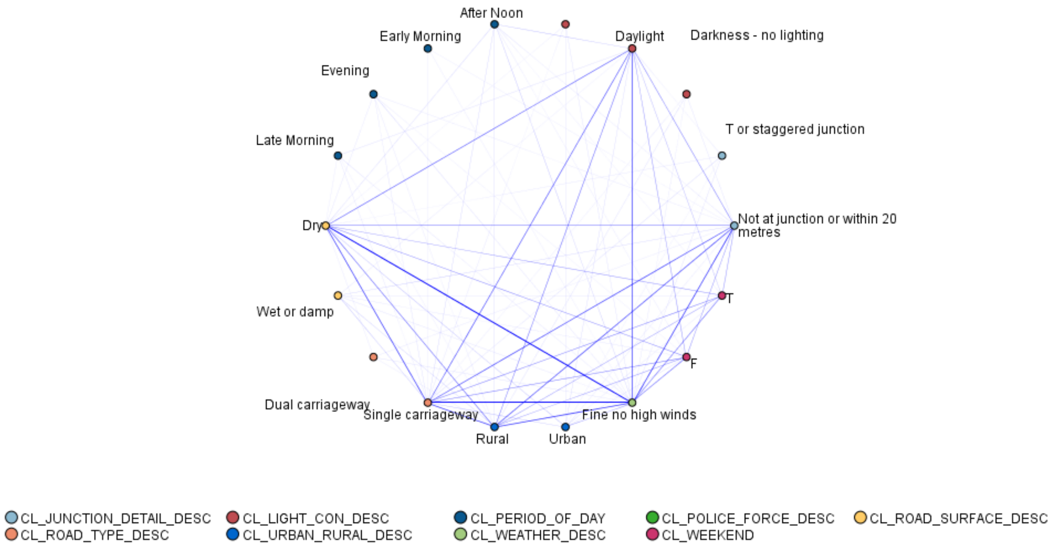


Figure 4. 11 Initial data relationships identified

The relationships identified, ordered by strength i.e. the number of times the relationship occurred, are as follows:

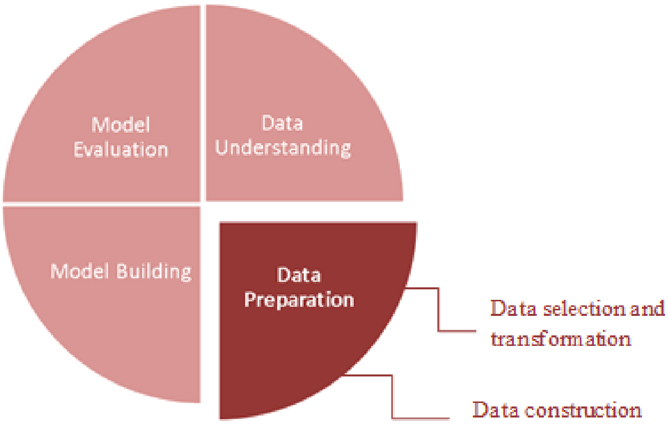
1. 11,653 fatalities occur in fine weather with no high winds on dry road surface conditions.
2. 11,370 fatalities occur on single carriageways in fine weather with no high winds.
3. 9,364 fatalities occur in fine weather with no high winds and do not occur at junctions.
4. 8,778 fatalities occur on single carriageways and not at junctions.
5. 8,735 fatalities occur in rural areas and not at junctions.
6. 8,257 fatalities occur in daylight and on single carriageways.
7. 7,094 fatalities occur in daylight and in dry road surface conditions.
8. 6,422 fatalities occur in day light and not at junctions.
9. 3,258 fatalities occur on single carriageways and at T or staggered junctions.
10. 2,675 fatalities occur in urban area when its dark and light are lit.
11. 2,625 fatalities occur on dual carriageways and not at junctions.

This exploration identifies individual strong relationships between two data fields. The limitation of this initial exploration is that it does not identify multi-layer relationships. It does, however, highlight significant data points which are useful for the model building phase. Some features in the STATS19 dataset, which seem to be good predictors of fatal traffic accidents, are similar to the features considered by (Wah, et al., 2012) as significant factors contributing to severe motor cycle accidents. These include weather, light and road conditions.

4.4 Data Quality and Cleanse

As discussed in chapter 3, the STATS19 dataset is widely used for traffic accident statistics in the UK and is considered good quality with regard to accuracy and completeness. Fatal accidents are specifically noted as being well recorded and complete. Given the STATS19 data quality had already been independently assessed as high, the requirements for standard data quality and data cleanse techniques was reduced for this experiment. As a result, fatal accident data was considered clean, well recorded and complete.

A review of fatal accident incidents was assessed for completeness and correctness. For fatal accident incidents in STATS19, all fields were complete, however, a small proportion of missing data had already been categorised as missing by the Department of Transport. From a completeness point of view, this was positive, however, missing and unknown data will add little to the predictive capability of the model. The SPSS data audit node was used to identify missing and correct data. No incorrect data was identified.



4.5 Data Selection and Transformation

Before feature selection, data relating to the observation period was grouped into fatal and non-fatal. Fatal were the fatal accidents and non-fatal were serious and slight accidents. A single target was created in SPSS for fatal and non-fatal. The target was required so Chaid could separate fatal and non-fatal and in addition group predictor information. The Chaid decision tree is based on the Chi-squared statistic of proportion which splits data into groups. In this experiment research, these groups are used to construct training data.

Once the target was defined, the next step was to select the data points which had the most predictive importance and group the data points more effectively to improve predictive performance. This phase is referred to as data transformation.

4.5.1 Selecting candidate features for data transformation

In chapter 3, Fig. 3.4 described the discretisation process followed for this experiment. Chaid was configured to target fatal and non-fatal in order to instruct Chaid which characteristics to target. The Chaid technique was run against the STATS19 data allowing the tool to identify patterns in the data relating to fatal and non-fatal. A decision tree was induced which presented the relationship of fatal and non-fatal to the key predictors. The key predictors at the top of the decision tree were noted together with the category levels at each branch and the process was repeated six times. SPSS Modeler decision trees are large with multiple branches and therefore are difficult to present in a single chart. Fig. 4.12 displays an extract of the target of fatal as 1 and non-fatal as 0, and showing "Road Surface" as the top of the tree as the key initial predictor.

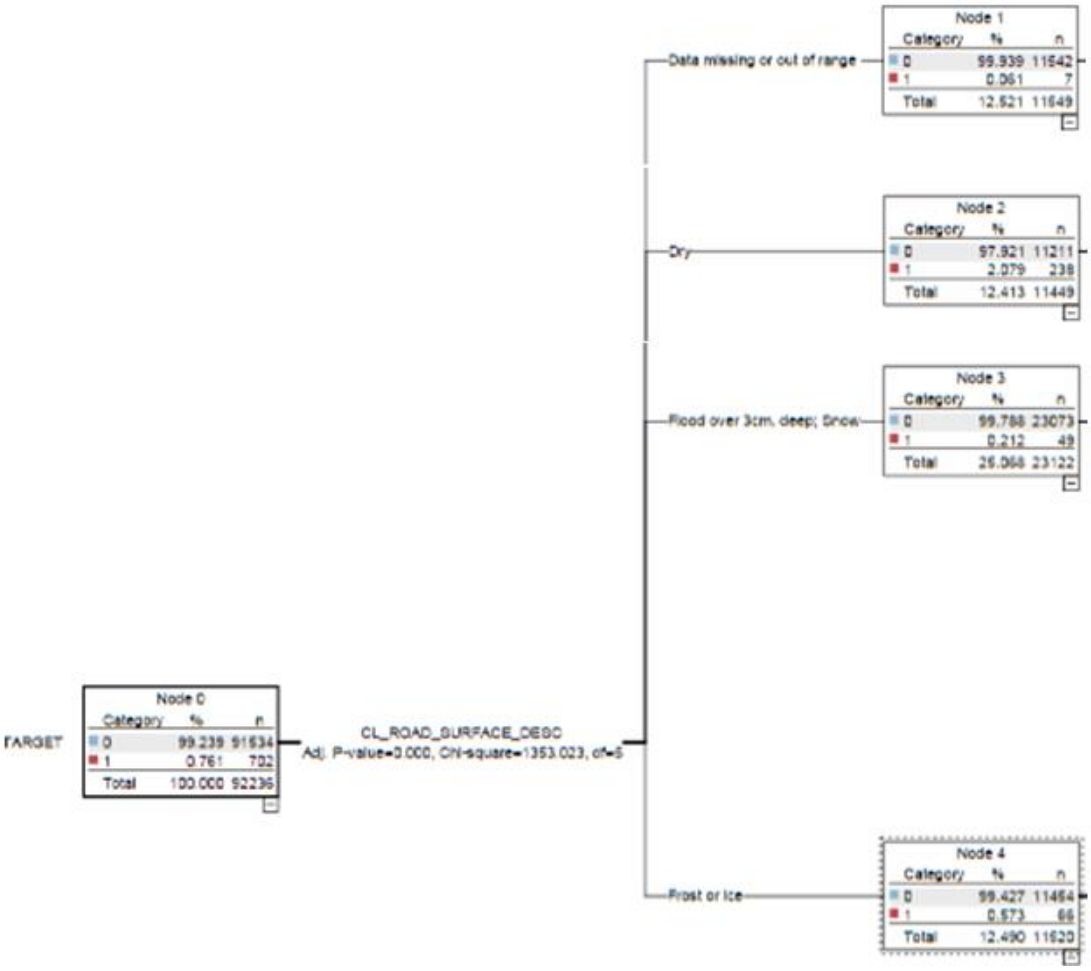
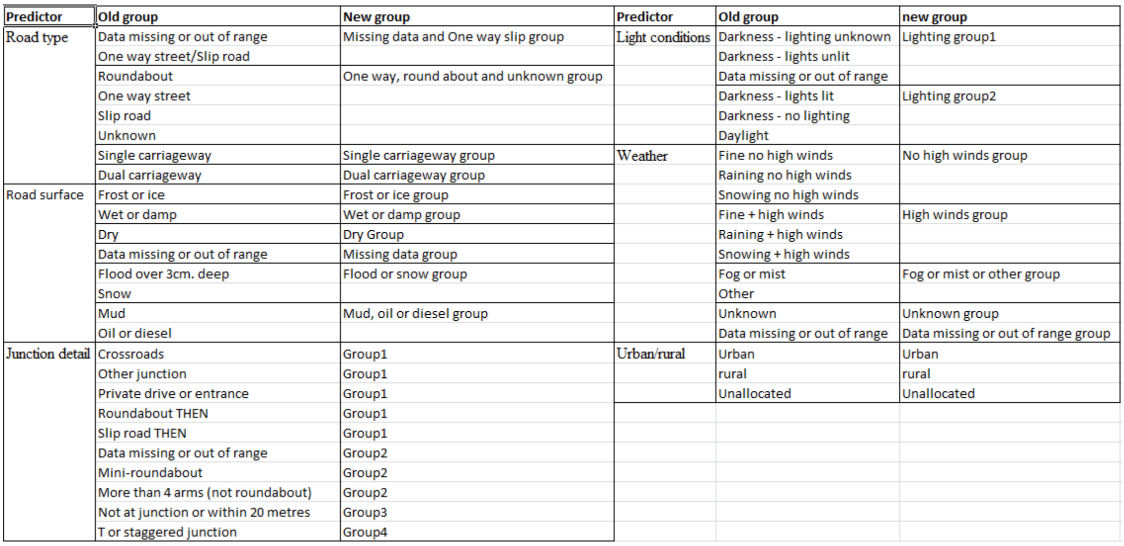


Figure 4. 12 Extract Chaid decision tree

After executing this procedure, a short list of candidate predictors and groups identified was compiled. The most important predictors identified for old groups and new groups are listed in Table 4.3. Old groups are the original groups provided in the STATS19 data and the new groups were identified using Chaid as being the most homogeneous based on their characteristics. In addition a combination of old and new groups was also considered in case the combination could help improve prediction accuracy and interpretability. In the next phase in data construction, training data was constructed using old groups, new groups as well as the combination of old and new groups. In the data construction stage, these are referred to as normal data i.e. no reduction, reduced data and semi reduced data respectively.

Table 4. 3 Most important predictors, old groups and new groups

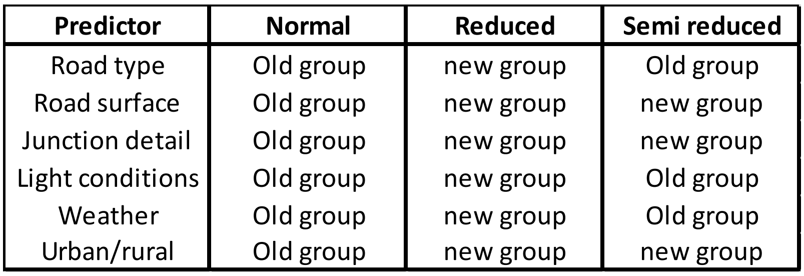


Predictor	Old group	New group	Predictor	Old group	new group
Road type	Data missing or out of range	Missing data and One way slip group	Light conditions	Darkness - lighting unknown	Lighting group1
	One way street/Slip road			Darkness - lights unlit	
	Roundabout	One way, round about and unknown group		Data missing or out of range	
	One way street			Darkness - lights lit	Lighting group2
	Slip road			Darkness - no lighting	
	Unknown			Daylight	
	Single carriageway	Single carriageway group	Weather	Fine no high winds	No high winds group
	Dual carriageway	Dual carriageway group		Raining no high winds	
Road surface	Frost or ice	Frost or ice group		Snowing no high winds	
	Wet or damp	Wet or damp group		Fine + high winds	High winds group
	Dry	Dry Group		Raining + high winds	
	Data missing or out of range	Missing data group		Snowing + high winds	
	Flood over 3cm. deep	Flood or snow group		Fog or mist	Fog or mist or other group
	Snow			Other	
	Mud	Mud, oil or diesel group		Unknown	Unknown group
	Oil or diesel			Data missing or out of range	Data missing or out of range group
Junction detail	Crossroads	Group1	Urban/rural	Urban	Urban
	Other junction	Group1		rural	rural
	Private drive or entrance	Group1		Unallocated	Unallocated
	Roundabout THEN	Group1			
	Slip road THEN	Group1			
	Data missing or out of range	Group2			
	Mini-roundabout	Group2			
	More than 4 arms (not roundabout)	Group2			
	Not at junction or within 20 metres	Group3			
	T or staggered junction	Group4			

4.6 Data Construction

Three datasets were constructed named normal, reduced and semi reduced data. Normal used old group data, reduced used new group data and semi reduced used a combination of old and new group data as defined in chapter 3.

Table 4. 4 Groups used to construct normal, reduced and semi reduced data



Predictor	Normal	Reduced	Semi reduced
Road type	Old group	new group	Old group
Road surface	Old group	new group	new group
Junction detail	Old group	new group	new group
Light conditions	Old group	new group	Old group
Weather	Old group	new group	Old group
Urban/rural	Old group	new group	new group

Table 4.4 describes the predictors and group combinations used to construct normal, reduced and semi reduced training data.

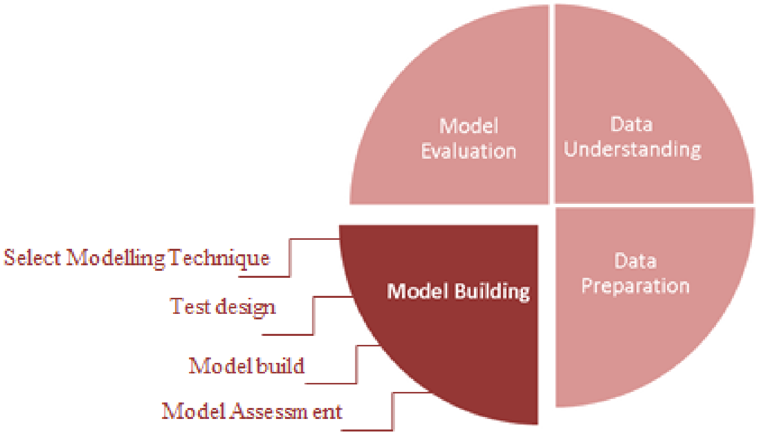
Training data construction

The process followed to construct each of the three training datasets is outlined below:

1. Application logic was built to return the “old group” or “new group” data for each predictor and its related factors as outlined in Table 4.4.
2. The type of training data was selected e.g. normal, reduced or semi-reduced.
3. Using the data selected in step 2, fatal traffic accidents only were filtered for the observation period, 2005-2012. A dataset for fatal cases only was constructed based on the six predictors outlined in Table 4.4.

4. Application logic was written to calculate the frequency count for each unique fatal accident factor combination. The frequency count was used to identify the number of times a particular combination occurred. The output of this step was a fatal accident dataset, with frequency count for each combination.
5. Application logic was then written to construct all factor combinations using the predictors in Table 4.4.
6. Application logic was written to construct an additional dataset which was the difference between the output of steps 4 and 5. This dataset was the non-fatal accident dataset. The frequency count for the non-fatal combinations was zero.
7. The output of steps 4 and step 6 were merged to form the training dataset.

In the steps above, where application logic is mentioned, the logic was constructed on an Oracle database and Oracle database views were created to get group values for normal, reduced and semi reduced data. Oracle stored procedures were used to construct the training data.



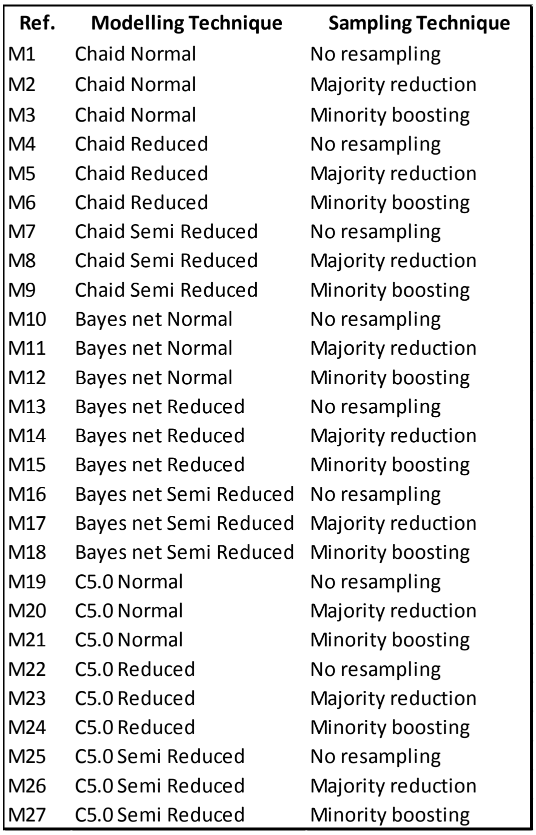
4.7 Model Build

For the purpose of this research, predictive models were built from three training datasets, normal, reduced and semi reduced, as described in the previous chapter. Nine models were constructed for each classification model for C5.0 decision tree, Chaid decision tree and Bayes net. The models created, with the reference modelling technique and relevant sampling technique, are listed in Table. 4.5.

Each dataset was noted as imbalanced in favour of the majority class which was non-fatal. It was therefore a key stage of this research to rebalance the datasets prior to building the predictive models. The techniques used for rebalancing were:

- Majority reduction
- Minority boosting

Table 4. 5 Experiment classification model listing



Ref.	Modelling Technique	Sampling Technique
M1	Chaid Normal	No resampling
M2	Chaid Normal	Majority reduction
M3	Chaid Normal	Minority boosting
M4	Chaid Reduced	No resampling
M5	Chaid Reduced	Majority reduction
M6	Chaid Reduced	Minority boosting
M7	Chaid Semi Reduced	No resampling
M8	Chaid Semi Reduced	Majority reduction
M9	Chaid Semi Reduced	Minority boosting
M10	Bayes net Normal	No resampling
M11	Bayes net Normal	Majority reduction
M12	Bayes net Normal	Minority boosting
M13	Bayes net Reduced	No resampling
M14	Bayes net Reduced	Majority reduction
M15	Bayes net Reduced	Minority boosting
M16	Bayes net Semi Reduced	No resampling
M17	Bayes net Semi Reduced	Majority reduction
M18	Bayes net Semi Reduced	Minority boosting
M19	C5.0 Normal	No resampling
M20	C5.0 Normal	Majority reduction
M21	C5.0 Normal	Minority boosting
M22	C5.0 Reduced	No resampling
M23	C5.0 Reduced	Majority reduction
M24	C5.0 Reduced	Minority boosting
M25	C5.0 Semi Reduced	No resampling
M26	C5.0 Semi Reduced	Majority reduction
M27	C5.0 Semi Reduced	Minority boosting

SPSS Modeler sampling capabilities were used to reduce the majority class so that both non-fatals and fatals would represent approximately 50% of the newly formed "majority reduction" dataset. A similar technique was applied to increase the minority class, fatals, to 50% of a new "minority boosting" dataset. Twenty seven training datasets were initially constructed from combinations of the following data and sampling techniques:

- Normal, reduced and semi-reduced data
- Majority reduction, minority boosting and no resampling

Each observation in normal, reduced or semi reduced data was classified as fatal or non-fatal. A positive frequency indicated fatal and a frequency of zero for non-fatal.

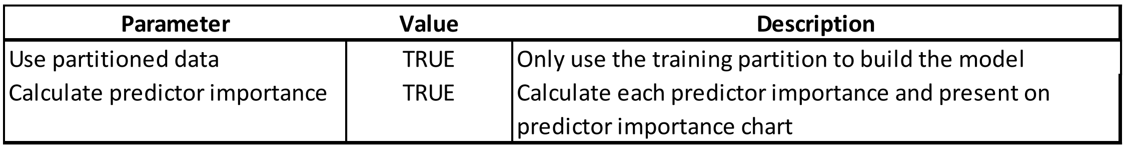
To build a classification model, an outcome variable or target is required. The outcome variable differentiates between fatal and non-fatal. Initially, a categorical outcome was created similar to the approach followed by (Wah, et al., 2012). Zero frequencies indicate non-fatal, low indicates fatal in the frequency range 1–20. All other frequencies are high. Modelling zero, low and high frequency accidents produced good prediction results, however, model results were difficult to interpret. It was therefore decided to treat accident outcome as a dichotomous outcome. This meant that the outcome was true or false. A positive frequency was true while a zero frequency was false. An accident outcome of true indicated fatal while an outcome of false indicated non-fatal. Using this approach, initial prediction results and model interpretability improved.

