Predicting Claim Cost Life Cycles in the Motor Insurance Industry

Rebecca Haughton, B.A., Elaine O'Dwyer, B.Sc.

A thesis submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business

August, 2015

Supervisors: Dr. Miguel Nicolau, UCD and Mark Coveney, FTI Consulting

Head of School: Professor Anthony Brabazon

Dedication

To our family and friends who supported us throughout the year.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Algorithms

Preface

This research is done in part fulfilment of the MSc Business Analytics in the Michael Smurfit Graduate Business School, University College Dublin. The work has been undertaken in conjunction with FTI Consulting, a consulting firm who are heavily involved in analytics projects across a number of industries with a keen interest in insurance.

Research in this paper is carried out in relation to the motor insurance industry; more specifically the prediction of claim payments is investigated. The Central Bank of Ireland (2014) reported that in 2013, non-life insurers in Ireland took over €230million in profit alone with over half of the insurance industry's assets relating to investments. Achieving even a small increase in the accuracy of predicting outstanding liabilities can result in huge gains for an insurer. The model presented in this paper takes a novel approach in forecasting claim payments, with the aim of taking a step towards this goal.

Rebecca Haughton

Elaine O'Dwyer

Dublin, August 2015

Acknowledgements

We wish to sincerely thank our academic supervisor, Dr. Miguel Nicolau, for his guidance and support right throughout the project. He was extremely generous with his time and knowledge and for this we are very grateful.

We are indebted to our project sponsor, FTI Consulting, for providing us with this business problem and the data upon which to base our research. In addition, we thank them for the opportunity afforded to us to engage in their office environment. We were warmly welcomed into the company and are hugely appreciative for the assistance and encouragement we received, in particular from our business supervisor Mr. Mark Coveney.

We would like to express a word of thanks to the course director and all the lecturers involved in the MSc. Business Analytics in UCD's Michael Smurfit Business School. They have opened our eyes to the diverse world of business analytics and shared their wisdom and experience with us.

We thank the developers of the R software environment and the extended R community whose packages we have used extensively for our work. Finally, a thank you to the authors whose research we have studied and referenced in our project.

Abstract

The main motivation behind this project is a desire to better understand the life cycle of costs for motor insurance claims. Accurately forecasting future obligations is a challenging task for insurance companies and one which this work aims to assist. Current techniques look to estimate a company's future liabilities across all claims for a given time period as well as the total cost of individual claims. However, the central aim of this project is to develop a model capable of making predictions about a newly reported claim in a more detailed manner.

In particular, we wish to estimate a claim's expected duration and the proportion of its total cost to be paid out in a given month. Rather than focusing on the actual payment amounts, this approach allows for claims which show similar payment patterns be grouped together, independent of the payment magnitude. Equipped with claims data from a large motor insurer in Ireland, transactional data is presented in matrix form detailing the proportion of the claims' total cost paid each month. Clustering techniques are applied with the most effective, the k-medoids algorithm CLARA, forming five stable clusters. These clusters are analysed in detail and the profile of claims within each group examined. Using claim characteristics other than transactional data, a classifier is developed which assigns a claim to one of the clusters thus allowing predictions to be made about its life cycle.

Testing has found that even the best performing classifier, generated using Support Vector Machines, is not sufficiently accurate to deem the model effective. Factors such as the class imbalance problem, covariance shift and the absence of possibly important variables relating to the incident are considered as contributors to this poor performance. Success at the clustering phase of the project does offer some insight into claim cost behaviour. In the future, improvements at the classification stage would allow for the opportunity to pair the model with an estimator of the total claim costs and convert the predicted proportions into costs.

List of important abbreviations

ANN	Artificial Neural Network
CLARA	Clustering Large Applications
FNOL	First Notice of Loss
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbour
SOTA	Self-Organising Trees Algorithm
SVM	Support Vector Machines

Chapter 1 - Introduction

1.1 Opening Remarks

This research sets out to investigate the feasibility and precision of a machine learning approach to the task of predicting the life cycle of insurance claim costs. In particular, an emphasis is placed on the proportion of a claim's total cost paid out in monthly time intervals. Focusing on monthly transactions in proportional form allows for common life cycles emerge, without the influence of the actual payment amount.