For this experiment accident outcome was dichotomous so must be either fatal or non-fatal. This outcome was set as the prediction target and was used to direct model training. Normal, reduced and semi reduced data was the base data used to build models. Sampling techniques complimented the base data by addressing class imbalance issues and shaped the data aiming to build better predictive models.

Predictive models were built following the test design described in chapter 3 and training, test and validation datasets were created. The SPSS Modeler partition node was used to create training and test data. Validation data, also referred to as hold out data, was created separately. Validation data was used to comprehensively test the model by evaluating performance measures using a dataset not previously seen by the model. Training data was used to build models, modelling performance was estimated on test data and validation data was used to provide an unbiased estimate of model performance. Data and modelling techniques were combined to build predictive models. Model performance was assessed using the confusion matrix. Additional measures such as accuracy, recall and precision were extrapolated. Accuracy was the accident classification rate for fatal and non-fatal accidents. Recall was the key measure of interest for this experiment and was used to measure the true fatal classification rate.

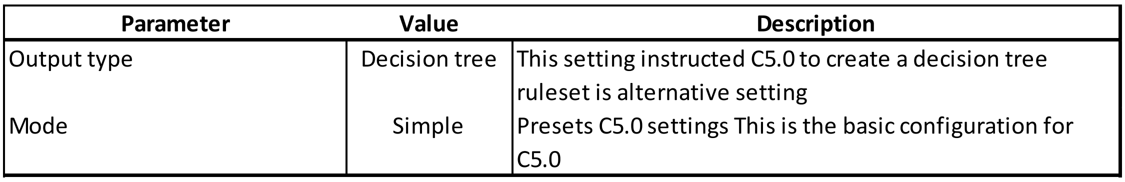
The parameter settings applied in SPSS Modeler for each modelling technique used for this research experiment are described below. The settings that are applicable to all models are described in Table 4.6 and settings specific to C5.0, Chaid and Bayes net are outlined in Tables 4.7, 4.8 and 4.9 respectively.

Table 4. 6 SPSS settings applicable to all models



Parameter	Value	Description
Use partitioned data	TRUE	Only use the training partition to build the model
Calculate predictor importance	TRUE	Calculate each predictor importance and present on predictor importance chart

Table 4. 7 SPSS C5.0 specific settings



Parameter	Value	Description
Output type	Decision tree	This setting instructed C5.0 to create a decision tree ruleset is alternative setting
Mode	Simple	Presets C5.0 settings This is the basic configuration for C5.0

Table 4. 8 SPSS Chaid specific settings

Parameter	Value	Description
Levels below root	5	Number of times the sample will be split
Alpha for splitting	0.05	Significance level for splitting nodes
Alpha for merging	0.05	Significance level for for merging categories
Maximim iterations for convergence	100	Maximum number of iteration before stopping even if convergence did not occur
Use Bonferroni adjustment	TRUE	Adjusted significance levels when testing category combinations
Chi-square	Pearson	Use Pearson to calculate the Chi-square statistic
Minimum records in parent branch	2%	Minimum proportion of records which should be in a parents node before splitting
Minimum records in child branch	1%	Minimum proportion of records which should be in a child node before splitting

Table 4. 9 SPSS Bayes net specific settings

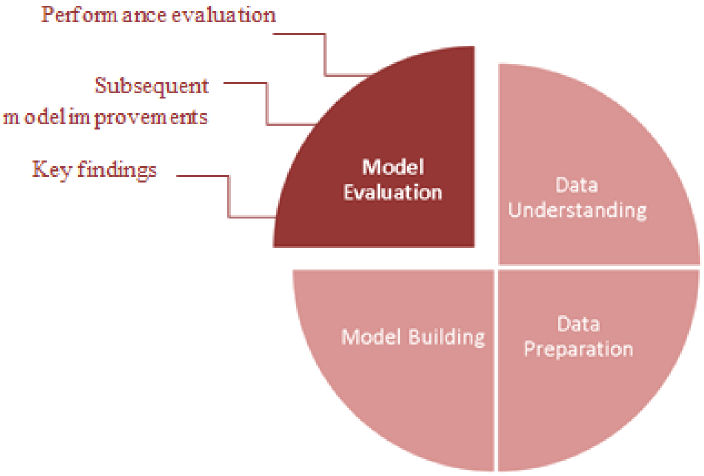
Parameter	Value	Description
Structure type	TAN	Build a Tree Augmented Naïve Bayes network model
Parameter learning method	Likelihood	Use likelihood to control estimating conditional probabilities between nodes
Mode	Simple	
Use only complete records	TRUE	This setting instructs Bayes net to only use complete records
Independence test	likelihood ratio	Use likelihood ratio to assess if paired observations are independent
Significance level	0.01	This setting is used by the independence test to set a cut off value

Initially, models were assessed for technical performance using accuracy, recall, precision measures and ROC curves. These measures were created for training, test and validation data. Validation data was based on data from STATS19 for 2013, however, this data was not used to train models. The results from validation data were therefore unbiased. The next chapter discusses the experiment evaluation. Models were assessed and compared based on performance and evaluation measures and interpretability.

4.8 Conclusion

This chapter describes the stages of the experiment implementation. An understanding of the data was developed by analysing the data in the STATS19 dataset. Initial relationships in the data were explored using SPSS Modeler web analysis node. The data quality assessment was completed prior to grouping the data into two target values, fatal and non-fatal. Data transformation was completed using Chaid to group data more effectively to improve the predictive performance of the target values. Training data was constructed for normal, reduced and semi reduced data. C5.0, Chaid and Bayes net models were built for each and sampling techniques for majority reduction, minority boosting and no resampling which were applied. Twenty seven models in total were built using SPSS Modeler.

5 EXPERIMENT EVALUATION



5.1 *Introduction*

A summary of twenty seven model results are presented in this chapter. Nine models were trained for each of the three training datasets, normal, reduced and semi reduced. Performance was assessed using technical measures including accuracy, recall and precision, which were previously discussed. The technical performance for the twenty seven models is initially assessed and the results are summarised in a confusion matrix, together with an overview of the models by classification techniques and ROC curves.

Validation performance results are based on unbiased data and therefore were a true measure of performance. The best performing model for each classification techniques in the assessment phase were selected and the three selected models are further evaluated using validation data based on 2013 STATS19 which the model had not previously seen. Evaluation involved technical and non-technical performance testing and was completed based on accuracy, recall and precision in line with the model assessment. Non-technical evaluation assessed the models interpretability and checked if the model addresses the experiment objectives. Following consideration of the evaluation results, model improvements were tested to identify if improved performance may be possible and considered for future research. Finally, the key findings from the experiment evaluation were discussed and the conclusion summarises the chapter.

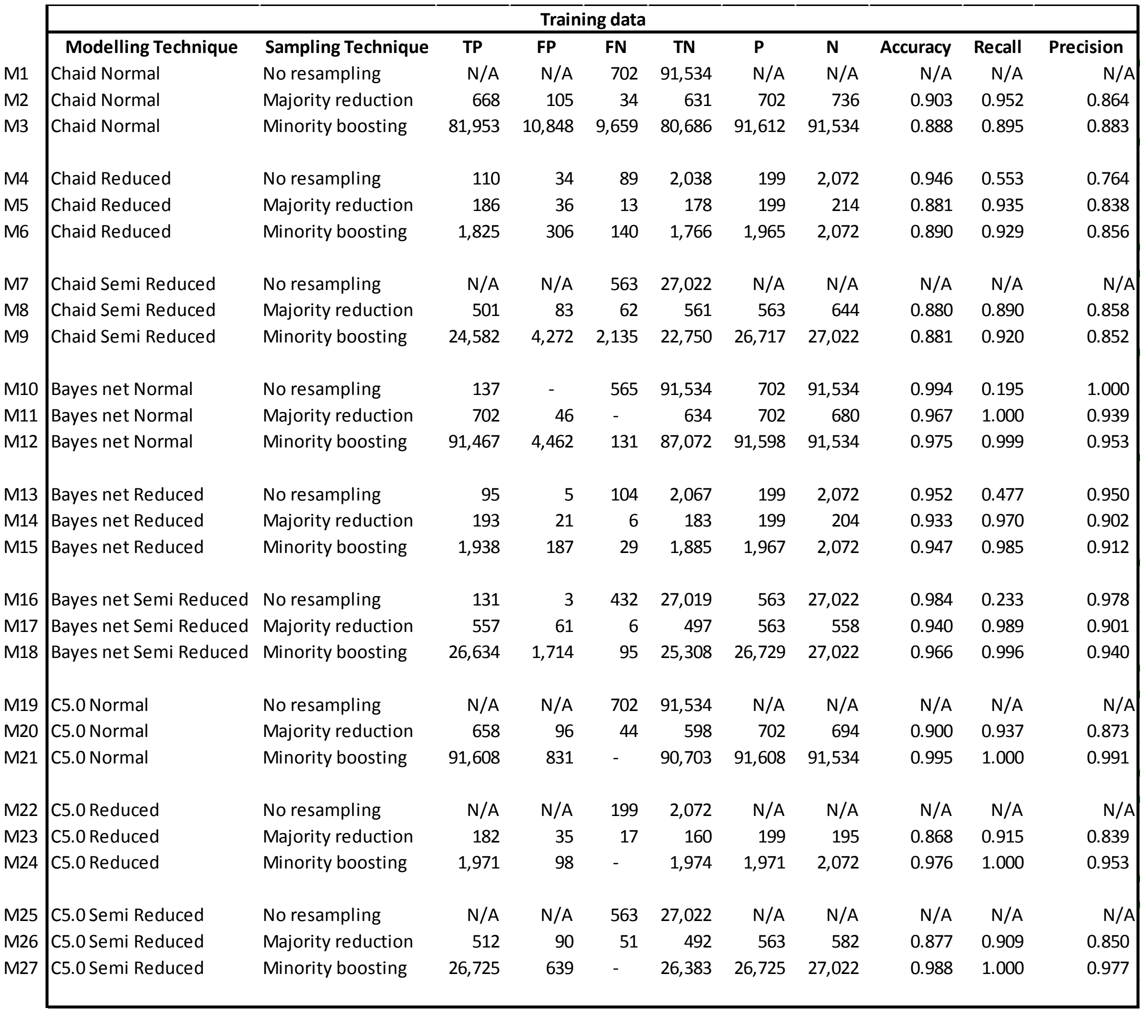
5.1 *Model Assessment on Training and Test Data*

A summary of twenty seven model results are presented in this chapter. Nine models were trained for each of the three training datasets, normal, reduced and semi reduced. These twenty seven models were initially assessed for their technical performance and the results were summarised. The initial technical performance of the models was assessed based on both the training and test data, giving two results per model. A confusion matrix was used to extrapolate accuracy and recall performance. The confusion matrix and an ROC curve were used to visually assess the difference in performance between training and test data.

5.1.1 Assessment confusion matrix

A confusion matrix was produced from SPSS Modeler analysis node for training and test data for each modelling technique and sampling technique and for the three datasets, normal, reduced and semi reduced. Table 5.1 and 5.2 below summarise the confusion matrices produced by SPSS Modeler and shows the accuracy and recall performance measures for the twenty seven models for training and test data respectively.

Table 5. 1 Confusion matrix and performance measures for training data



	Training data											
	Modelling Technique	**Sampling Technique**	**TP**	**FP**	**FN**	**TN**	**P**	**N**	**Accuracy**	**Recall**	**Precision**
M1	Chaid Normal	No resampling	N/A	N/A	702	91,534	N/A	N/A	N/A	N/A	N/A
M2	Chaid Normal	Majority reduction	668	105	34	631	702	736	0.903	0.952	0.864
M3	Chaid Normal	Minority boosting	81,953	10,848	9,659	80,686	91,612	91,534	0.888	0.895	0.883
M4	Chaid Reduced	No resampling	110	34	89	2,038	199	2,072	0.946	0.553	0.764
M5	Chaid Reduced	Majority reduction	186	36	13	178	199	214	0.881	0.935	0.838
M6	Chaid Reduced	Minority boosting	1,825	306	140	1,766	1,965	2,072	0.890	0.929	0.856
M7	Chaid Semi Reduced	No resampling	N/A	N/A	563	27,022	N/A	N/A	N/A	N/A	N/A
M8	Chaid Semi Reduced	Majority reduction	501	83	62	561	563	644	0.880	0.890	0.858
M9	Chaid Semi Reduced	Minority boosting	24,582	4,272	2,135	22,750	26,717	27,022	0.881	0.920	0.852
M10	Bayes net Normal	No resampling	137	-	565	91,534	702	91,534	0.994	0.195	1.000
M11	Bayes net Normal	Majority reduction	702	46	-	634	702	680	0.967	1.000	0.939
M12	Bayes net Normal	Minority boosting	91,467	4,462	131	87,072	91,598	91,534	0.975	0.999	0.953
M13	Bayes net Reduced	No resampling	95	5	104	2,067	199	2,072	0.952	0.477	0.950
M14	Bayes net Reduced	Majority reduction	193	21	6	183	199	204	0.933	0.970	0.902
M15	Bayes net Reduced	Minority boosting	1,938	187	29	1,885	1,967	2,072	0.947	0.985	0.912
M16	Bayes net Semi Reduced	No resampling	131	3	432	27,019	563	27,022	0.984	0.233	0.978
M17	Bayes net Semi Reduced	Majority reduction	557	61	6	497	563	558	0.940	0.989	0.901
M18	Bayes net Semi Reduced	Minority boosting	26,634	1,714	95	25,308	26,729	27,022	0.966	0.996	0.940
M19	C5.0 Normal	No resampling	N/A	N/A	702	91,534	N/A	N/A	N/A	N/A	N/A
M20	C5.0 Normal	Majority reduction	658	96	44	598	702	694	0.900	0.937	0.873
M21	C5.0 Normal	Minority boosting	91,608	831	-	90,703	91,608	91,534	0.995	1.000	0.991
M22	C5.0 Reduced	No resampling	N/A	N/A	199	2,072	N/A	N/A	N/A	N/A	N/A
M23	C5.0 Reduced	Majority reduction	182	35	17	160	199	195	0.868	0.915	0.839
M24	C5.0 Reduced	Minority boosting	1,971	98	-	1,974	1,971	2,072	0.976	1.000	0.953
M25	C5.0 Semi Reduced	No resampling	N/A	N/A	563	27,022	N/A	N/A	N/A	N/A	N/A
M26	C5.0 Semi Reduced	Majority reduction	512	90	51	492	563	582	0.877	0.909	0.850
M27	C5.0 Semi Reduced	Minority boosting	26,725	639	-	26,383	26,725	27,022	0.988	1.000	0.977

Table 5. 2 Confusion matrix and performance measures for test data

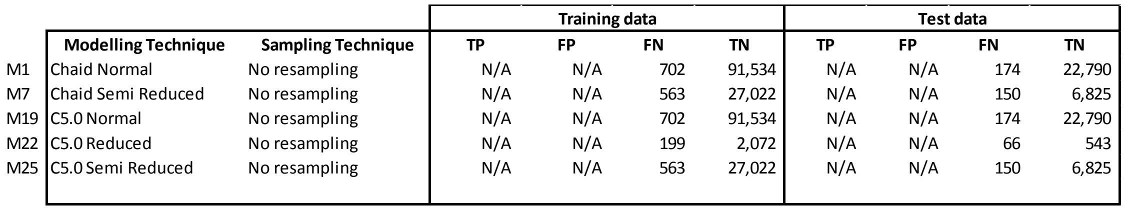
Test data										
Modelling Technique	**Sampling Technique**	**TP**	**FP**	**FN**	**TN**	**P**	**N**	**Accuracy**	**Recall**	**Precision**
Chaid Normal	No resampling	N/A	N/A	174	22,790	N/A	N/A	N/A	N/A	N/A
Chaid Normal	Majority reduction	164	2,879	10	19,911	174	22,790	0.874	0.943	0.054
Chaid Normal	Minority boosting	152	2,772	22	20,018	174	22,790	0.878	0.874	0.052
Chaid Reduced	No resampling	28	12	38	531	66	543	0.918	0.424	0.700
Chaid Reduced	Majority reduction	61	75	5	468	66	543	0.869	0.924	0.449
Chaid Reduced	Minority boosting	61	88	5	455	66	543	0.847	0.924	0.409
Chaid Semi Reduced	No resampling	N/A	N/A	150	6,825	N/A	N/A	N/A	N/A	N/A
Chaid Semi Reduced	Majority reduction	126	870	24	5,955	150	6,825	0.872	0.840	0.127
Chaid Semi Reduced	Minority boosting	131	1,135	19	5,690	150	6,825	0.835	0.873	0.103
Bayes net Normal	No resampling	24	-	150	22,790	174	22,790	0.993	0.138	1.000
Bayes net Normal	Majority reduction	167	1,207	7	21,583	174	22,790	0.947	0.960	0.122
Bayes net Normal	Minority boosting	167	1,095	7	21,695	174	22,790	0.952	0.960	0.132
Bayes net Reduced	No resampling	29	-	37	543	66	543	0.939	0.439	1.000
Bayes net Reduced	Majority reduction	64	70	2	473	66	543	0.882	0.970	0.478
Bayes net Reduced	Minority boosting	64	69	2	474	66	543	0.883	0.970	0.481
Bayes net Semi Reduce	No resampling	31	-	119	6,825	150	6,825	0.983	0.207	1.000
Bayes net Semi Reduce	Majority reduction	143	664	7	6,161	150	6,825	0.904	0.953	0.177
Bayes net Semi Reduce	Minority boosting	144	405	6	6,420	150	6,825	0.941	0.960	0.262
C5.0 Normal	No resampling	N/A	N/A	174	22,790	N/A	N/A	N/A	N/A	N/A
C5.0 Normal	Majority reduction	159	3,004	15	19,786	174	22,790	0.869	0.914	0.050
C5.0 Normal	Minority boosting	122	375	52	22,415	174	22,790	0.981	0.701	0.245
C5.0 Reduced	No resampling	N/A	N/A	66	543	N/A	N/A	N/A	N/A	N/A
C5.0 Reduced	Majority reduction	56	126	10	417	66	543	0.777	0.848	0.308
C5.0 Reduced	Minority boosting	53	46	13	497	66	543	0.903	0.803	0.535
C5.0 Semi Reduced	No resampling	N/A	N/A	150	6,825	N/A	N/A	N/A	N/A	N/A
C5.0 Semi Reduced	Majority reduction	123	1,208	27	5,617	150	6,825	0.823	0.820	0.092
C5.0 Semi Reduced	Minority boosting	117	185	33	6,640	150	6,825	0.969	0.780	0.387

As described in chapter 3, fatality events were grouped into six counts as follows:

1	True positives (TP):	4	False negatives (FN):
2	False positives (FP):	5	Total positves (P):
3	True negatives (TN):	6	Total negatives (N):

Twenty two models succeeded in classifying accidents for all the six counts TP, FP, FN, and TN for both training and test data. However, five models, as summarised in Table 5.3, were unable to extract prediction for fatals accidents and TP and FP are shown as N/A. All these cases occurred with no resampling, resulting in over half of the no resampling models failing to classify fatal accidents. This suggests underfitting where the models failed to capture the underlying trend in the data and failed to extract prediction. These five models were therefore excluded from any further analysis in this experiment.

Table 5. 3 Models with unsuccessful prediction



			Training data				Test data			
	Modelling Technique	Sampling Technique	TP	FP	FN	TN	TP	FP	FN	TN
M1	Chaid Normal	No resampling	N/A	N/A	702	91,534	N/A	N/A	174	22,790
M7	Chaid Semi Reduced	No resampling	N/A	N/A	563	27,022	N/A	N/A	150	6,825
M19	C5.0 Normal	No resampling	N/A	N/A	702	91,534	N/A	N/A	174	22,790
M22	C5.0 Reduced	No resampling	N/A	N/A	199	2,072	N/A	N/A	66	543
M25	C5.0 Semi Reduced	No resampling	N/A	N/A	563	27,022	N/A	N/A	150	6,825

For the six Chaid and C5.0 no resampling models, only Chaid reduced produced a prediction. All Bayes net models produced a prediction result for all three sampling techniques. For training data, TP and TN counts far exceeded FP and FN counts indicating that in general the models classified fatals and non-fatals correctly. For the test data, although TP was generally classified accurately, the model identified a high count of FP indicating that non-fatal accidents in the test data held similar characteristics to fatal accidents. Positively FN counts were low indicating that the model does not often misclassify fatal as non-fatals.

Accuracy

Accuracy is the proportion of accidents classified correctly. For the 27 models, referenced M1 to M27, accuracy was calculated based on the formula:

$$\text{Accuracy} = (\text{Count of TP} + \text{Count of TN}) / (\text{Count of P} + \text{Count of N})$$

Accuracy performance reported on the training data, as outlined in Table 5.1, ranged from 0.868 to 0.995 indicating that in general the models are classifying accidents to a high level of accuracy. For test data accuracy ranged from 0.823 to 0.993, however there was one outlier at 0.777. In similar recent research by (Wah, et al., 2012) classifying motor cycle accident occurrences using CART decision tree, training accuracy was reported as 0.8337 and test accuracy was 0.7812. Accuracy ranges for this experiment result were in line with (Wah, et al., 2012) and on this basis overall model accuracy for training and test data was considered acceptable and no further remodelling was completed. Interestingly, accuracy for test data was lower than training data in (Wah, et al., 2012) experiment as well as in this experiment which indicates slight overfitting.

Recall

Recall, sometimes referred to as the true positive rate, is the proportion of fatals classified as fatals and is calculated using the following formula:

$$\text{Recall} = (\text{Count of TP})/((\text{Count of FN})+(\text{Count of TP}))$$

No resampling recall results were poor for this experiment as the base data favours the majority class or non-fatal and recall focuses on fatal accident results. Where the majority reduction and minority boosting sampling techniques were applied, recall rates significantly improved with rates for training data ranging from 0.890 to 1.000 and for test data ranging from 0.701 to 0.970. This indicates that rebalancing in favour of the fatal class was effective.

Precision

Precision is a measure of exactness being the percentage of fatals classified as fatals which are actually fatals or true positives. Precision is calculated using the following formula:

$$\text{Precision} = (\text{Count of TP})/((\text{Count of TP})+(\text{Count of FP}))$$

For training data precision ranged from 0.838 to 1.000 where the model has extracted patterns from the data. However, for test data precision rates dropped significantly ranging from 0.052 to 0.700, with the exception of the three Bayes net no resampling models which achieved 1.000. Where precision was low, the test data produced good recall results which is in line with (Han, et al., 2011, p. 368) who identified the inverse relationship between precision and recall.

5.1.2 Model assessment

Chaid model review

Two Chaid models were eliminated from the model assessment as they did not produce fatal accident prediction. Five of the seven remaining Chaid models scored 0.88 to 0.89 accuracy for training data, with the highest Chaid models achieving 0.903 and 0.946. Recall for Chaid models generally produced good results for training data with the lowest of 0.890 and the highest of 0.952 with the exception of “M4 reduced no resampling”. M4 produced recall of only 0.553 although it had the highest accuracy for Chaid models of 0.946. Chaid models accuracy for test data ranged from 0.835 to 0.918 which is marginally less than the training data result. As with training data, recall achieved good results with test data recall ranging from 0.840 to 0.943 with the exception of M4 whose recall was 0.424 in keeping with the training data results.

Bayes net model review

Bayes net models performed well for accuracy with training data results ranging from lowest of 0.933 to highest of 0.994. All Bayes net models succeeded in fatal accident prediction. No resampling models produced the highest level of accuracy for Bayes net models, however, recall was poor at a low of 0.195 and a high of 0.477. Positively, both majority reduction and minority boosting sampling performed well for training data for both accuracy and recall with ranges of 0.933 to 0.975 and 0.970 and 1.000 respectively. For test data, Bayes net models performed well for accuracy although accuracy for reduced models was slightly lower for test data. Accuracy results for test data ranged from 0.882 to 0.993. Recall results generally ranged from 0.953 to 0.970, however, no resampling also performed poorly as in the training data ranging from 0.138 to 0.439.

C5.0 model review

Three Chaid models were eliminated from the model assessment as they did not produce fatal accident prediction. The remaining six C5.0 models scored 0.868 to 0.995 accuracy, with "M21 normal minority boosting" achieving the highest of 0.995.

Minority boosting sampling produced 1.000 accuracy for all three models for training data and these three models also achieved the highest accuracy from 0.976 to 0.995. The remaining models also performed well for recall ranging from 0.909 to 0.937. For test data, accuracy scores were slightly less, ranging from 0.777 to 0.981, although still good. As with training data, "M21 normal minority boosting" achieved the highest result. The modelling techniques and sampling techniques achieved the exact same order of accuracy in training and test data as outlined in Table 5.4.

Table 5. 4 C5.0 models accuracy results

Ranking	Modelling Technique	Sampling Technique	Taining Data Accuracy	Test Data Accuracy
1	C5.0 Normal	Minority boosting	0.995	0.981
2	C5.0 Semi Reduced	Minority boosting	0.988	0.969
3	C5.0 Reduced	Minority boosting	0.976	0.903
4	C5.0 Normal	Majority reduction	0.900	0.869
5	C5.0 Semi Reduced	Majority reduction	0.877	0.823
6	C5.0 Reduced	Majority reduction	0.868	0.777

For test data, recall for “minority boosting” models did not perform as well as in the training data and in contrast the three models were lowest ranging from 0.701 to 0.914. This suggests an overfitting situation where the classifier perfectly fits the training data and therefore the model can lose capability to generalise to situations not presented in the training data (Wah, et al., 2012, p. section B). As expected model results for the test data generally under-performed the training data as the modelling techniques had not previously seen the test data.

5.1.3 ROC assessment

The ROC curve shows the trade-off between the proportion of fatal accidents correctly classified as fatal and the proportion of non-fatal accidents incorrectly classified as fatal. This is commonly described as the true positive rate or sensitivity against the false positive rate or 1-specificity. The closer the curve follows the Y axis and then tails off to the right, the more accurately the model classifies fatal accidents and the less likely to incorrectly classify a non-fatal accident as fatal (Han, et al., 2011, p. 374). Similarly, the larger the space between the 45 degree line and curve, the more accurate a model is. This space is referred to as the area under the curve (AUC). A high AUC indicates good recall. In general the ROC results for the 22 models performed well for ROC assessments and the ROC curves for training and test data were not significantly different. Fig. 5.1 is an example of the ROC results for both training and test data presented for C5.0 models for normal data and majority reduction sampling techniques. Both charts demonstrate that the true positive rate is high and the false positive rate is low, therefore indicating the model classifies fatalities well and unlikely to misclassify non-fatals as fatals.

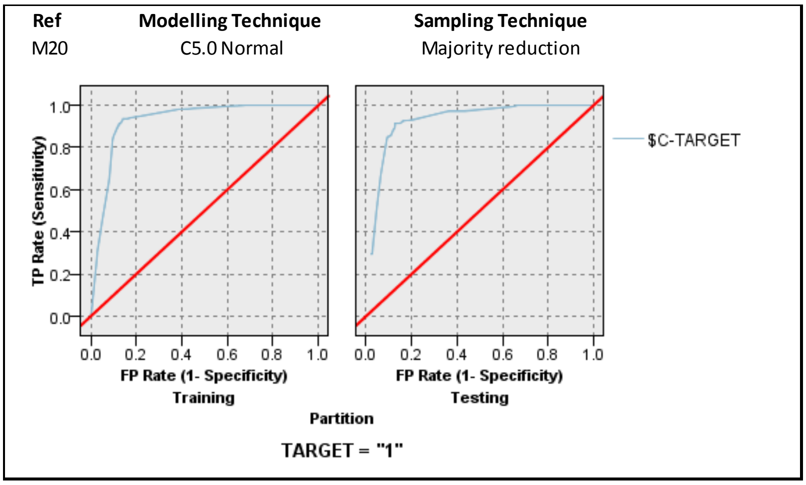


Figure 5. 1 Accurate ROC curve

Although some ROC curves did not perform as well for test data, as in Fig. 5.2 below, the result still indicates good performance although the risk of false positives increased.

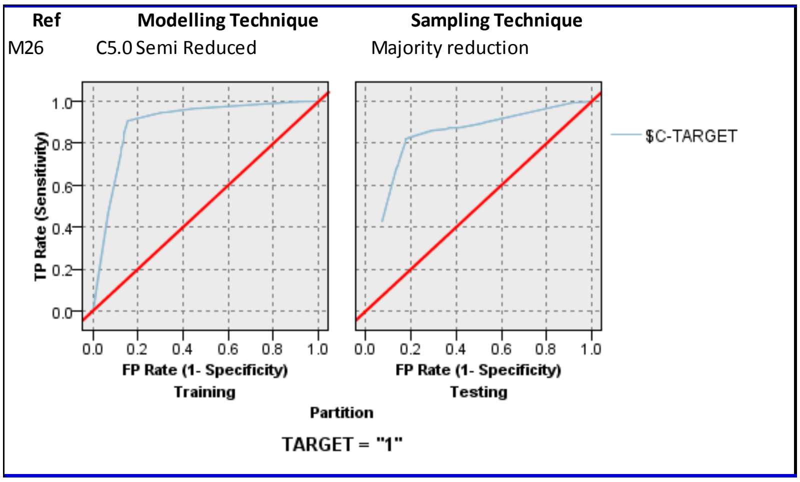
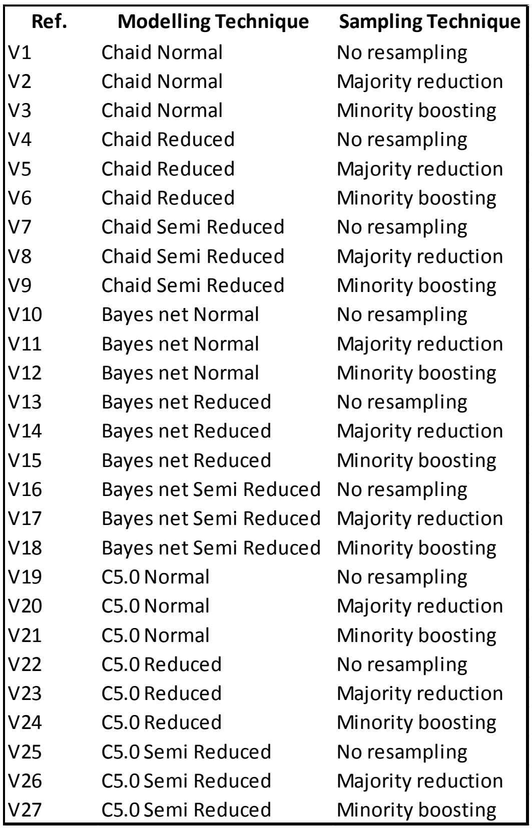


Figure 5. 2 Less accurate ROC curve

5.2 Model Evaluation with Validation Data

Once the model assessment was complete, further evaluation was performed to test the usefulness of the models towards achieving the research experiment goal of identifying the factors most likely to accurately predict fatal traffic accidents. The prediction accuracy was further examined by testing the models using a 2013 STATS19 accident dataset or validation data. The 27 models described in chapter 4, which were the basis of training and test data assessment, were copied to create new models in order to test the 2013 validation data.

Table 5. 5 Validation model listing



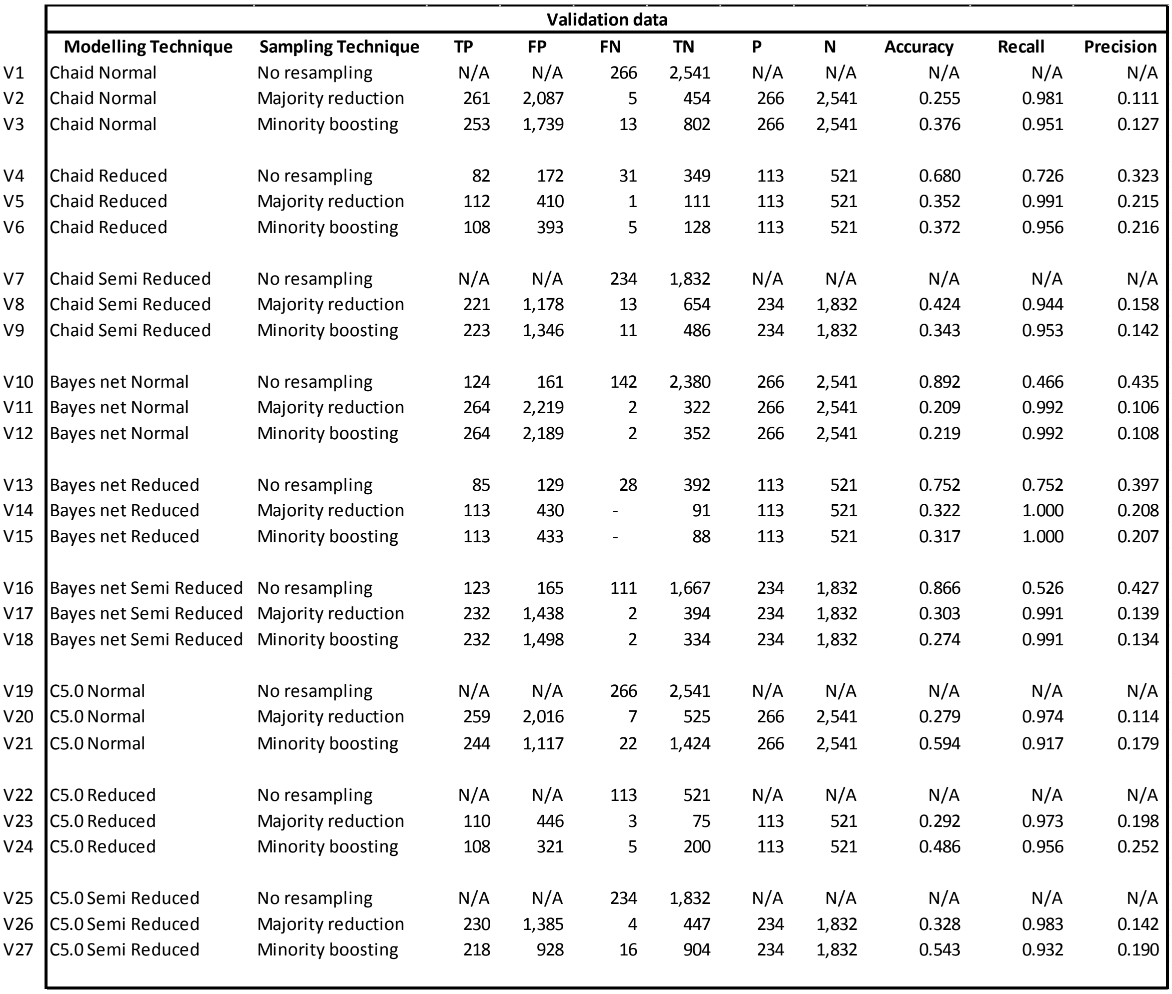
Ref.	Modelling Technique	Sampling Technique
V1	Chaid Normal	No resampling
V2	Chaid Normal	Majority reduction
V3	Chaid Normal	Minority boosting
V4	Chaid Reduced	No resampling
V5	Chaid Reduced	Majority reduction
V6	Chaid Reduced	Minority boosting
V7	Chaid Semi Reduced	No resampling
V8	Chaid Semi Reduced	Majority reduction
V9	Chaid Semi Reduced	Minority boosting
V10	Bayes net Normal	No resampling
V11	Bayes net Normal	Majority reduction
V12	Bayes net Normal	Minority boosting
V13	Bayes net Reduced	No resampling
V14	Bayes net Reduced	Majority reduction
V15	Bayes net Reduced	Minority boosting
V16	Bayes net Semi Reduced	No resampling
V17	Bayes net Semi Reduced	Majority reduction
V18	Bayes net Semi Reduced	Minority boosting
V19	C5.0 Normal	No resampling
V20	C5.0 Normal	Majority reduction
V21	C5.0 Normal	Minority boosting
V22	C5.0 Reduced	No resampling
V23	C5.0 Reduced	Majority reduction
V24	C5.0 Reduced	Minority boosting
V25	C5.0 Semi Reduced	No resampling
V26	C5.0 Semi Reduced	Majority reduction
V27	C5.0 Semi Reduced	Minority boosting

The same process for data construction was followed as outlined in chapter 4. The new models created are listed in Table 5.5. A confusion matrix was produced for the validation data for each modelling technique and for the three datasets, normal, reduced and semi reduced. A technical and non-technical evaluation was completed for the validation data. The technical evaluation was based on accuracy and recall as calculated based on the confusion matrix summary. The non-technical evaluation was based on a review of the interpretability of the models.

5.2.1 Confusion matrix evaluation

The confusion matrix for the validation data and the accuracy, recall and precision performance measures calculated are outlined in Table 5.6. As in training and test assessment, five models were unable to extract prediction for fatals accidents.

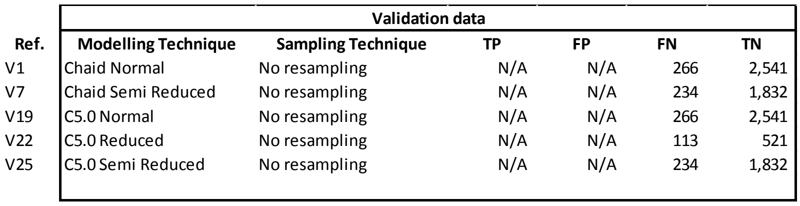
Table 5. 6 Confusion matrix and performance measures for validation data



	Validation data										
	Modelling Technique	**Sampling Technique**	**TP**	**FP**	**FN**	**TN**	**P**	**N**	**Accuracy**	**Recall**	**Precision**
V1	Chaid Normal	No resampling	N/A	N/A	266	2,541	N/A	N/A	N/A	N/A	N/A
V2	Chaid Normal	Majority reduction	261	2,087	5	454	266	2,541	0.255	0.981	0.111
V3	Chaid Normal	Minority boosting	253	1,739	13	802	266	2,541	0.376	0.951	0.127
V4	Chaid Reduced	No resampling	82	172	31	349	113	521	0.680	0.726	0.323
V5	Chaid Reduced	Majority reduction	112	410	1	111	113	521	0.352	0.991	0.215
V6	Chaid Reduced	Minority boosting	108	393	5	128	113	521	0.372	0.956	0.216
V7	Chaid Semi Reduced	No resampling	N/A	N/A	234	1,832	N/A	N/A	N/A	N/A	N/A
V8	Chaid Semi Reduced	Majority reduction	221	1,178	13	654	234	1,832	0.424	0.944	0.158
V9	Chaid Semi Reduced	Minority boosting	223	1,346	11	486	234	1,832	0.343	0.953	0.142
V10	Bayes net Normal	No resampling	124	161	142	2,380	266	2,541	0.892	0.466	0.435
V11	Bayes net Normal	Majority reduction	264	2,219	2	322	266	2,541	0.209	0.992	0.106
V12	Bayes net Normal	Minority boosting	264	2,189	2	352	266	2,541	0.219	0.992	0.108
V13	Bayes net Reduced	No resampling	85	129	28	392	113	521	0.752	0.752	0.397
V14	Bayes net Reduced	Majority reduction	113	430	-	91	113	521	0.322	1.000	0.208
V15	Bayes net Reduced	Minority boosting	113	433	-	88	113	521	0.317	1.000	0.207
V16	Bayes net Semi Reduced	No resampling	123	165	111	1,667	234	1,832	0.866	0.526	0.427
V17	Bayes net Semi Reduced	Majority reduction	232	1,438	2	394	234	1,832	0.303	0.991	0.139
V18	Bayes net Semi Reduced	Minority boosting	232	1,498	2	334	234	1,832	0.274	0.991	0.134
V19	C5.0 Normal	No resampling	N/A	N/A	266	2,541	N/A	N/A	N/A	N/A	N/A
V20	C5.0 Normal	Majority reduction	259	2,016	7	525	266	2,541	0.279	0.974	0.114
V21	C5.0 Normal	Minority boosting	244	1,117	22	1,424	266	2,541	0.594	0.917	0.179
V22	C5.0 Reduced	No resampling	N/A	N/A	113	521	N/A	N/A	N/A	N/A	N/A
V23	C5.0 Reduced	Majority reduction	110	446	3	75	113	521	0.292	0.973	0.198
V24	C5.0 Reduced	Minority boosting	108	321	5	200	113	521	0.486	0.956	0.252
V25	C5.0 Semi Reduced	No resampling	N/A	N/A	234	1,832	N/A	N/A	N/A	N/A	N/A
V26	C5.0 Semi Reduced	Majority reduction	230	1,385	4	447	234	1,832	0.328	0.983	0.142
V27	C5.0 Semi Reduced	Minority boosting	218	928	16	904	234	1,832	0.543	0.932	0.190

The five models, which failed to extract prediction for fatal accidents, had no resampling technique applied to the data and the five models, listed in Table 5.7, were excluded from any further analysis.

Table 5. 7 Validation models with unsuccessful prediction



	Validation data					
Ref.	**Modelling Technique**	**Sampling Technique**	**TP**	**FP**	**FN**	**TN**
V1	Chaid Normal	No resampling	N/A	N/A	266	2,541
V7	Chaid Semi Reduced	No resampling	N/A	N/A	234	1,832
V19	C5.0 Normal	No resampling	N/A	N/A	266	2,541
V22	C5.0 Reduced	No resampling	N/A	N/A	113	521
V25	C5.0 Semi Reduced	No resampling	N/A	N/A	234	1,832

The performance metrics for accuracy and recall for each of the remaining 22 models is graphically presented in Fig. 5.3. It is clear from the graph that there is a significant variance between the results achieved for accuracy and recall for the models when applied to the validation data.