One of the most prevalent and challenging problems in the insurance industry surrounds methods of claims (loss) reserving (Boland, 2007). Predicting future liabilities, i.e. outstanding claim costs, is important not only in the pricing of future premiums but also in identifying surplus cash used for investments. Last year's figures tell us that just over 50% of income for Aviva Ireland came from investments while 71% of their assets were made of financial investments (Aviva plc, 2014).

Thus, accuracy in claim reserving can have a knock on effect throughout the business. This project looks at supplementing the challenge of loss reserving by extending our understanding of expected payment patterns and forecasting for monthly time intervals on a claim-by-claim basis.

1.2 Research Context

Data mining can be described as a step in the Knowledge Discovery in Databases (KDD) process which enables the extraction of novel and potentially useful knowledge. It has the ability to uncover previously unknown and even unanticipated patterns in highly complex data through a variety of methods. Within data mining, machine learning techniques are among those widely called upon and used. These techniques operate by learning rules from the instances provided and then use them in the prediction of future behaviours (Kotsiantis, 2007).

Modern technology has allowed for more and more information to be collected and stored by insurance companies in relation to all aspects of the business. Therefore, applying machine learning techniques to their data has the potential to give insurance companies a business advantage in a highly competitive market and to either augment or replace traditional actuarial methods.

Within the industry, research involving data driven approaches has been rendered successful in a number of areas by enhancing old models and discovering new, significant relationships (Devale, 2012). Applications of machine learning in this sphere appear to follow two major pathways: methods which aim to predict the end of claim outcomes and methods which predict the most appropriate class to which a claim belongs. Examples include:

- Predicting the duration of claims relating to income protection insurance (Liu et al, 2014).
- The use of clustering in the task of risk analysis to identify customer segments according to their distinct levels of risk (Kumar & Singh, 2012; Yeo et al, 2001);
- Fraud detection in motor insurance claims (Viaene et al, 2002);

Prior to experiments with machine learning techniques, stochastic methods were, and still are, widely applied in the insurance industry. One such application is in the task of predicting the total settlement cost of claims. As stressed by Grize (2015), estimating these future costs is vital as insurance companies need to reserve adequate amounts to cover these impending claims. Reserving is not only a factor in maintaining customer satisfaction, it also contributes to the determination of future premiums and is imperative due to restrictions most currently detailed in Solvency II, an EU directive which consolidates insurance regulation across the region (Central Bank of Ireland, 2015).

Unfortunately, these actuarial techniques have to date focused on aggregated data for all claims over a certain period of time. For these techniques to be applied and their validity justified, Grize (2015) warns of the necessity to assume that all claims are homogeneous in their payment process and stationary in their development over time. An additional drawback is the inability of these methods to provide details on factors such as claim duration or the occurrences of payments and their sizes (Plat & Antonio, 2014).

As discussed in Section 2.3, evidence suggests that claim behaviour varies considerably according to different characteristics, e.g. claim type or involvement of third parties. The method proposed in this paper focuses on prediction at a micro level by attempting to estimate the proportion of a claim paid at each monthly time

interval. This is an aspect of claim prediction not previously addressed in the literature but one which would provide insurance companies with a more in depth and detailed method of assessing future liabilities.

1.3 Research Question

With the previous points in mind, we pose the question: can machine learning methodologies develop our understanding, and the prediction, of the life cycle of motor insurance claims? We specifically analyse the stages of payments through which a claim passes, from when it is reported until fully settled, and ask if there are any emergent groups of similar claims. In uncovering such clusters, we aim to develop a model capable of estimating the life cycle of a new claim and predicting the proportion of its total cost payable in future months based on its cluster membership.

1.4 Scope of Research

Insurance companies offer policies covering a wide range of scenarios which can be broadly divided into two distinct categories, namely life insurance and non-life insurance. Due to varying durations and methods of processing claims of both types, a decision was made to focus on non-life insurance data. In conjunction with our business sponsors, FTI Consulting, motor insurance was selected as the particular area of research. This was partly due to the availability of data and also related to their specific area of interest.

The predictive model developed from this research project looks to forecast proportions of the total claim cost payable in a given month and not the payment amounts themselves. This is in keeping with the aim of grouping claims with similar payment paths, regardless of the monetary value of the monthly transaction total.