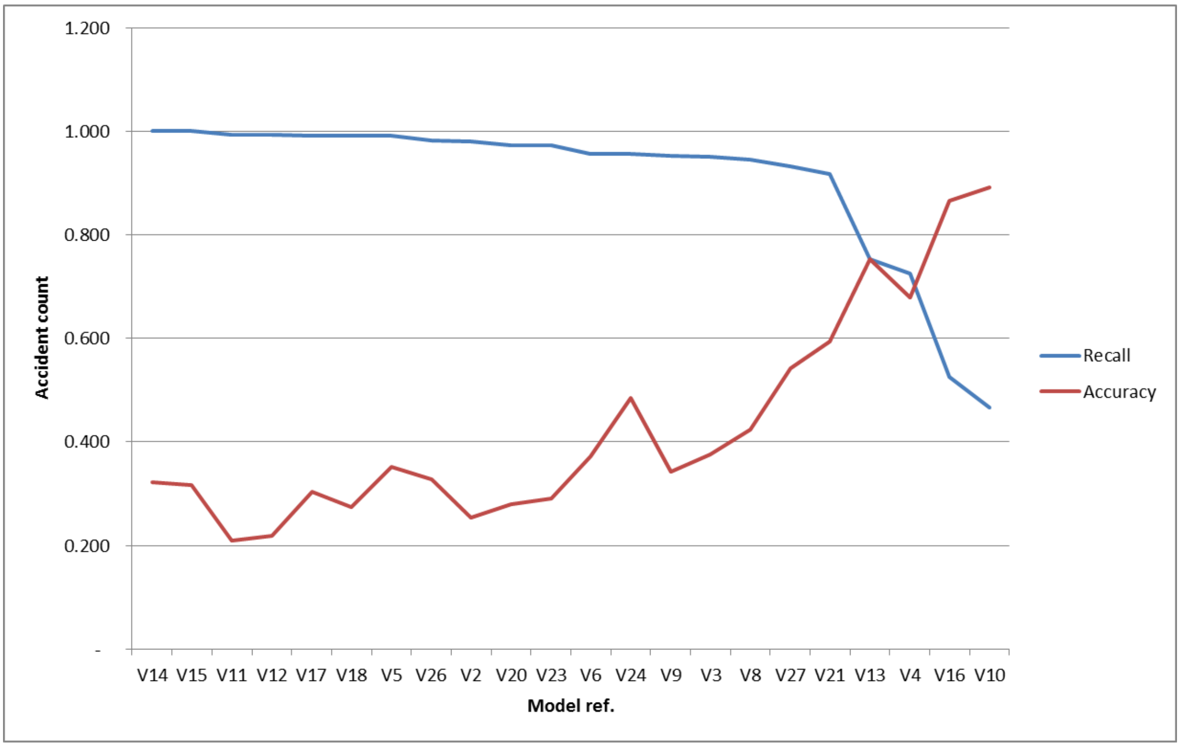


Figure 5. 3 Validation recall and accuracy results

Accuracy

Accuracy relates to ability of the model to classify fatal and non-fatal accidents correctly. As discussed in chapter 3, accuracy rates for infrequent events such as fatal accidents may be low where sampling techniques are applied given the models are built to focus on fatal accident prediction. Accuracy rates for the validation model have a vast range from 0.209 to 0.892. The top performing models are all Bayes net models and all have no resampling. “V10 normal”, “V16 semi reduced” and “V13 reduced” performed most accurately with 0.892, 0.866 and 0.752 respectively. Sixteen models achieved less than 50% accuracy with accuracy ranging from 0.209 to 0.486. These results mean that the models had some difficulty correctly classifying accidents as fatal or non-fatal with unseen data. The model classifies fatals quite well and the number of false negatives is relatively low. Where the model struggles is in accurately classifying the non-fatals, with a large number of false positives being identified. The accuracy on training and test data was in general good so the poor performance on validation data suggests additional factors in the unseen data which the model could not recognise. Fig. 5.4 presents the count of accidents classified by each model for TP, FP and FN from the validation data.

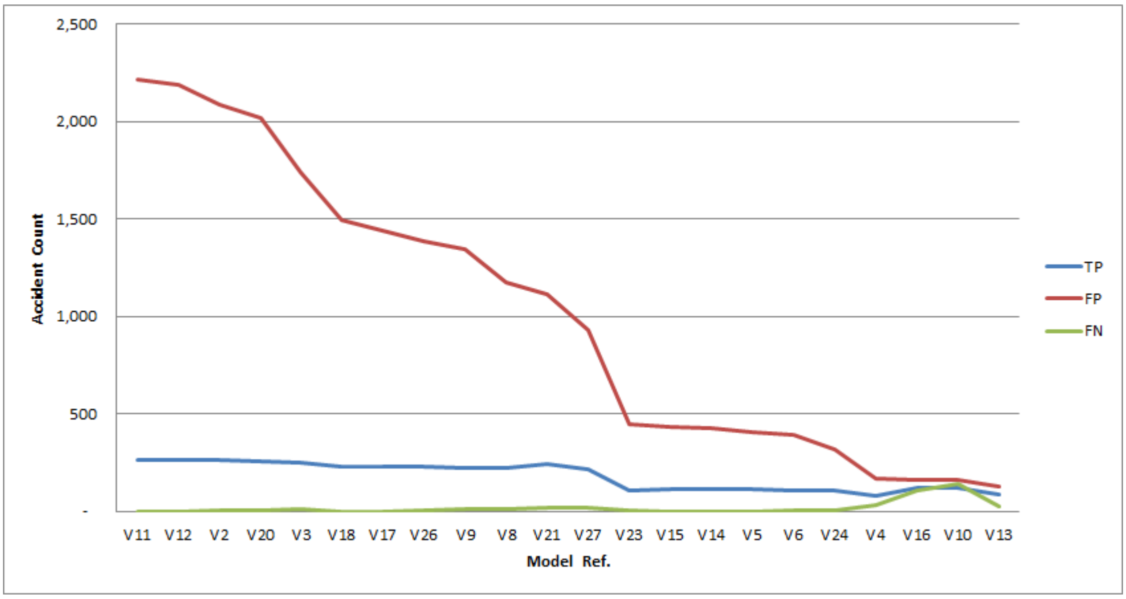


Figure 5. 4 Validation count TP, FP & FN

The data construction was focussed on fatal accident classification and the models worked well in achieving that task as can be seen from the low number of false negatives. However, data points for fatal and non-fatal accidents were quite similar and, based on the low accuracy performance, further review and analysis of the data points in the STATS19 may help to identify additional data groupings or characteristics which could reduce the misclassification of non-fatals as fatals. Further understanding of the data points would require consultation with subject matter experts to add deeper data understanding.

Recall

Eighteen of the twenty two models had recall over 0.9 indicating that most models had a good fatal accident classification rate. The models which did not perform as well did not have sampling techniques applied, indicating that data rebalance was important for classifying infrequent events like fatal accidents. The models which produced the best recall on validation data were “V14 Bayes reduced majority reduction”, “V12 Bayes normal minority boosting” and “V18 Bayes semi reduced minority boosting” with 1.000, 0.992 and 0.9991 respectively. For each of these models accuracy was noted as poor but recall was good with the cause of poor accuracy being high false positive counts.

Overall the models performed well for recall for validation data meaning most fatal accidents were correctly classified as fatal by each model. Given recall of fatal accidents was a key focus of the data preparation phase, the high level of recall is a positive result. Fig. 5.5 outlines the overall recall performance across all models for validation data.

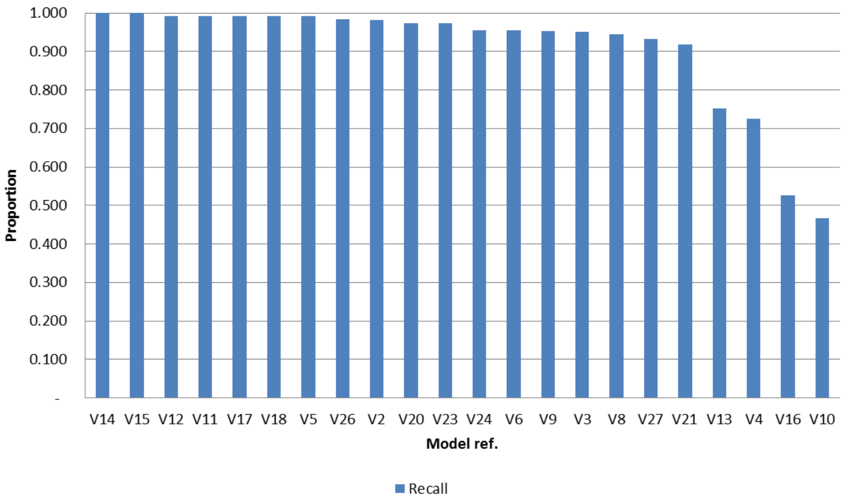


Figure 5. 5 Recall performance by model

Precision

For the validation data, evidence of the inverse relationship between precision and recall was displayed. Precision rates were consistently low and ranged from 0.106 to 0.435. For most models recall was high so the low precision performance was expected. (Weiss & Hirsh, 2000) outlined that for infrequent events relatively low precision rates may be considered acceptable as long as many of the target events are predicted. For this experiment, the precision rates were considered acceptable due to the high recall rates achieved.

ROC Curve

The ROC curve shows the trade-off between the proportion of fatal accidents correctly classified as fatal and the proportion of non-fatal accidents incorrectly classified as fatal. Unsurprisingly the ROC curves, were quite different to the curves produced for training and test data. Fig. 5.6 shows the ROC curve for the best performing model "V14 Bayes reduced majority reduction". As the false positives were more significant in the validation data, the curve indicated less accurate prediction as there was less area under the curve (AUC).

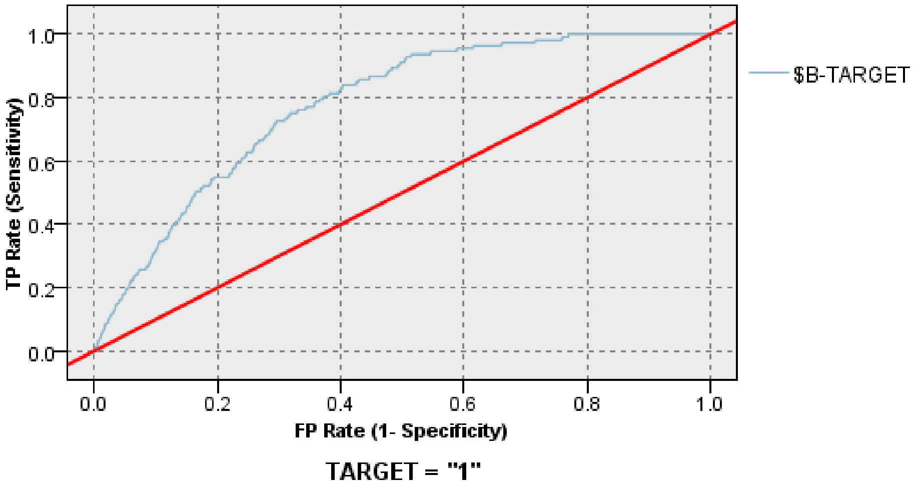


Figure 5. 6 ROC curve model V14

Top Performing Models

Table 5.8 summarises the top performing models from this experiment based on their combined performance in relation to accuracy and recall. Although the experiment was focussed on fatal accident prediction, the models ability to accurately identify fatals and non-fatals correctly was also important to minimise the false positives and negatives. The top performing models are discussed and analysed in more detail in the following section.

Table 5. 8 Confusion matrix and performance measures for top performing models

	Modelling Technique	Sampling Technique	TP	FP	FN	TN	P	N	Accuracy	Recall
V13	Bayes net Reduced	No resampling	85	129	28	392	113	521	0.752	0.752
V21	C5.0 Normal	Minority boosting	244	1,117	22	1,424	266	2,541	0.594	0.917
V4	Chaid Reduced	No resampling	82	172	31	349	113	521	0.680	0.726

5.2.2 Model 1: V13 Bayes net reduced no resampling

Technical evaluation

"V13 Bayes net reduced no resampling" was selected as a top performing model as it was balanced between accuracy and recall. Accuracy of 0.752 means the model was good at classifying fatal and non-fatal accidents. Recall, being the proportion of fatal accidents which were classified as fatal, was 0.752 and means the model was good at classifying fatal accidents as fatal. This model misclassified 28 fatals and 129 non-fatals. This represented 25% of fatals and 25% of non-fatals meaning the model classified the majority of fatals and non-fatals correctly. A higher recall would have resulted in improved fatal accident prediction, however, based on the results of other models, this may have led to a higher proportion of false positives. ROC curves for training and test are presented in Fig. 5.7. The validation ROC in Fig. 5.8 suggests the model overfitted the data as the TP and FP rates were not as good as training and test. The validation result was good with a large area under the curve indicating fatals are classified well but false positives risk was apparent.

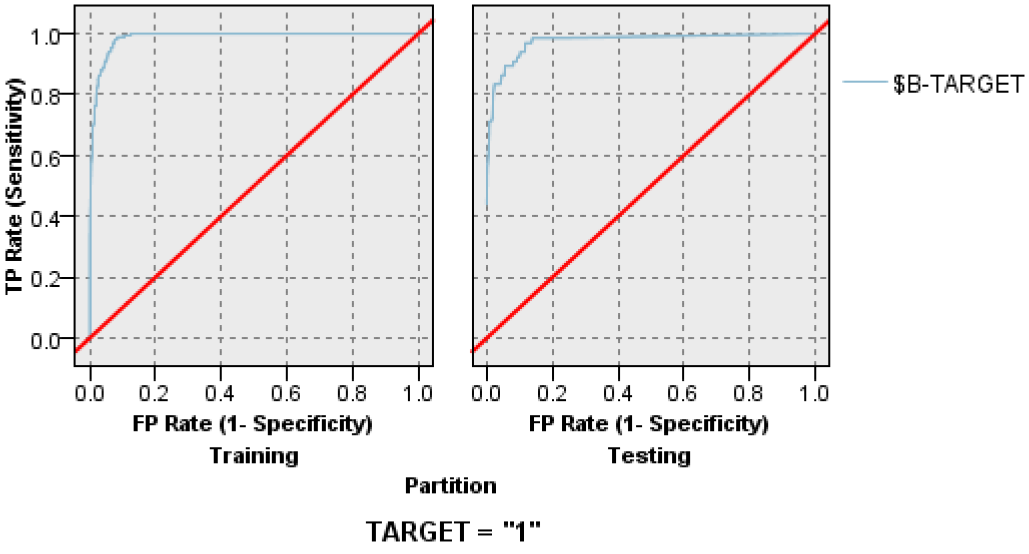


Figure 5. 7 Training & test ROC for Bayes net reduced no resampling

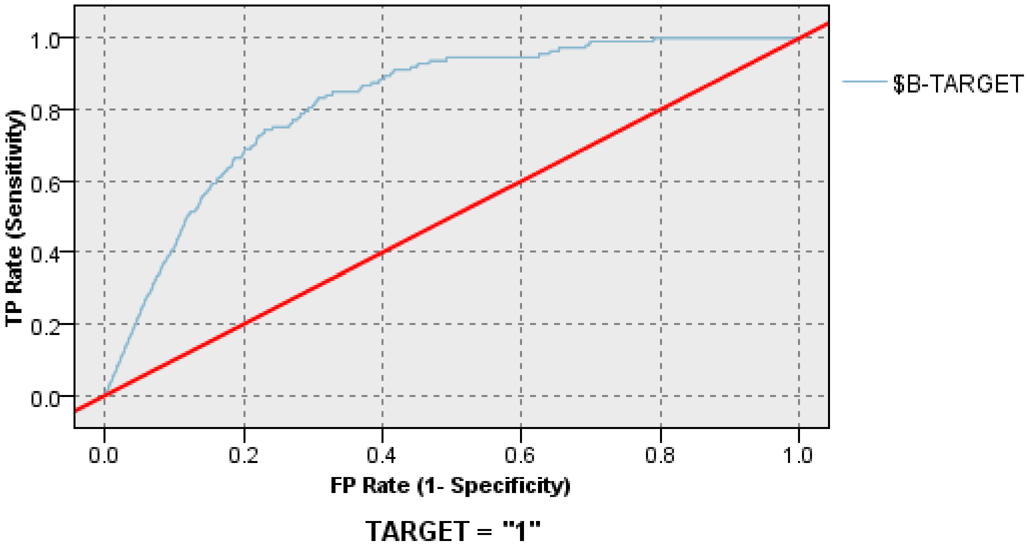


Figure 5. 8 Validation ROC for Bayes net reduced no resampling

Non-Technical evaluation

Fig. 5.9 presents importance of individual predictors for “V13 Bayes net reduced no resampling”. Road type and road surface account for 33% and 27% importance respectively. Urban/rural account for 16% of importance and junction detail 12% with weather conditions at 9% and light at 2%. For this model, road type and road surface were the most important individual predictors and were the most likely factors for fatal accident occurrence.

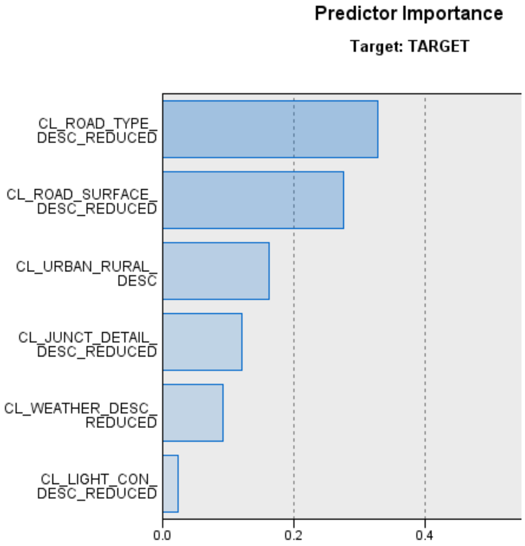


Figure 5. 9 Bayes net reduced no resampling predictor importance

Fig. 5.10 presents the Bayesian network for the model and highlights the strongest relationships as deeper colour nodes on the graphical model. Road surface and road type were the darkest coloured predictors and therefore the most important predictors of fatal accident for this model.

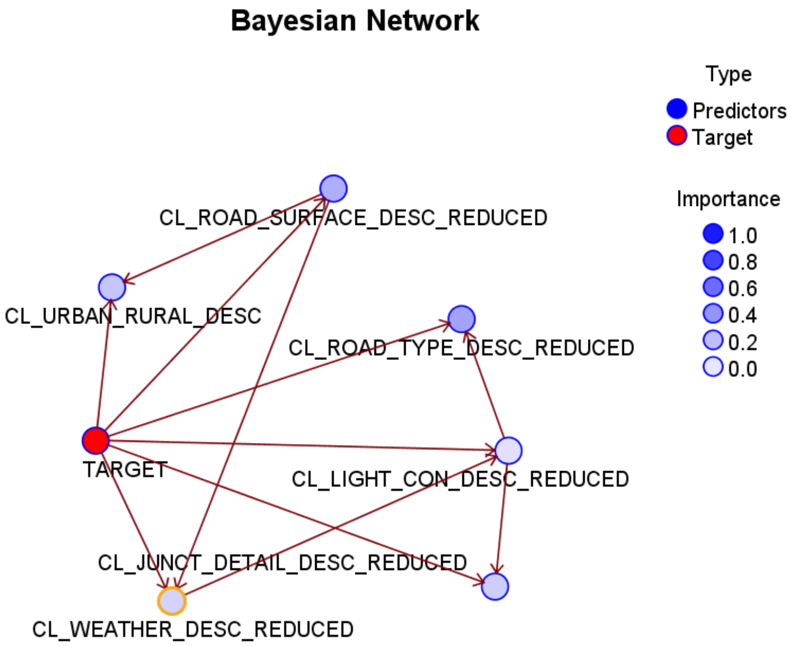
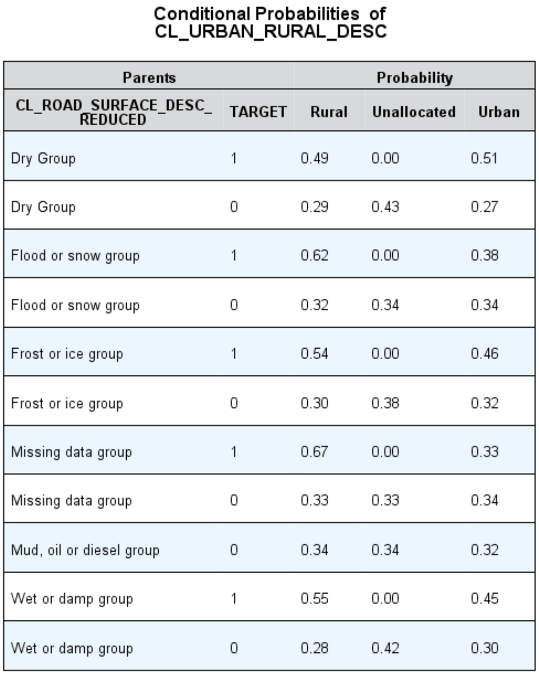


Figure 5. 10 Bayes net reduced with no resampling

The Bayes net model identified three relationships as identified by the directional arrows. Road type and road surface were identified as the key individual indicators of fatal accident therefore the two related relationships will be further discussed.

The first relationship linked road surface to urban/rural as outlined in Table 5.9. A target of 1 represents a fatal accident.

Table 5. 9 V13 road surface and urban/rural conditional probability



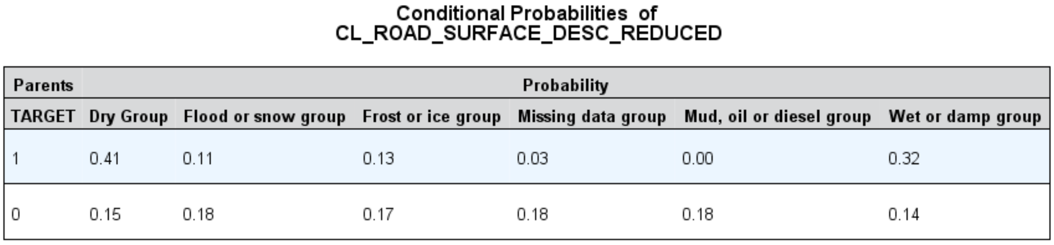
Conditional Probabilities of CL_URBAN_RURAL_DESC

Parents		Probability		
CL_ROAD_SURFACE_DESC_REDUCED	TARGET	Rural	Unallocated	Urban
Dry Group	1	0.49	0.00	0.51
Dry Group	0	0.29	0.43	0.27
Flood or snow group	1	0.62	0.00	0.38
Flood or snow group	0	0.32	0.34	0.34
Frost or ice group	1	0.54	0.00	0.46
Frost or ice group	0	0.30	0.38	0.32
Missing data group	1	0.67	0.00	0.33
Missing data group	0	0.33	0.33	0.34
Mud, oil or diesel group	0	0.34	0.34	0.32
Wet or damp group	1	0.55	0.00	0.45
Wet or damp group	0	0.28	0.42	0.30

Table 5.10 presents the conditional probability of road surface causing an accident. “Dry group” at 41% had the highest probability of contributing to the cause of a fatal

accident. “Wet or damp group” at 32% was the second most significant contributing factor.

Table 5. 10 V13 road surface conditional probability



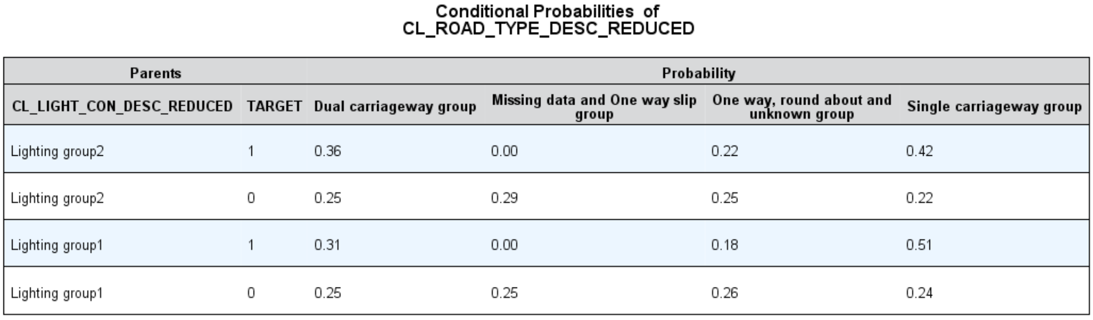
Conditional Probabilities of
CL_ROAD_SURFACE_DESC_REDUCED

Parents	Probability					
TARGET	Dry Group	Flood or snow group	Frost or ice group	Missing data group	Mud, oil or diesel group	Wet or damp group
1	0.41	0.11	0.13	0.03	0.00	0.32
0	0.15	0.18	0.17	0.18	0.18	0.14

The key relationship predictors from the model can be deduced by combining the results in Tables 5.9 and 5.10. “Dry group” road surfaces had the highest probability of occurrences and 51% were most likely in urban with 49% in rural areas. For “wet or damp group”, the reverse relationship existed with rural more probable at 55% and urban at 45%. It could therefore be deduced that fatal accidents are more probable in urban areas where the road surface is dry and in rural areas when the road surface is wet or damp.

The second relationship linked light conditions to road type. As outlined in Table 5.11, single carriageway groups were identified as the most probable for lighting group 1 and 2 at 42% and 51% respectively. Unfortunately, although a strong relationship was identified extraction of a meaningful insight was difficult as the lighting groups as described in Table 5.12, do not provide any distinct factors. Lighting groups were selected using Chaid decision tree to identify most homogeneous groupings, however, group 2 relates to darkness-lights lit, darkness-no lighting or daylight, which would cover the vast majority of lighting conditions and therefore too generalised to extract insight. In order to establish usable insights, data groupings would need to be revisited and could be improved with the knowledge of a subject matter expert as described in future work and research in chapter 6.

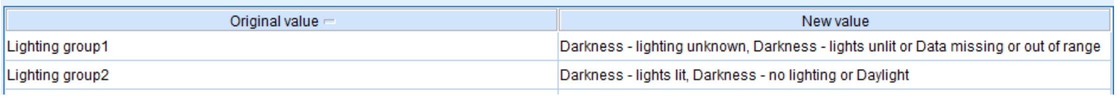
Table 5. 11 V13 light condition and road type conditional probability



Conditional Probabilities of
CL_ROAD_TYPE_DESC_REDUCED

Parents		Probability			
CL_LIGHT_CON_DESC_REDUCED	TARGET	Dual carriageway group	Missing data and One way slip group	One way, round about and unknown group	Single carriageway group
Lighting group2	1	0.36	0.00	0.22	0.42
Lighting group2	0	0.25	0.29	0.25	0.22
Lighting group1	1	0.31	0.00	0.18	0.51
Lighting group1	0	0.25	0.25	0.26	0.24

Table 5. 12 Lighting groups description



Original value	New value
Lighting group1	Darkness - lighting unknown, Darkness - lights unlit or Data missing or out of range
Lighting group2	Darkness - lights lit, Darkness - no lighting or Daylight

5.2.3 Model 2: V21 C5.0 normal minority boosting

Technical evaluation

"V21 C5.0 normal minority boosting" was selected as a top performing model as it had very good recall at 0.917 and better accuracy than most models at 0.594. The high recall meant the model was very good at classifying fatal accidents as fatal, however, it did not perform so well classifying non-fatals correctly. The model identified a significant number of false positives although classification of fatals was much better with a smaller number of false negatives. This model misclassified 1,117 non-fatals or 44% and 22 fatals or 8%. Overall the model performed well at classifying fatal accidents as fatal. However, the large volume of misclassified non-fatals was the main reason for lower accuracy. While rebalancing the data in favour of the rare class led to higher recall, it also meant that the model had difficulty identifying fatal and non-fatal accidents correctly.

ROC curves for training, test and validation are presented below in Fig. 5.11 and 5.12. In Fig. 5.12, the line tails off to the right earlier than in Fig. 5.11 and the area under the curve in Fig. 5.12 is less than in Fig. 5.11 indicating that, proportionally, there were more FPs or non-fatals misclassified as fatals in the validation data than in training or test data. The validation ROC in Fig. 5.12 suggests the model overfitted the data as the TP and FP rates were not as good as training and test. Reducing the FP rate would require revisiting the data construction stage by looking at alternative data groupings. Guidance from a subject matter expert could greatly increase the identification of relevant groupings.

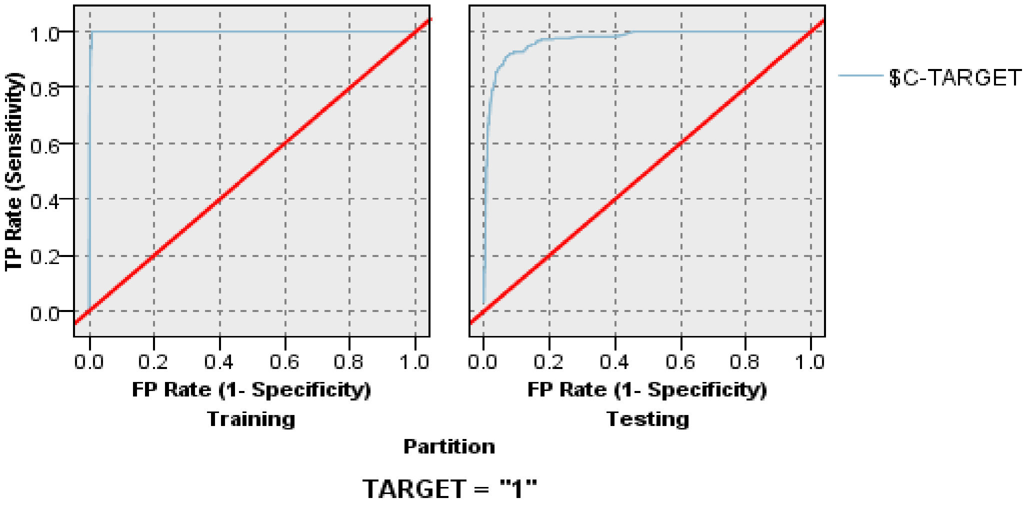


Figure 5. 11 Training & test ROC for C5.0 normal minority boosting

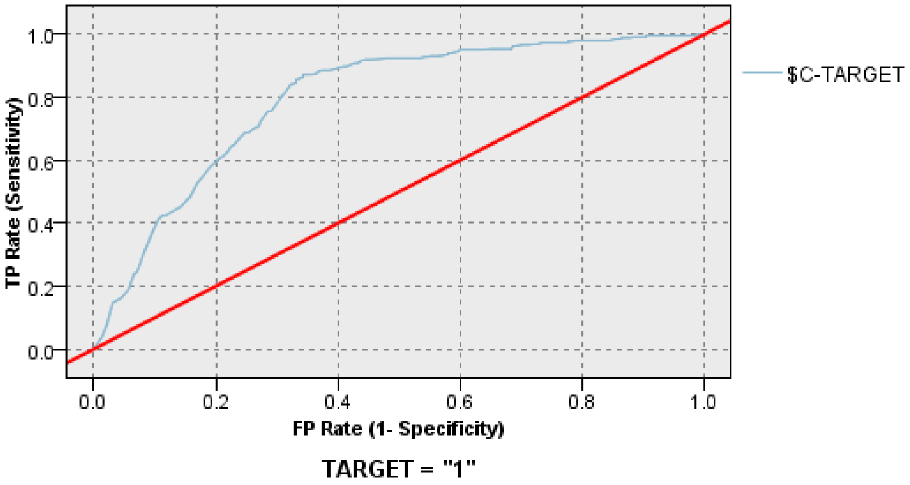


Figure 5. 12 Validation ROC for C5.0 normal minority boosting

Non-Technical evaluation

Fig. 5.13 presents importance of individual predictors for model “V21 C5.0 normal with minority boosting”. This model identified only 2 important predictors being road type at 54% and junction detail at 46%.

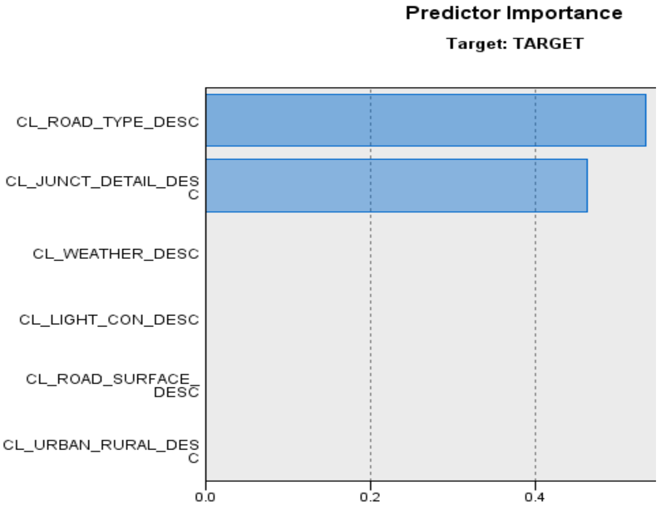


Figure 5. 13 V21 C5.0 normal minority boosting predictor importance

V21 C5.0 model learns rules from the data and presents them in a decision tree format. These rules can then be used to make predictions by scoring new or validation data against the decision tree model. A limitation of SPSS Modeler is, although its decision tree functionality is strong, extraction of the decision tree hierarchical presentation is difficult, especially where large numbers of nodes exist, a similar limitation was experienced by (Wah, et al., 2012). Fig. 5.14 displays the V21 C5.0 model with the most significant node or root node presented on the left i.e. fatal (1) and non-fatal (0). The first predictor was then identified and for this model was urban and rural. Next the most important nodes for both urban and rural were identified, being road surface and road type respectively. For each node the proportion of fatal (1) and non-fatal (0) were presented. The decision tree continues in a similar fashion until the last predictor in the model was identified and was presented as the final node on the branch.

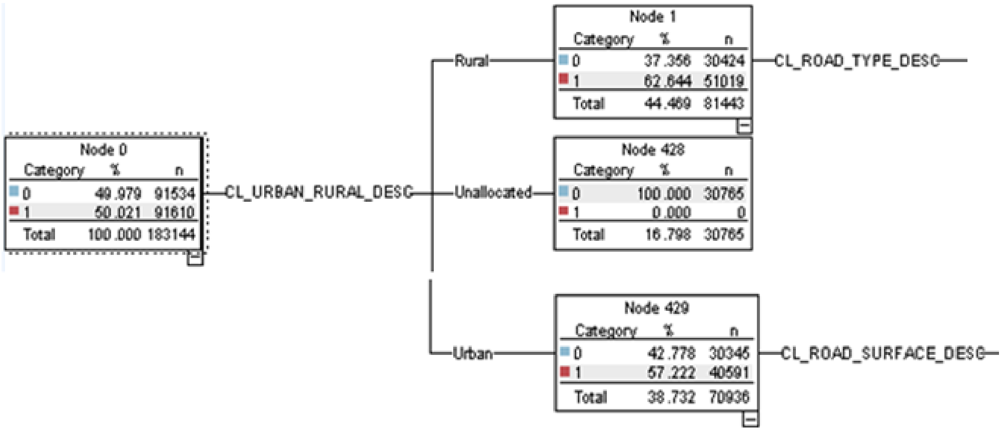


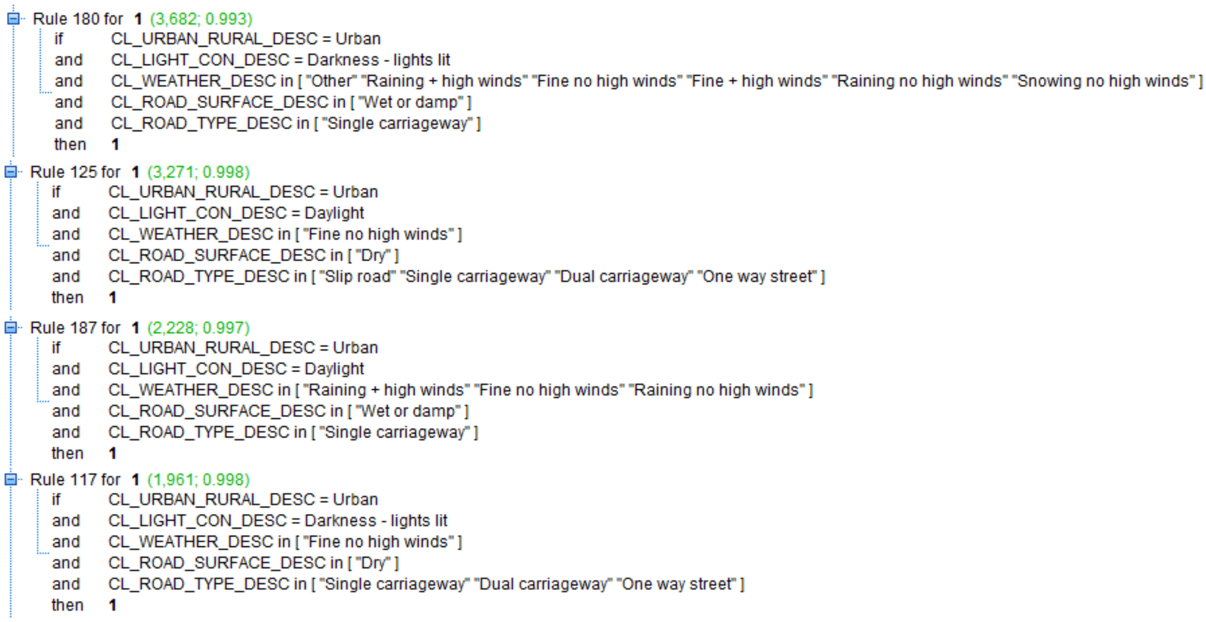
Figure 5. 14 Extract C5.0 normal minority boosting decision tree

The top rules for fatal accidents extracted from the V21 model are listed in Table 5.13. The ranking and frequency count are based on the training data as this is the driver for the model rule definitions and these are used to classify fatal accidents for validation data. The rule indicates if the conditions are met, a fatal accident is likely to occur.

Table 5. 13 V21 top rules based on training data

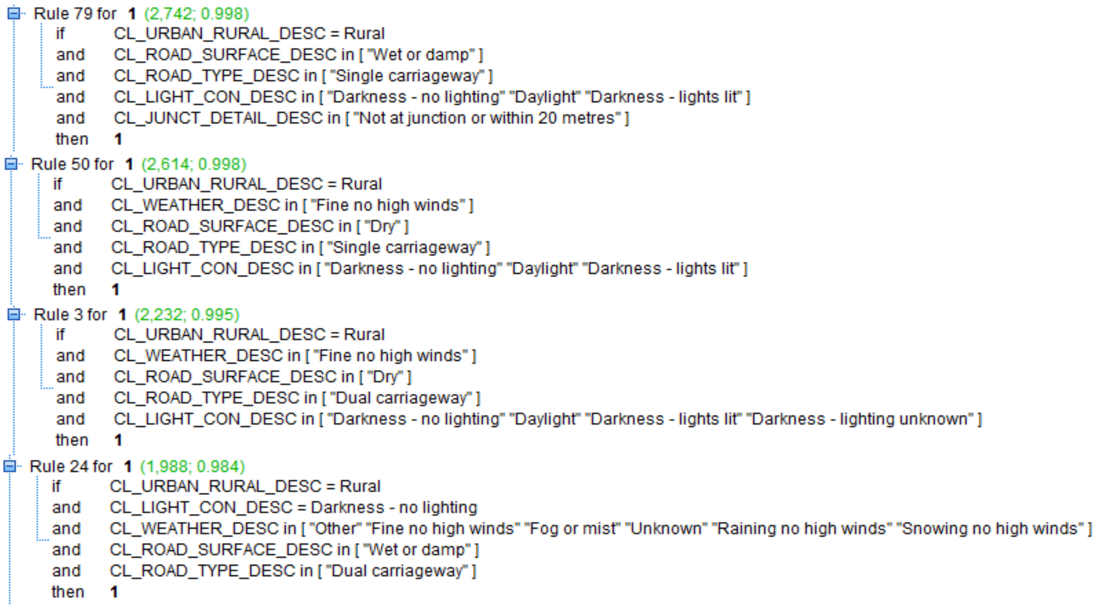
Ranking	Rule Ref.	Training Data Frequency	Confidence	Urban/ Rural
Rank 1	Rule 180 for 1	3,682	0.993	Urban
Rank 2	Rule 125 for 1	3,271	0.998	Urban
Rank 3	Rule 79 for 1	2,742	0.998	Rural
Rank 4	Rule 50 for 1	2,614	0.998	Rural
Rank 5	Rule 3 for 1	2,232	0.995	Rural
Rank 6	Rule 187 for 1	2,228	0.997	Urban
Rank 7	Rule 24 for 1	1,988	0.984	Rural
Rank 8	Rule 117 for 1	1,961	0.998	Urban

The top four urban and four rural rules which indicate the likelihood of a fatal accident are summarised in Fig. 5.15 and 5.16 respectively.



```
Rule 180 for  1 (3,682; 0.993)
   if      CL_URBAN_RURAL_DESC = Urban
   and     CL_LIGHT_CON_DESC = Darkness - lights lit
   and     CL_WEATHER_DESC in [ "Other" "Raining + high winds" "Fine no high winds" "Fine + high winds" "Raining no high winds" "Snowing no high winds" ]
   and     CL_ROAD_SURFACE_DESC in [ "Wet or damp" ]
   and     CL_ROAD_TYPE_DESC in [ "Single carriageway" ]
   then    1
Rule 125 for  1 (3,271; 0.998)
   if      CL_URBAN_RURAL_DESC = Urban
   and     CL_LIGHT_CON_DESC = Daylight
   and     CL_WEATHER_DESC in [ "Fine no high winds" ]
   and     CL_ROAD_SURFACE_DESC in [ "Dry" ]
   and     CL_ROAD_TYPE_DESC in [ "Slip road" "Single carriageway" "Dual carriageway" "One way street" ]
   then    1
Rule 187 for  1 (2,228; 0.997)
   if      CL_URBAN_RURAL_DESC = Urban
   and     CL_LIGHT_CON_DESC = Daylight
   and     CL_WEATHER_DESC in [ "Raining + high winds" "Fine no high winds" "Raining no high winds" ]
   and     CL_ROAD_SURFACE_DESC in [ "Wet or damp" ]
   and     CL_ROAD_TYPE_DESC in [ "Single carriageway" ]
   then    1
Rule 117 for  1 (1,961; 0.998)
   if      CL_URBAN_RURAL_DESC = Urban
   and     CL_LIGHT_CON_DESC = Darkness - lights lit
   and     CL_WEATHER_DESC in [ "Fine no high winds" ]
   and     CL_ROAD_SURFACE_DESC in [ "Dry" ]
   and     CL_ROAD_TYPE_DESC in [ "Single carriageway" "Dual carriageway" "One way street" ]
   then    1
```

Figure 5. 15 V21 C5.0 top urban rules



```
Rule 79 for  1 (2,742; 0.998)
    if      CL_URBAN_RURAL_DESC = Rural
    and     CL_ROAD_SURFACE_DESC in [ "Wet or damp" ]
    and     CL_ROAD_TYPE_DESC in [ "Single carriageway" ]
    and     CL_LIGHT_CON_DESC in [ "Darkness - no lighting" "Daylight" "Darkness - lights lit" ]
    and     CL_JUNCT_DETAIL_DESC in [ "Not at junction or within 20 metres" ]
    then    1
Rule 50 for  1 (2,614; 0.998)
    if      CL_URBAN_RURAL_DESC = Rural
    and     CL_WEATHER_DESC in [ "Fine no high winds" ]
    and     CL_ROAD_SURFACE_DESC in [ "Dry" ]
    and     CL_ROAD_TYPE_DESC in [ "Single carriageway" ]
    and     CL_LIGHT_CON_DESC in [ "Darkness - no lighting" "Daylight" "Darkness - lights lit" ]
    then    1
Rule 3 for  1 (2,232; 0.995)
    if      CL_URBAN_RURAL_DESC = Rural
    and     CL_WEATHER_DESC in [ "Fine no high winds" ]
    and     CL_ROAD_SURFACE_DESC in [ "Dry" ]
    and     CL_ROAD_TYPE_DESC in [ "Dual carriageway" ]
    and     CL_LIGHT_CON_DESC in [ "Darkness - no lighting" "Daylight" "Darkness - lights lit" "Darkness - lighting unknown" ]
    then    1
Rule 24 for  1 (1,988; 0.984)
    if      CL_URBAN_RURAL_DESC = Rural
    and     CL_LIGHT_CON_DESC = Darkness - no lighting
    and     CL_WEATHER_DESC in [ "Other" "Fine no high winds" "Fog or mist" "Unknown" "Raining no high winds" "Snowing no high winds" ]
    and     CL_ROAD_SURFACE_DESC in [ "Wet or damp" ]
    and     CL_ROAD_TYPE_DESC in [ "Dual carriageway" ]
    then    1
```

Figure 5. 16 V21 C5.0 top rural rules

Once the rules from the training model were identified, the validation data was scored against the model. Table 5.14 summarises the top ten rules with the highest prediction for fatal traffic accidents extracted from the validation dataset, based on the highest frequency counts, and compared to the rule ranking for the training data. Confidence indicates the likelihood of the predicted outcome once all of the conditions in the rule are true.

Table 5. 14 14 V21 C5.0 top 10 rules for validation data

Validation Ranking	Target Value	Predicted Value	Rule Ref.	1st Predictor	Validation Frequency Count	Confidence	Training data ranking
Rank 1	1	1	1_50	Rural	501	0.998	Rank 4
Rank 2	1	1	1_125	Urban	230	0.998	Rank 2
Rank 3	1	1	1_79	Rural	211	0.998	Rank 3
Rank 4	1	1	1_3	Rural	148	0.995	Rank 5
Rank 5	1	1	1_117	Urban	119	0.998	Rank 8
Rank 6	1	1	1_180	Urban	63	0.993	Rank 1
Rank 7	1	1	1_88	Rural	37	0.998	
Rank 8	1	1	1_187	Urban	34	0.997	Rank 6
Rank 9	1	1	1_24	Rural	28	0.984	Rank 7
Rank 10	1	1	1_26	Rural	27	0.987	

The V21 decision tree, outlined in Fig. 5.14, identified urban and rural as the 1st predictors. All of the top rules ranked in the training data remained strong predictors for the validation data although the order had changed. As training data was based on data from 2005 to 2012 some alteration in prevalent predictors would be expected as fatal traffic accident characteristics change over time. However, it is positive that there has not been a fundamental change in the top predictors.

The top two rules for urban are outlined in Table 5.15. In both cases the target and the predicted value were both fatal. Rule 1_125 predicts that in urban areas, in daylight, where fine weather and no high winds with dry roads that fatal accidents are most likely to occur on single carriageway. In 2013 STAT19 data, this represented 197 counts of fatal accidents from the total 230 identified in the data. Rule 1_117 predicts that in urban areas, in darkness but lights lit, where fine weather and no high winds on dry roads that, again, fatal accidents are most probable on single carriageway, with 94 of the total fatal accident count for this rule.

Table 5. 15 V21 C5.0 top urban rule description

Rule Ref.	1st Predictor	2nd Predictor	3rd Predictor	4th Predictor	5th Predictor	Validation Frequency Count	Confidence
1_125	**Urban**	Daylight	**Fine no high winds**	Dry	Dual carriageway	28	
					One way street	4	
					Single carriageway	197	
					Slip road	1	
						230	0.998
1_117	**Urban**	Darkness - lights lit	**Fine no high winds**	Dry	Dual carriageway	23	
					One way street	2	
					Single carriageway	94	
						119	0.998

The top two rules for rural are outlined in Table 5.16. Rule 1_50 predicts that in rural areas, where fine weather and no high winds, with dry roads on single carriageway that fatal accidents are most likely to occur in daylight represented by 374 counts of fatal accidents from the total 501 identified in the data. Rule 1_79 predicts that in rural areas, with wet or damp roads, on single carriageway, not at or within 20 metres of a junction that fatal accidents are most probable during daylight, with 110 of the total fatal accident count of 211 for this rule.

Table 5. 16 V21 C5.0 top rural rule description

Rule Ref.	1st Predictor	2nd Predictor	3rd Predictor	4th Predictor	5th Predictor	Validation Frequency Count	Confidence
1_50	Rural	Fine no high winds	Dry	Single carriageway	Darkness - lights lit	31	
					Darkness - no lighting	96	
					Daylight	374	
						501	0.998
1_79	Rural	Wet or damp	Single carriageway	Darkness - lights lit	Not at junction/<20 metres	18	
				Darkness - no lighting	Not at junction/<20 metres	83	
				Daylight	Not at junction/<20 metres	110	
						211	0.998

The V21 decision tree is quite broad on first review with many of the predictors further down the branches appearing general and difficult to extract specific factors to predict fatal traffic accidents. However, reviewing the top rules and comparing against the validation data provides insights which are clearer to understand and a more meaningful link between predictors. Although the C5.0 normal minority boosting model, did not perform well on accuracy, recall was very good and the rules extracted provided a clear understanding of key factors that can predict fatal traffic accidents. The model rules created based on training data performed well when scored against the 2013 STATS19 validation data.

5.2.4 Model 3: V4 Chaid reduced no resampling

Technical evaluation

”V4 Chaid reduced no resampling” was selected as a top performing model as it had better accuracy than most models and good recall at 0.680 and 0.726 respectively. Recall and accuracy results were not significantly different but the model performed well at classifying fatal accidents as fatal. Misclassifications were lower than the previously discussed V21 C.50 model, however 172 or 33% of non-fatals were misclassified as fatal and 31 or 27% of fatals were misclassified as non-fatal. Data reduction succeeded in extracting predictions from the data which was not originally possible. The model performed well and produced largely balanced results for accuracy. ROC curves for training, test and validation are presented for this model in Figs. 5.17 and 5.18. The V4 ROC curves behaved quite similarly to the V21 C5.0 model, with evidence that the model overfitted the validation data, represented by a smaller AUC.

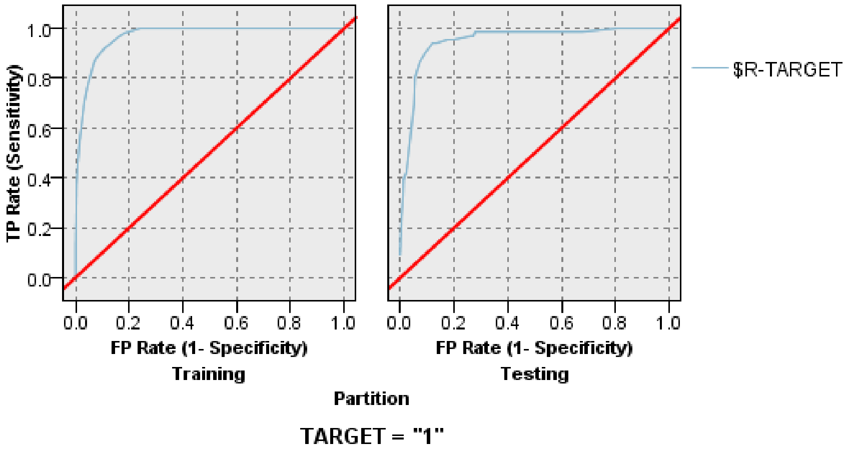


Figure 5. 17 Training & test ROC for Chaid reduced no resampling

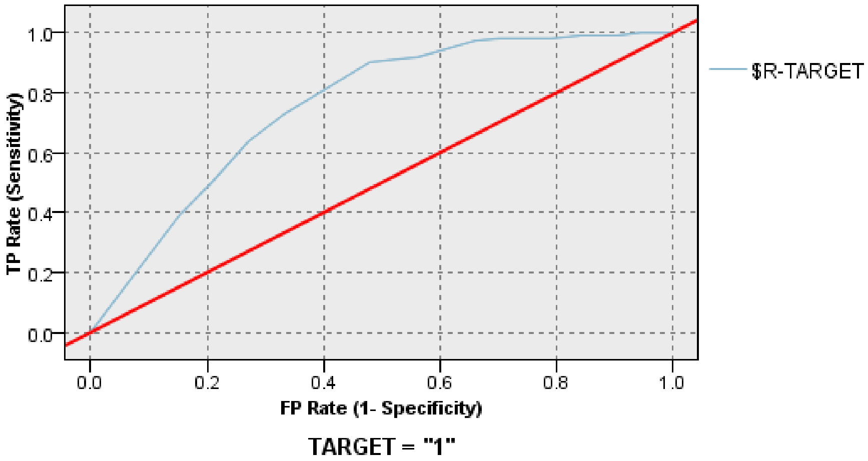


Figure 5. 18 Validation ROC for Chaid reduced no resampling

Non-Technical evaluation

Fig. 5.19 presents individual predictor importance for “V4 Chaid reduced with no resampling”. Road surface and light conditions were the most important predictors accounting for 34% and 22% respectively. Road type accounted for 18%, junction detail 12%, weather 7% and urban/rural the remaining 6%.

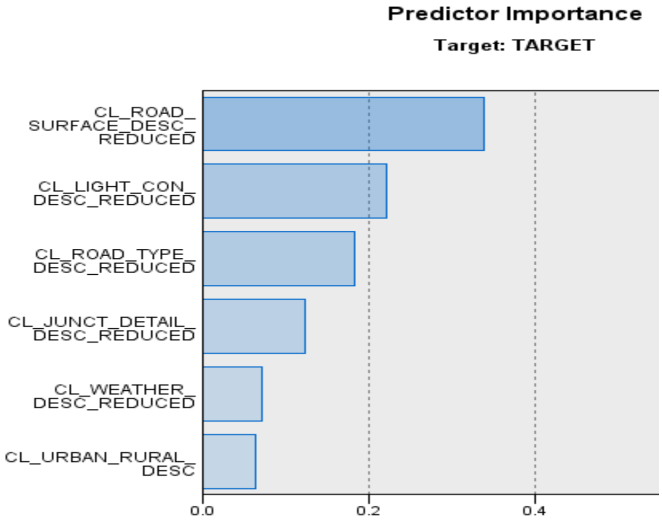
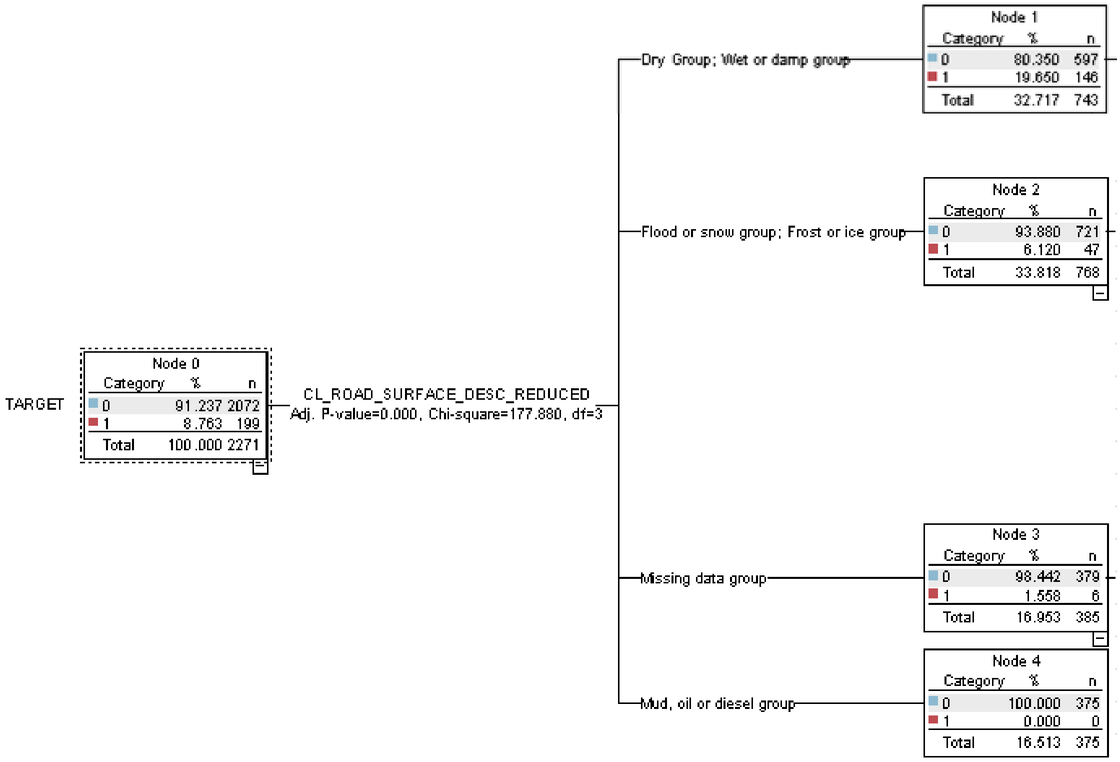


Figure 5. 19 V4 Chaid reduced no resampling predictor importance

As with V21, V4 model learns rules from the data and presents them in a decision tree format. Fig. 5.20 displays the V4 Chaid model with the most significant node or root node presented on the left i.e. fatal (1) and non-fatal (0). The first predictor was then identified and for this model was road surface.


TARGET
CL_ROAD_SURFACE_DESC_REDUCED
Adj. P-value=0.000, Chi-square=177.880, df=3
Dry Group; Wet or damp group
Flood or snow group; Frost or ice group
Missing data group
Mud, oil or diesel group

Node 0

Category	%	n
🟦 0	91.237	2072
🟥 1	8.763	199
Total	100.000	2271

Node 1

Category	%	n
🟦 0	80.350	597
🟥 1	19.650	146
Total	32.717	743

Node 2

Category	%	n
🟦 0	93.880	721
🟥 1	6.120	47
Total	33.818	768

Node 3

Category	%	n
🟦 0	98.442	379
🟥 1	1.558	6
Total	16.953	385

Node 4

Category	%	n
🟦 0	100.000	375
🟥 1	0.000	0
Total	16.513	375

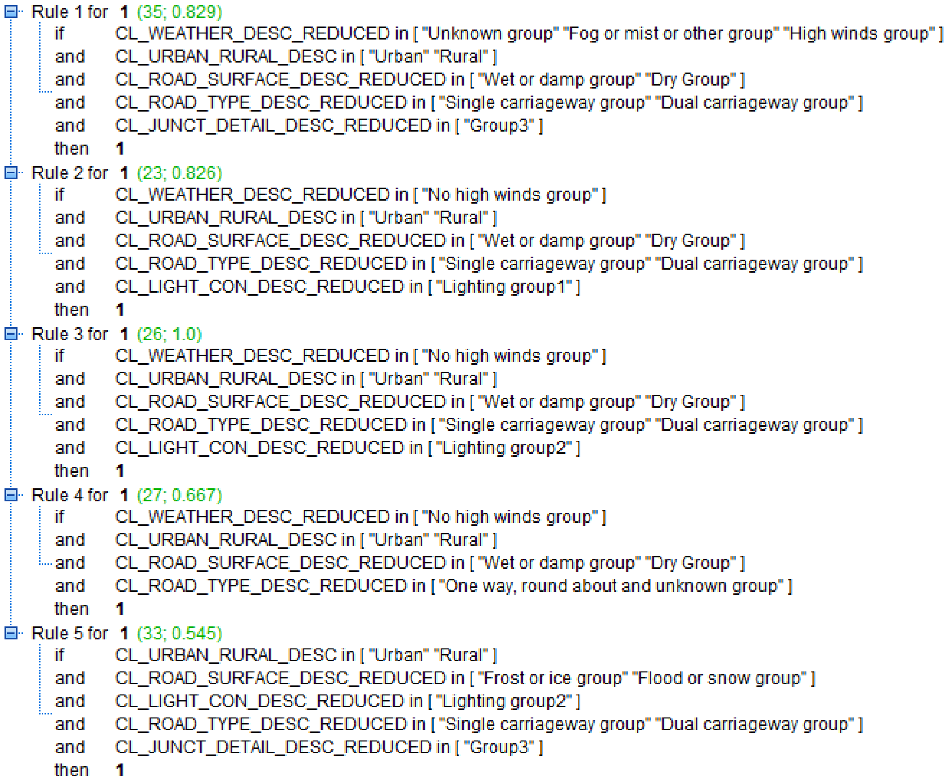
Figure 5. 20 Extract Chaid reduced no resampling decision tree

The V4 Chaid model only produced five rules for fatal accidents extracted from training data as listed in Table 5.17. The frequency count for this model are low as the model is based on less training data due to the reduction technique applied and no sampling technique was applied to correct any imbalance.

Table 5. 17 V4 top rules based on training data

Ranking	Rule Ref.	Training Data Frequency	Confidence
Rank 1	Rule 1 for 1	35	0.829
Rank 2	Rule 5 for 1	33	0.545
Rank 3	Rule 4 for 1	27	0.667
Rank 4	Rule 3 for 1	26	1.000
Rank 5	Rule 2 for 1	23	0.826

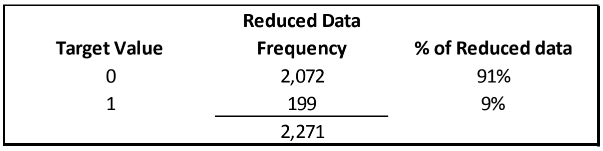
The five rules identified which indicate the likelihood of a fatal accident are summarised in Fig. 5.21.



```
Rule 1 for  1 (35; 0.829)
    if      CL_WEATHER_DESC_REDUCED in [ "Unknown group" "Fog or mist or other group" "High winds group" ]
    and     CL_URBAN_RURAL_DESC in [ "Urban" "Rural" ]
    and     CL_ROAD_SURFACE_DESC_REDUCED in [ "Wet or damp group" "Dry Group" ]
    and     CL_ROAD_TYPE_DESC_REDUCED in [ "Single carriageway group" "Dual carriageway group" ]
    and     CL_JUNCT_DETAIL_DESC_REDUCED in [ "Group3" ]
    then    1
Rule 2 for  1 (23; 0.826)
    if      CL_WEATHER_DESC_REDUCED in [ "No high winds group" ]
    and     CL_URBAN_RURAL_DESC in [ "Urban" "Rural" ]
    and     CL_ROAD_SURFACE_DESC_REDUCED in [ "Wet or damp group" "Dry Group" ]
    and     CL_ROAD_TYPE_DESC_REDUCED in [ "Single carriageway group" "Dual carriageway group" ]
    and     CL_LIGHT_CON_DESC_REDUCED in [ "Lighting group1" ]
    then    1
Rule 3 for  1 (26; 1.0)
    if      CL_WEATHER_DESC_REDUCED in [ "No high winds group" ]
    and     CL_URBAN_RURAL_DESC in [ "Urban" "Rural" ]
    and     CL_ROAD_SURFACE_DESC_REDUCED in [ "Wet or damp group" "Dry Group" ]
    and     CL_ROAD_TYPE_DESC_REDUCED in [ "Single carriageway group" "Dual carriageway group" ]
    and     CL_LIGHT_CON_DESC_REDUCED in [ "Lighting group2" ]
    then    1
Rule 4 for  1 (27; 0.667)
    if      CL_WEATHER_DESC_REDUCED in [ "No high winds group" ]
    and     CL_URBAN_RURAL_DESC in [ "Urban" "Rural" ]
    and     CL_ROAD_SURFACE_DESC_REDUCED in [ "Wet or damp group" "Dry Group" ]
    and     CL_ROAD_TYPE_DESC_REDUCED in [ "One way, round about and unknown group" ]
    then    1
Rule 5 for  1 (33; 0.545)
    if      CL_URBAN_RURAL_DESC in [ "Urban" "Rural" ]
    and     CL_ROAD_SURFACE_DESC_REDUCED in [ "Frost or ice group" "Flood or snow group" ]
    and     CL_LIGHT_CON_DESC_REDUCED in [ "Lighting group2" ]
    and     CL_ROAD_TYPE_DESC_REDUCED in [ "Single carriageway group" "Dual carriageway group" ]
    and     CL_JUNCT_DETAIL_DESC_REDUCED in [ "Group3" ]
    then    1
```

Figure 5. 21 V4 Chaid model rules

The rules produced from the V4 model were very general and did not provide a clear insight into the key predictors of fatal traffic accidents. In addition the rules produced did not follow the decision tree key predictors and confidence scores are inconsistent. This indicates that fatal traffic accidents was not well represented in the training data and, without the application of a sampling technique, the imbalance was not corrected and therefore the rules produced were limited and broad in scope. On review of the reduced data set, only 9% of data related to fatal accidents and given the overall sample size this means although rules were identified in the data they may not generalise well.



Target Value	Reduced Data Frequency	% of Reduced data
0	2,072	91%
1	199	9%
	2,271	

The V4 Chaid reduced no resampling decision tree, although performed better than many other models, did not produce actionable insights and would be unlikely to generalise well to a larger dataset. For decision trees to be effective, the non-technical evaluation is just as important, as without rules which can be well understood and generalised, good accuracy and recall cannot be actioned.

5.3 Subsequent Model Improvements

Although the focus of this experiment was on fatal accident recall, the accuracy results were very poor for many models evaluated. As previously discussed in chapter 3, when prediction is focussed on infrequent events, sampling techniques applied to improve the prediction of the infrequent event can negatively impact accuracy. In order to identify whether changes in parameters could improve model accuracy, two further parameter settings were selected to rebuild a sample of models. Boosting and Likelihood ratio instead of Pearson were selected as parameters for the model rebuild. Boosting can be used to enhance model accuracy by building models in sequence and learning from misclassifications to improve subsequent models and weighting to produce one overall prediction. Boosting impacts on the training time, however, for decision trees can significantly improve accuracy and the parameter is available in both C5.0 and Chaid.[12] For Chaid, SPSS Modeler offers Pearson and Likelihood ratio to calculate the Chi-squared ratio. Pearson was used for the original model build and is generally a faster calculation. Likelihood ratio is considered more robust and is the preferred method for small samples. Initial consideration was given to selecting models for rebuild based on the top performing models. However, “V13 Bayes net reduced no resampling” performed well for accuracy and recall so was not included in the rebuild. “V21 C5.0 normal minority boosting” was selected to rebuild using boosting. “V4 Chaid reduced no resampling” was selected to rebuild using boosting and Likelihood ratio instead of Pearson. As the initial rebuild of V4 using Likelihood ratio did not produce any changes in results, “V5 Chaid reduced majority reduction” and “V6 Chaid reduced minority boosting” were selected to rebuild using Likelihood ratio, to identify if the parameter would impact where resampling techniques had been applied. Table 5.18 outlines the results for accuracy and recall for the rebuilt models and the variance in results when compared to the original models.

[12] IBM, 2012. SPSS Modeler C5.0 Node Model Options. http://www-01.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50_modeltab.htm, Accessed 10 11 2014].

RB4 Boosting Chaid reduced no resampling

When boosting was applied to M4/V4 models, a new model “RB4 Boosting Chaid reduced no resampling” was created. All other parameters remained as per the original models. As outlined Table 5.18 results for training and test data improved for both accuracy, recall and precision. For validation data, accuracy dropped 0.079 but led to a substantial improvement in recall by 0.212 to 0.938 with only a slight reduction in precision of 0.022.

Table 5. 18 Boosting and likelihood ratio results

Boosting												
				Training data			Test data			Validation data		
Model Ref.	Modelling Technique	Sampling Technique	Parameter Setting	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
RB4	Chaid Reduced	No resampling	Boosting	0.985	0.915	0.910	0.946	0.712	0.770	0.601	0.938	0.301
M4/V4	Chaid Reduced	No resampling	N/A	0.946	0.553	0.764	0.918	0.424	0.700	0.680	0.726	0.323
	Variance Boosting vs original M4/V4			0.039	0.362	0.146	0.028	0.288	0.070	- 0.079	0.212	- 0.022
RB21	C5.0 Normal	Minority boosting	Boosting	0.999	1.000	0.999	0.992	0.661	0.494	0.684	0.880	0.215
M21/V2	C5.0 Normal	Minority boosting	N/A	0.995	1.000	0.991	0.981	0.701	0.245	0.594	0.917	0.179
	Variance Boosting vs original M21/V21			0.004	-	0.008	0.011	- 0.040	0.248	0.090	- 0.038	0.036

Likelihood												
				Training data			Test data			Validation data		
Model Ref.	Modelling Technique	Sampling Technique	Parameter Setting	Accuracy	Recall	Precision	Accuracy	Recall	Precision	Accuracy	Recall	Precision
RL4	Chaid Reduced	No resampling	Likelihood	0.946	0.553	0.764	0.918	0.424	0.700	0.680	0.726	0.323
M4/V4	Chaid Reduced	No resampling	Pearson	0.946	0.553	0.764	0.918	0.424	0.700	0.680	0.726	0.323
	Variance Likelihood vs original M4/V4			-	-	-	-	-	-	-	-	-
RL5	Chaid Reduced	Majority reduction	Likelihood	0.888	0.899	0.865	0.854	0.894	0.418	0.385	0.973	0.221
M5/V5	Chaid Reduced	Majority reduction	Pearson	0.881	0.935	0.838	0.869	0.924	0.449	0.352	0.991	0.215
	Variance Likelihood vs original M5/V5			0.006	- 0.035	0.027	- 0.015	- 0.030	- 0.030	0.033	- 0.018	0.007
RL6	Chaid Reduced	Minority boosting	Likelihood	0.896	0.874	0.910	0.887	0.879	0.487	0.446	0.956	0.238
M6/V6	Chaid Reduced	Minority boosting	Pearson	0.890	0.929	0.856	0.847	0.924	0.409	0.372	0.956	0.216
	Variance Likelihood vs original M6/V6			0.007	- 0.055	0.053	0.039	- 0.045	0.078	0.074	-	0.022

RB21 Boosting C5.0 normal minority boosting

Boosting was applied to M21/V21 models, where minority boosting sampling technique was applied, and “RB21 Boosting C5.0 normal minority boosting” was built. Variances for RB21 were not as significant as for RB4 with improvements in accuracy and precision in training and test data but reduction in recall for test data. Accuracy and precision improved by 0.090 and 0.036 respectively for validation data although recall reduced by 0.038 to 0.880.

RL4 Likelihood Chaid reduced no resampling

Likelihood ratio parameter was selected for M4/V4 instead of Pearson and a new model “RL4 Likelihood Chaid reduced no resampling” was created, however, the change in parameter had no impact on the model results. In order to assess if the parameter could impact the two other Chaid reduced models, M5/V5 and M6/V6 were selected for rebuild.

RL5 Likelihood Chaid reduced majority reduction

A new model "RL5 Likelihood Chaid reduced majority reduction" was created for M5/V5. Although a slight improvement in accuracy was produced for training data and validation data, recall for training, test and validation reduced slightly. Precision improved slightly for training and validation data with a slight reduction for test data.

RL6 Likelihood Chaid reduced minority boosting

A new model "RL6 Likelihood Chaid reduced minority boosting" was created for M6/V6. Accuracy and precision improved for all three data sets with the largest improvement of accuracy 0.074 in validation data. There was no change to the validation recall result, however, a slight reduction for training and test data.

Selection of optimal models, sampling techniques and parameters to identify the best prediction models is an iterative process with trial and error and repeated evaluation. The focus of this experiment was to identify whether C5.0 and Chaid decision trees and Bayes net, using three sampling techniques could extract prediction for fatal traffic accidents. Due to time constraints, three modelling techniques and three sampling techniques were selected applying SPSS Modeler standard parameters. As identified from the new models built and evaluated in this chapter, changes in parameters settings can have a positive impact on the results achieved. Suggested further work will be discussed in future work and research in chapter 6.

5.4 Key Findings

Based on the validation data evaluation, the following key findings were extracted from the experiment results:

- Decision trees did not perform well when no sampling technique was applied, with only one of six models predicting fatal accidents.
- Where no resampling was applied and prediction was achieved, accuracy results were higher than average although recall was significantly lower.
- C5.0 models where minority boosting was applied achieved very good recall and although accuracy rates were low they were better than all other models with sampling.
- Chaid reduced models with sampling techniques achieved the highest precision across models and maintained high recall results and although accuracy was low it was above the average accuracy for models with sampling.
- Bayes net models achieved prediction results for all nine models. Where sampling techniques were applied recall was excellent, however, accuracy rates and precision rates were significantly reduced.

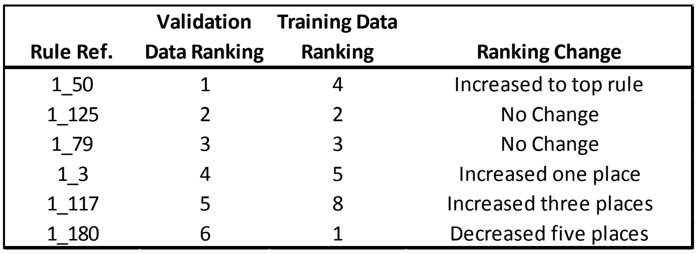
- For most models the application of sampling techniques to improve recall resulted in the classification of high volumes of false positives and therefore consistently low precision rates.

Key contributory factors identified

The approach taken to identify the key contributory factors was to review the top performing models. As mentioned in chapter 2, contributory factors identified depend on the characteristics of the data. The STATS19 accident dataset contains mainly environmental characteristics.

For the decision trees the model rules were extracted and reviewed. Unfortunately, although rules were extracted from the V4 Chaid model, they were quite general. As no resampling was completed, this model found it difficult to identify meaningful underlying patterns. The "V21 C5.0 normal minority boosting" decision tree model identified 196 rules for fatal accidents from the training data. When the model was applied to validation data 73 rules for fatal traffic accidents were highlighted. Table 5.19 displays the ranking of the most prevalent rules for validation and training. Interestingly the top five rules for training data were in the top six validation data rules although ranking had changed.

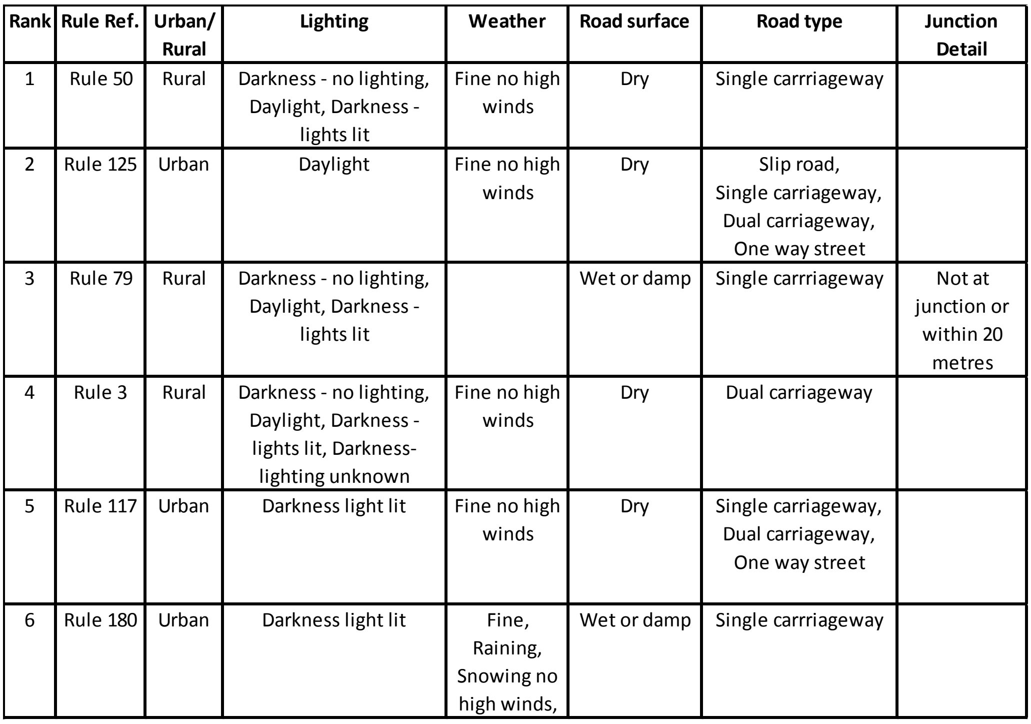
Table 5. 19 V21 C5.0 top ranked rules for validation and training data



Rule Ref.	Validation Data Ranking	Training Data Ranking	Ranking Change
1_50	1	4	Increased to top rule
1_125	2	2	No Change
1_79	3	3	No Change
1_3	4	5	Increased one place
1_117	5	8	Increased three places
1_180	6	1	Decreased five places

The contributory factors identified for the top six rules are summarised in Table 5.20. Fatal accidents occurred most frequently on rural single carriageways, on dry roads with fine weather with no high winds. Interestingly as discussed in the literature review in chapter 2, (Wah, et al., 2012) identified clear weather and dry road surface condition as being the strongest predictors of serious and fatal traffic accidents for motorbikes in Malaysia. Although lighting features are included in the rule there are no distinguishing features. This rule implies that in rural areas road surface and weather conditions are not the key cause of fatal accidents as most accidents occur when conditions are favourable. In urban areas, fatal accidents occur most frequently during daylight, on dry road surfaces and when weather is fine with no high winds. When multiple road types are listed it is difficult to identify a strong contributory factor. As with rural areas, fatal accidents are most likely to occur when favourable weather and road surface conditions exist.

Table 5. 20 V21 C5.0 top rules for validation data



Rank	Rule Ref.	Urban/ Rural	Lighting	Weather	Road surface	Road type	Junction Detail
1	Rule 50	Rural	Darkness - no lighting, Daylight, Darkness - lights lit	Fine no high winds	Dry	Single carrriageway	
2	Rule 125	Urban	Daylight	Fine no high winds	Dry	Slip road, Single carriageway, Dual carriageway, One way street	
3	Rule 79	Rural	Darkness - no lighting, Daylight, Darkness - lights lit		Wet or damp	Single carrriageway	Not at junction or within 20 metres
4	Rule 3	Rural	Darkness - no lighting, Daylight, Darkness - lights lit, Darkness- lighting unknown	Fine no high winds	Dry	Dual carriageway	
5	Rule 117	Urban	Darkness light lit	Fine no high winds	Dry	Single carriageway, Dual carriageway, One way street	
6	Rule 180	Urban	Darkness light lit	Fine, Raining, Snowing no high winds,	Wet or damp	Single carrriageway	

As discussed in 5.2.2 “V13 Bayes net reduced no resampling” identified road surface as the key meaningful indicators of fatal accident. Fatal accidents are most probable on dry road surfaces and only marginally more likely to occur in urban areas. A second contributory factor was identified where for wet or damp road surfaces fatal accidents the risk of a fatal accident increased for rural areas.

When comparing the both the C5.0 and Bayes net models, the contributory factors identified are consistent in that road surface and urban/rural are identified as the strongest predictors for both models. Although C5.0 provides more details on the relationship between factors, the ranking of key factors ties with the Bayes net key probabilities of factors contributing to fatal accident. Some factors were grouped into higher level groups which meant some rules are quite general resulting in a loss of meaningful insight. Further work could focus on ensuring more meaningful groups are assigned in the data preparation phase which could enhance the insights from the models.

5.5 Conclusion

The focus of this experiment was to build models which could predict fatal accidents and to identify contributory factors to fatal accidents. Models were built to focus on fatal accident classification. In this chapter, the models were assessed using training and test data and evaluated using STAT19 2013 data which the models had not previously seen. Most models achieved a high level of recall and correctly classified fatal accidents. By focussing the models on fatal classification, many misclassified non-fatal accidents as fatal where similar features existed and therefore low accuracy and precision rates were produced. Boosting and likelihood ratio were tested for the top models and the parameter changes resulted in some performance improvements. Applying further sampling techniques may improve the models overall performance, however, due to time constraints they were not included in the scope of this experiment.

The key findings are based on the evaluation phase and highlight the effectiveness of sampling techniques in achieving high fatal accident recall. The key contributory factors for C5.0 and Bayes net are consistent and were road surface and urban/rural. The rules identified from the C5.0 decision tree are easy to understand and could provide insight into the relationships between factors. Further work should focus on grouping data into more meaningful features which could identify more actionable insights from the experiment.

6 CONCLUSION AND FURTHER RESEARCH

6.1 Introduction

This chapter summarises the research completed as part of the experiment. The scope and objectives of the research are revisited and the achievement of those objectives and contributions to the body of knowledge are briefly discussed. The experiment approach is evaluated and limitations are discussed. Future work and research which could enhance the experiment is also briefly discussed.

6.2 Research Definition & Research Overview

The main objective of this research was to apply three classification techniques, C5.0 and Chaid decision trees and Bayes net, to predict fatal road traffic accidents based on a UK road safety dataset and to evaluate the model performance. Secondly, to identify the key contributory factors of fatal traffic accidents from the predictive models.

The research completed as part of this dissertation, commenced by reviewing academic literature related to road traffic accidents, data mining and predictive analytics. The understanding gained from this research, was incorporated into the design, implementation and evaluation of the research experiment and methodology adopted was based on CRISP-DM. The focus of the model design and build was to classify fatal traffic accidents and sampling techniques were adopted to improve fatal accident recall. The research achieved the following aims:

- data mining and predictive analytics literature was reviewed to identify suitable predictive and evaluation techniques relevant to traffic accident and infrequent event prediction
- the STATS19 data was analysed and data transformation was performed in order to prepare the data for modelling
- the design and implementation of the C5.0 and Chaid decision trees and Bayes net models achieved prediction for fatal traffic accidents
- the model performance was evaluated and changes to model parameters were tested and evaluated
- key findings from the experiment were identified and the model results were interpreted to extract the key contributory factors identified.

6.3 Contributions to the Body of Knowledge

After conducting the experiment and evaluating the results achieved, some findings could contribute to the body of knowledge.

- Using Chaid decision tree for supervised discretisation proved effective in identifying homogeneous subgroups in the data to a standard which would be useful for subject matter expert consideration. Chaid would serve as a good first level data transformation and the results would be presented in a manner which non-technical individuals would easily understand.

- The experiment demonstrated that classification techniques can be used to predict infrequent events once sampling techniques are applied.

- SPSS Modeler proved an effective tool for the experiment implementation for the data preparation phase and model build.

- Applying sampling techniques to classification models to address class imbalance can be effective in improving recall, however, consideration must be given to the resulting impact on accuracy and precision.

- Ranking the rules extracted from decision trees and summarising the key predictors in a standard format proved an effective means of understanding, interpreting and comparing key predictors. The literature review was limited with regard to methodologies or approaches to non-technical evaluation or interpretability of classification models.

6.4 Experimentation, Evaluation and Limitation

The intention of this experiment was to establish whether classification techniques would be effective in the prediction of fatal traffic accidents. In order to assess whether C5.0, Chaid or Bayes net performed better at meeting the objective, the scope of the experiment was limited to applying consistent parameters throughout the experiment.

This initial experiment design and approach to data preparation and model build proved successful in meeting the objective of predicting fatal accidents as was shown by the evaluation of the validation data. Recall results for validation data were very good, however, accuracy and precision performance was poor in many cases, with classification of non-fatals as fatals or false positives being the main performance issue.

A decision made as part of the initial experiment design was to focus on fatal accident prediction. This decision guided the data preparation and model build implementation. At the data discretisation stage, two target values of fatal and non-fatal were set. Two classes of STATS19 data, serious and slight accidents, were grouped as non-fatal. Given the extent of the false positives in the evaluation, the model struggled to classify accurately when similar features existed for fatal and non-fatal. It is likely that some of the serious accident class would have similar characteristics to fatal and this may be the cause of the high degree of false positives. It would have been better to rerun the experiment with three target values, fatal, serious and slight to establish if the model accuracy and precision rates would have improved, however, time constraints did not permit the rerun.

Instead of the experiment rerun, five additional models were built to test if changing parameters in the model build would improve the performance. Boosting was applied to the top C5.0 and Chaid models and Likelihood ratio instead of Pearson was applied to three Chaid models. The results showed some improvements in accuracy and precision and highlight that applying other sampling techniques or parameter settings may have improved accuracy and precision.

Literature review proved more challenging and time consuming than initially planned. Research identified generally related to specific narrow research questions or with a focus on statistical techniques. Methodologies, definitions and best practice papers were difficult to identify which limited the scope for relevant references with books providing the main source of explanation and research papers providing evidence based specific research.

A key limitation of the data understanding and preparation phase was the lack of consultation with a subject matter expert. Road traffic accidents characteristics and causes are widely varied and more meaningful groupings could have been extracted from the STATS19 dataset with practical knowledge of the field. Reliance was therefore placed on Chaid to identify homogenous groupings. It is likely that identification of meaningful contributory factors was limited by the data groupings. At the evaluation stage, interpretation of some decision tree rules proved difficult where factors within the rule were so wide ranging as to lose meaning for example lighting.

6.5 Future Work & Research

Future work and research is based on limitations identified as part of the experiment implementation and possible techniques to overcome them. An opportunity to apply the experiment to another research problem is also considered.

Changes to current experiment design

As previously discussed, rerunning the experiment with three target values instead of two may help to improve the false positive issue identified in this experiment. This would involve commencing the experiment from the initial data selection stage and redefining the target values as fatal, serious and slight instead of fatal and no-fatal as in the current research. In addition consideration of the data groups by a subject matter expert may provide additional insights which could help extract more meaningful contributory factors.

The experiment could also be expanded to consider the results obtained for serious and slight accidents and identify the related key predictors and contributory factors. This experiment extracted data from the STATS19 accident dataset. There are two other STATS19 datasets maintained vehicle and casualty and integrating the three datasets may provide additional insights. Unfortunately due to time constraints and insufficient data knowledge, it was not considered as part of this research experiment.

Support vector machine (SVM)

Research has demonstrated that SVM has been successful in improving the accuracy of cancer classification where clustering was applied before classification (Wahed, et al., 2012). As accuracy was the key limitation of this research experiment evaluation, this technique is of interest. Similarly if clustering was applied to traffic accident data, a rare event like cancer, before applying a classification technique, like SVM, the prediction accuracy may be improved. SVM classification is available in SPSS Modeler.

Consider applying the experiment to Irish road accidents

The experiment was completed based on the UK STATS19 data due to the availability, quality and wide use of the data. However, the experiment could also be applied to the Irish road accidents, although the scope may need to be widened as fatal accident volumes may not be sufficient. Road safety trends are in line with trends in the UK, as outlined in Fig. 6.1, with similar proportion of deaths by road user group.

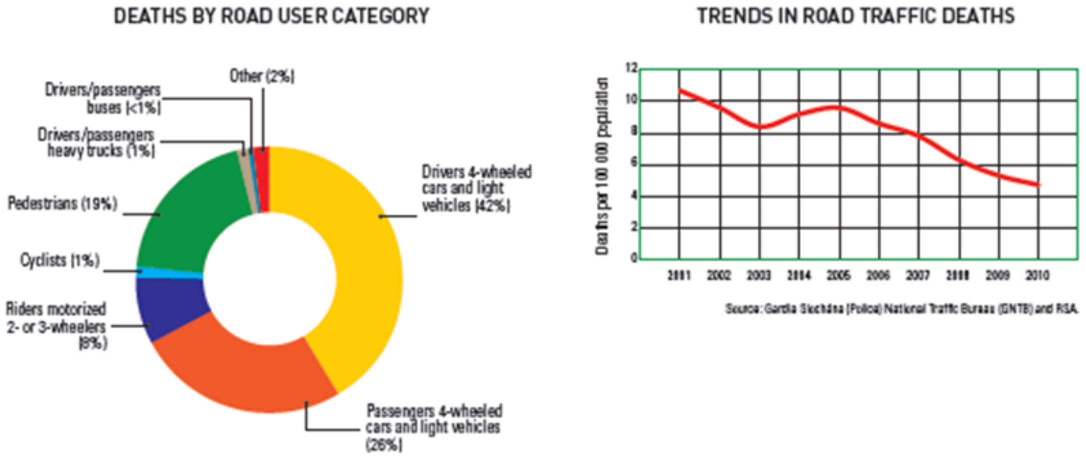


Figure 6. 1 Trends in Ireland road traffic accident deaths

Source: (The World Health Organisation, 2013, p. 130)

In order to assess the readiness in Ireland to meet the experiment requirements, a brief questionnaire was prepared and forwarded to a road safety professional in Ireland. The results of the questionnaire are presented in Appendix 1. From the reply it appears that road safety data is consistently recorded and reported and some consideration has already been given to the application of predictive analytics to road safety in Ireland.

6.6 Conclusion

This final chapter considers the experiment completed as part of the research and results achieved. The initial objectives achieved are outlined, together with contributions to the body of knowledge identified during the course of the research. The experiment achievements and limitations are discussed. Future work which could help overcome limitations in this experiment or add to the research learning is considered.

The experiment met many of the initial objectives and although accuracy performance was poorer than expected, fatal traffic accidents prediction was successful. Consideration has been given to further work which could improve the experiment results.

BIBLIOGRAPHY

Abugessaisa, I., 2008. *Analytical tools and information-sharing methods supporting road safety organizations,* Linköping: LiU-Tryck.

Aounallah, M., Quirion, S. & Mineau, G. W., 2004. Distributed Data Mining vs. Sampling Techniques: A Comparison. *Lecture Notes in Computer Science,* Volume 3060, pp. 454 - 460.

Baguley, C., 2001. *'The importance of a road accident data system and its utilization', International Symposium on Traffic Safety Strengthening and Accident Prevention, 28-30 November,* Nanjing: pp. 1-20.

Berry , M. J. & Linoff, G. S., 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* 2 ed. s.l.:Wiley.

Beshah, T. et al., 2013. Mining pattern from road accident data: Role of road user's behaviour and implications for improving road safety. *International journal of tomography and simulation,* 22(1), pp. 73 - 86.

Chapman P., et al, 2000. *CRISP-DM 1.0: Step by step data mining guide,* s.l.: SPSS.

Chong, M., Abraham, A. & Paprzycki, M., 2005. Traffic Accident Analysis Using Machine Learning Paradigms. *The International Journal of Computing and Informatics,* Volume 29, pp. 89 - 98.

David, S. & Branche, S., 2004. Road safety is no accident. *Journal of Safety Research,,* pp. 173 - 174.

Department of Transport UK, 2013. *Reported Road Casualties in Great Britain: Main Results 2013,* s.l.: Department of Transport.

Department of Transport UK, 2013. *Reported Road Casualties in Great Britain: Summary Results 2013,* s.l.: Department of Transport.

Eckerson, W., 2007. *Predictive Analytics: Extending the Value of Your Data Warehousing Investment,* s.l.: The Data Warehousing Institute.

Frawley, W., Piatetsky-Shapiro, G. & Matheus, C., 1992. Knowledge Discovery in Databases: An Overview. *AI Magazine,* Volume 13, pp. 213-228.

Grove, J., 2014. *Vehicle Licensing Statistics: 2013,* s.l.: Department of Transport UK.

Guillet, F. & Hamilton, H., 2007. Quality Measures in Data Mining. *Computational Intelligence and Complexity,* Volume 43, p. 120.

Han, J., Kamber, M. & Pei, J., 2011. *Data Mining: Concepts and Techniques.* Third ed. s.l.:Morgan Kaufmann.

He, H. & Garcia, E., 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering,* Volume 21, pp. 1263 - 1284.

Hermans, E., Brijs, T. & Geert, W., 2009. *Elaborating an Index Methodology for Creating an Overall Road Safety Performance Score for a Set of Countries.* Seoul, s.n.

Japkowicz, N., 2000. *Learning from Imbalanced Data Sets: A Comparison of Various Strategies,* s.l.: AAAI Press.

Kashani, A., Mohaymany, A. & Ranjbari, A., 2012. Analysis of factors associated with traffic injury severity on rural roads in Iran. *Journal of Injury and Violence Research,* 4(1), pp. 36-41.

Konstantinos, T. & Chorianopoulos, A., 2009. *Data Mining Techniques in CRM: Inside Customer Segmentation.* s.l.:Wiley.

Ling, C. & Li, C., 1998. *Data Mining for Direct Marketing: Problems and Solutions.* s.l., AAAI Press , pp. 73-79.

Lord, D. & Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* 44(5), pp. 291 - 305.

McCue, . C., 2007. *Data Mining and Predictive Analysis Intelligence Gathering and Crime Analysis.* s.l.:Butterworth-Heinemann; 1 edition.

Miner, G., Nisbet , R. & Elder, J., 2009. *Handbook of Statistical Analysis and Data Mining Applications.* 1 ed. s.l.:Academic Press.

Mitchell, K., 2002. Collaboration and information sharing: an ROI perspective?. *The Public Manager,* pp. 59 - 61.

Monfared, A. B. et al., 2013. Prediction of Fatal Road Traffic Crashes in Iran Using The Box-Jenkins Time Series Model. *Journal of Asian Scientific Research,* pp. 425-430.

Nyce, C., 2007. *Predictive Analytics White Paper,* s.l.: American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America.

Shearer, C., 2000. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of data warehousing,* 5(4), pp. 13-22.

Simoncic, M., 2004. A Bayesian Network Model of Two-Car Accidents. *Journal of Transportation and Statistics,* 7(23), pp. 13-25.

Souza, J., Matwin, S. & Japkowicz, N., 2002. *Evaluating data mining models: a pattern language.* s.l., s.n.

Tesema, T., Abraham, A. & Grosan, C., 2005. Rule Mining and Classification of Road Traffic Accidents using Adaptive Regression Trees. *International Journal of Simulation: Systems Science & Technology,* Volume 6, pp. 80-94.

The International Transport Forum, 2013. *Road Safety Annual Report 2013,* s.l.: OECD.

The International Transport Forum, 2014. *Road Saftey Annual Report 2014,* s.l.: OECD.

The Irish Road Safety Authority, 2013. *The Irish Road Safety Strategy 2013 - 2020.* [Online]
Available at: http://www.rsa.ie/Documents/About%20Us/RSA_STRATEGY_2013-2020%20.pdf
[Accessed 22 09 2014].

The Police Chief, 2005. Sobriety Checkpoints: An Effective Tool to Reduce DWI Fatalities. *The Police Chief Magazine*, 72(2).

The SAS Institute, 1998. *Data mining and the case for sampling,* s.l.: The SAS Institute.

The World Health Organisation, 2013. *Global status report on road safety,* s.l.: The World Health Organisation.

The World Health Organisation, 2013. *Global Status Report on road safety 2013 supporting a decade of action,* s.l.: The World Health Organization.

The World Health Organization, 2004. *World report on road traffic injury prevention,* Geneva: The World Health Organization.

Wahed, E., Emam, I. & Badr, A., 2012. Feature Selection for Cancer Classification: An SVM based Approach. *International Journal of Computer Applications,* 46(8), pp. 20 - 26.

Wah, Y., Nasaruddin, N., Voon, W. & Lazim, M., 2012. Decision Tree Model for Count Data. *Proceedings of the World Congress on Engineering,* 4 7.

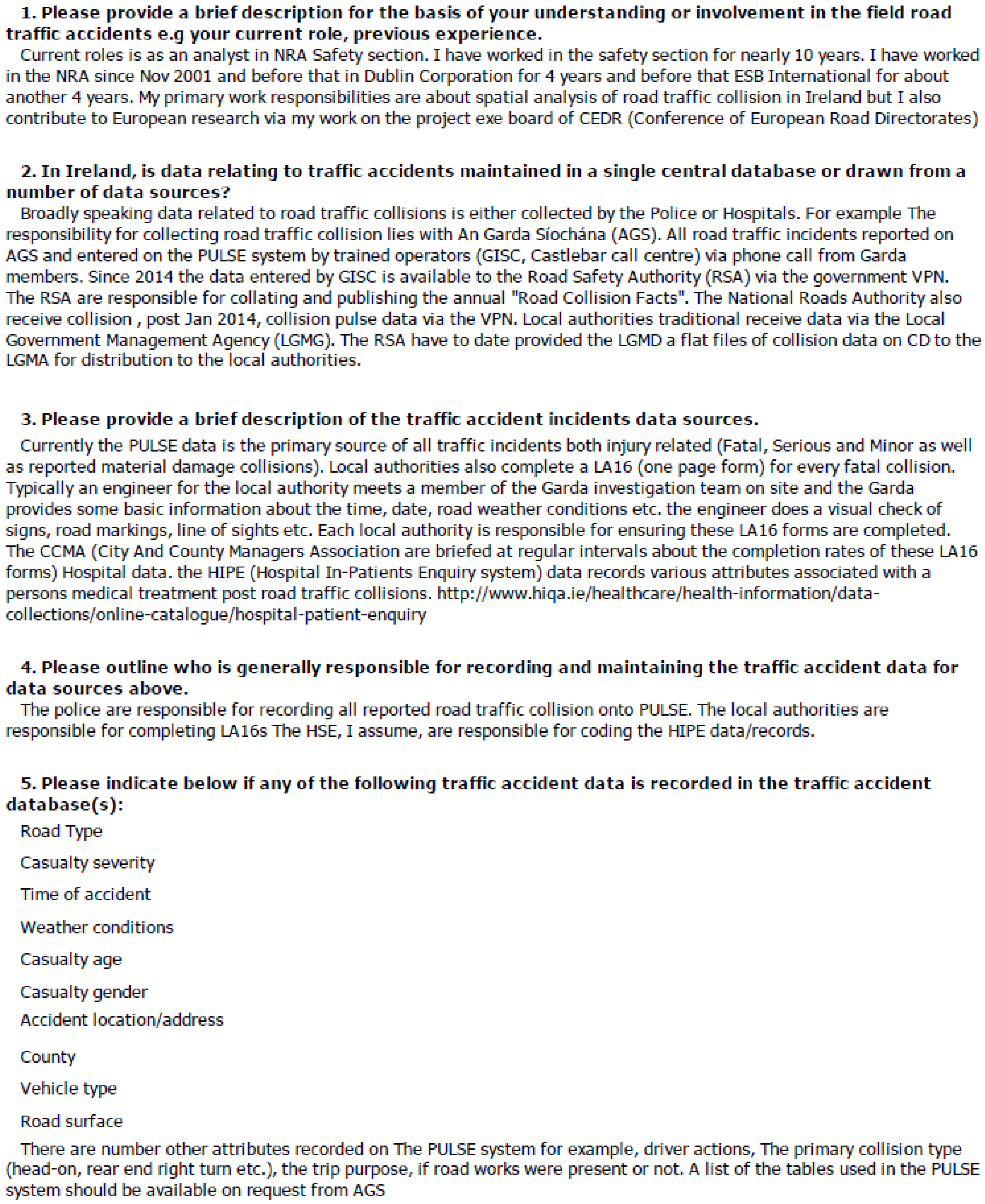
Weiss, G. & Hirsh, H., 2000. *'Learning to Predict Extremely Rare Events', AAAI Technical Report WS-00-05, Papers from the AAAI Workshop,* Menlo Park: AAAI Press.

Witten, I. H., Frank, E. & Hall, M. A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* 3 ed. s.l.:Morgan Kaufmann.

Zhang, S., Zhang, C. & Yang, Q., 2003. Data preparation for data mining. *Applied Artificial Intelligence: An International Journal,* 17(5-6), pp. 375-381.

APPENDIX 1 – ROAD SAFETY QUESTIONNAIRE

This questionnaire was completed by a road safety professional in Ireland and was intended to assess the whether this experiment could be applied to Irish road traffic accident data.



1. Please provide a brief description for the basis of your understanding or involvement in the field road traffic accidents e.g your current role, previous experience.

Current roles is as an analyst in NRA Safety section. I have worked in the safety section for nearly 10 years. I have worked in the NRA since Nov 2001 and before that in Dublin Corporation for 4 years and before that ESB International for about another 4 years. My primary work responsibilities are about spatial analysis of road traffic collision in Ireland but I also contribute to European research via my work on the project exe board of CEDR (Conference of European Road Directorates)

2. In Ireland, is data relating to traffic accidents maintained in a single central database or drawn from a number of data sources?

Broadly speaking data related to road traffic collisions is either collected by the Police or Hospitals. For example The responsibility for collecting road traffic collision lies with An Garda Síochána (AGS). All road traffic incidents reported on AGS and entered on the PULSE system by trained operators (GISC, Castlebar call centre) via phone call from Garda members. Since 2014 the data entered by GISC is available to the Road Safety Authority (RSA) via the government VPN. The RSA are responsible for collating and publishing the annual "Road Collision Facts". The National Roads Authority also receive collision , post Jan 2014, collision pulse data via the VPN. Local authorities traditional receive data via the Local Government Management Agency (LGMG). The RSA have to date provided the LGMD a flat files of collision data on CD to the LGMA for distribution to the local authorities.

3. Please provide a brief description of the traffic accident incidents data sources.

Currently the PULSE data is the primary source of all traffic incidents both injury related (Fatal, Serious and Minor as well as reported material damage collisions). Local authorities also complete a LA16 (one page form) for every fatal collision. Typically an engineer for the local authority meets a member of the Garda investigation team on site and the Garda provides some basic information about the time, date, road weather conditions etc. the engineer does a visual check of signs, road markings, line of sights etc. Each local authority is responsible for ensuring these LA16 forms are completed. The CCMA (City And County Managers Association are briefed at regular intervals about the completion rates of these LA16 forms) Hospital data. the HIPE (Hospital In-Patients Enquiry system) data records various attributes associated with a persons medical treatment post road traffic collisions. http://www.hiqa.ie/healthcare/health-information/data-collections/online-catalogue/hospital-patient-enquiry

4. Please outline who is generally responsible for recording and maintaining the traffic accident data for data sources above.

The police are responsible for recording all reported road traffic collision onto PULSE. The local authorities are responsible for completing LA16s The HSE, I assume, are responsible for coding the HIPE data/records.

5. Please indicate below if any of the following traffic accident data is recorded in the traffic accident database(s):

Road Type

Casualty severity

Time of accident

Weather conditions

Casualty age

Casualty gender

Accident location/address

County

Vehicle type

Road surface

There are number other attributes recorded on The PULSE system for example, driver actions, The primary collision type (head-on, rear end right turn etc.), the trip purpose, if road works were present or not. A list of the tables used in the PULSE system should be available on request from AGS

6. Are you aware if data is maintained or reported specific to fatal road traffic accidents in Ireland?

Yes

7. If Yes, please provide a brief description of the fata traffic accident data maintained or reported?

For each fatal collision a forensic investigation is conducted of trained Garda and compiled into reports including detailed scales drawings of the collision scene. I assume GNTB collate these documents. An LA16 is to be completed by the local authority for each fatal collision within their administrative area. The senior engineer or director of transport services is responsible for ensuring these forms are completed. The PUlSE data is update by the Police (via Castlebar call centre). If a serious injury unfortunately turns out to be a fatality (International 3 day definition protocol) then the investigating Garda will amend the PULSE record accordingly.

8. In your knowledge, has the data quality of the road traffic accident database been reviewed or analysed?

Yes

9. If Yes, please provide a brief description of your understanding of the data quality analysis prepared?

The NRA sponsored research into the data collected by Ireland on road traffic collision and benchmarked the type of information collected in other countries A summary of the reports are available online ...
http://www.nra.ie/safety/research/irish-collision-data-revi/Collision-Data-and-International-Benchmarking.pdf
http://www.nra.ie/safety/research/irish-collision-data-revi/Collision-Contributory-Factors.pdf

10. Based on your knowledge or use of traffic accident date, please outline your assessment of the quality of road traffic accident data?

Ireland road collision data is improving. The completeness of records has greatly improved since the introduction of the new 2014 PULSE system. The type and detail of the information sought by Police is amongst the best (as noted in the Risk Solution report cited above in Q9) I understand that there have been attempts to like the hospital and Police data in the past and this has not been as successful as hoped. At a European Level there is a move towards a common definition of road injuries using an existing trauma scale called MIAS (Maximum Abbreviated Injury Sale). However despite the current deficiencies in the ability to link hospital data to police data there have been papers produced providing very useful insights using Irish data ... http://www.itrn.ie/uploads/Short%20and%20Caulfield.pdf

11. Is there a formal reporting format for traffic accident data?

Yes

12. If Yes, please provide a brief description of the format for reporting traffic accident data?

I'm not entirely clear what you mean by this question. Since 2014 the trained operators in GISC "talk" the investigation Garda member through the collision asking and prompting, where appropriate, through the PULSE screens and complete the data entry for each collision reported to AGS. There are incidents where the Police are not informed of ... see Short and Caulfield paper cited above in Q10

13. Is there a formal reporting frequency or timeline for traffic accident data?

Yes

14. If Yes, please provide a brief description of the frequency or timeline for reporting traffic accident data.

As above ...

15. Who is generaly responsible for reporting traffic accident data?

The RSA are responsible for collating the raw collision data (from those road traffic collisions reported to AGS) into an annual publication. e.g. http://www.rsa.ie/Documents/Road%20Safety/Crash%20Stats/2011_Road_Collision_Fact_Book.pdf

16. In line with your understanding, please list the main users of traffic accident data?

Within the NRA and from an engineering road safety perspective the road collision data is used (in conjunction with exposure data in the form of veh km travelled) to rank the safety of the road network. see link below for further details http://www.nra.ie/safety/design-manual-roads-and-b/nra-hd15-network-safety-r/ AGS use the collision data to target speed enforcement and the location of "Safety Camera Zones". The RSA use the road collision data for updating and improving police around all aspects of road safety.

17. In Ireland, are traffic accident statistics published?

Yes

18. If Yes, please provide a brief description of the statistics published?

The RSA publish annual statistics. See link in Q 15 The NRA publish road collision rates for the national road network on their web site. See link below http://www.nra.ie/safety/design-manual-roads-and-b/nra-hd15-network-safety-r/HD15_AvgCollisionRates.pdf

19. Please outline who is generally responsible for publishing traffic accident statistics?

With the current Road Safety Strategy (2013 to 2020) there are numerous actions related to the publication of data and the responsible agency. In general it is the RSA who are responsible for publishing the collision statistics

20. Are traffic accident data or statistics in Ireland analysed to identify contributory factors?

Yes

21.

If Yes, please provide a brief description of the analysis prepared to identify traffic accident contributory factors?

Annually the road collision facts contain a breakdown of collisions by contributory actions (e.g. exceeded safe speed, Improper overtaking, failed to signal etc.) as well as as Collision classified by Weather, Surface Condition, Road Character and by whether skidding occurred or not at the scene of the collision, to name but a few..

22. What are the main objectives of this analysis?

To help explain some of the event that make have contributed to the collisions. For example during the engineering review of the "Network Safety Ranking" - see link below - engineers review patterns of collision to help target where engineering countermeasure are most likely to reduce the severity and frequency of collisions occurring on the national road network. http://www.nra.ie/safety/design-manual-roads-and-b/nra-hd15-network-safety-r/

23. Who are the main users of this traffic accident data analysis?

Very similar to Q16 on the main users of the collision data. There and many different types of analysis that are conducted using the collision data and I've mentioned some published analysis both on the RSA and NRA sites. In general the analysis is used internally and by other stakeholders within road safety as well as some users from academia.

24. What tools are used to analyse traffic accidents?

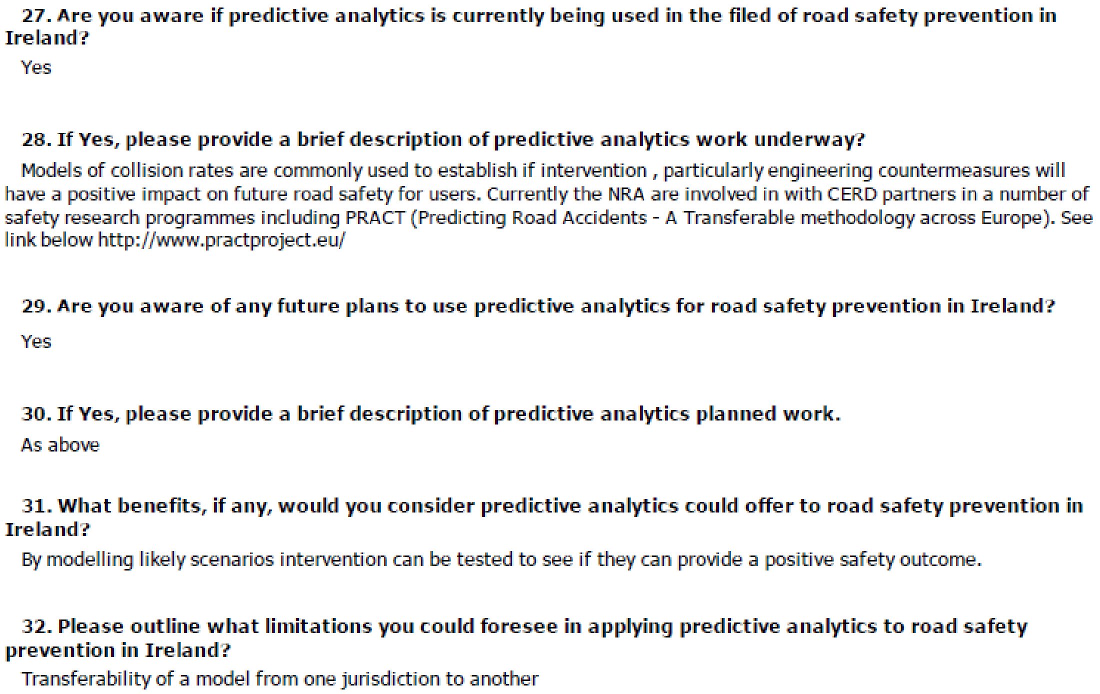
In HD15 there as spatial tools used to manipulate the various data sources such as GIS but also the is Excel and Access employed to produce the HD15 reports. these are distributed via SharePoint o the road safety engineers for review. SQL sever is used to create table views of the PUSLE data in the first instance.

25. Are traffic accident statistics reported to the European Union?

Yes

26. If Yes, please provide a brief description of the statistics reported to the European Union and the format?

Again the road Safety Strategy covers a number of actions and these include reporting to Eurostat. IRTAD and the ETSC are also commonly update with Irish road collision details.



27. Are you aware if predictive analytics is currently being used in the filed of road safety prevention in Ireland?

Yes

28. If Yes, please provide a brief description of predictive analytics work underway?

Models of collision rates are commonly used to establish if intervention , particularly engineering countermeasures will have a positive impact on future road safety for users. Currently the NRA are involved in with CERD partners in a number of safety research programmes including PRACT (Predicting Road Accidents - A Transferable methodology across Europe). See link below http://www.practproject.eu/

29. Are you aware of any future plans to use predictive analytics for road safety prevention in Ireland?

Yes

30. If Yes, please provide a brief description of predictive analytics planned work.

As above

31. What benefits, if any, would you consider predictive analytics could offer to road safety prevention in Ireland?

By modelling likely scenarios intervention can be tested to see if they can provide a positive safety outcome.

32. Please outline what limitations you could foresee in applying predictive analytics to road safety prevention in Ireland?

Transferability of a model from one jurisdiction to another