During the research process, we look to build a classification model using only First Notice of Loss (FNOL) information, i.e. information known at the time a claim is reported. Such a restriction is necessary if the classifier is to be incorporated into the prediction process at the earliest stage possible. Furthermore, as a short written description of the incident is usually included in this first report, the application of text mining techniques could enhance the classification stage. However, it was

decided that text mining analytics is out of the scope of this paper but is an area which could merit further investigation.

1.5 Outline of the thesis

In this paper we take the following approach:

- A review of existing literature is conducted in Chapter 2 with two particular purposes in mind. We firstly analyse actuarial methods used in the area of claim cost predictions and highlight the absence of work relating to the stages of payments of a claim. We also critique the techniques of clustering and classification, as these are the machine learning tools used to develop our model, and examine the areas of insurance to which they have been applied.
- A thorough analysis of the motor claims dataset is conducted in Chapter 3 with details relating to its cleansing and pre-processing described. This phase of descriptive analytics is necessary in order to develop an understanding of the data and ensure it is in a suitable form for the remaining steps.
- The methodology of the project is contained in Chapter 4 with the results of the main stages of clustering and classification being reported in Chapter 5 and 6 respectively. In Chapter 6 we also disclose our findings relating to the overall model and compare its outcome to that of the actual data.
- Finally, the significance of our findings is discussed in Chapter 7 before concluding the paper and proposing possible improvements and extensions to the project in Chapter 8.

1.6 Contributions

The aim of this research was to use a novel approach to predict claim cost life cycles in the motor insurance industry. In doing so we focused on clustering claim transactions in proportional form which allowed us determine any naturally evolving clusters of claims. Following a study of the key characteristics of each cluster, a classification model was developed to determine the cluster to which a new claim belongs. A cluster-specific predictive model then estimates the proportion of the total claim paid out in subsequent months. The central assumption underlying this approach is the existence of a relationship between the payment patterns of a claim and other information relating to the incident.

We found that, when clustered on their monthly proportional payments, claims do fall into groups of similar life cycles. The best performing algorithm uncovered 5 such clusters and appeared to be strongly influenced by the month in which the majority payment was made. Unfortunately, to this point, no satisfactory classifier has been developed capable of assigning claims to one of these clusters with a high degree of accuracy.

Once we determine the group to which any new claim belongs, estimations regarding its life cycle can be made. The research conducted in this paper is novel in its approach to analysing claim costs, both through its proportional expression of costs and focus on monthly transactions. The successful outcome of the clustering stage enhanced our project sponsor's understanding of claim payment behaviours and has the potential to improve the forecasting of future obligations.

Chapter 2 - Literature Review

2.1 Overview

We begin this chapter by outlining current actuarial techniques used to predict claim costs in the insurance industry. Research which approaches the task of claim prediction by grouping claims is then documented. The machine learning techniques of clustering and classification are reviewed with an emphasis on evaluating the findings of previous research relating to insurance. By critiquing the works of others we aim to identify methods appropriate for our project.

2.2 Current Practices

Several actuarial and statistical techniques exist which serve to predict claim costs. Grize (2015) informs us that these methods have been classically deterministic. He also declares that, with the exception of a small number of isolated works, it is only since the turn of this century and the seminal paper by England & Verrall (2002) that more stochastic estimation models are being used. This section outlines some of the existing prediction techniques in order to better understand current practices and identify their potential shortfalls.

2.2.1 Chain Ladder Method

Even the most recent literature, such as that by Grize (2015), acknowledges the chain ladder method as the traditional classical technique to estimate payments in future years. Deterministic in nature, this actuarial method is still very commonly used, partly owing to the fact that is distribution free and so widely applicable (Wüthrich & Merz, 2008). Literature has seen many variations of the chain ladder method proposed, e.g. the double chain ladder by Martínez-Miranda et al (2012). Others have developed stochastic versions of the method such as a bootstrapping approach which requires simulation (England & Verrall, 2002).

In its most basic form, the chain ladder method requires the claims data to be in a two-dimensional table where the rows correspond to the year in which a claim originated. This ‘origin year’ is often referred to as the ‘accident year’ given the extensive historical use of this method in the motor insurance industry (Boland,

2007). The columns of the table represent the 'development year', i.e. the year in which (partial) payments were made.

Data is usually aggregated over a yearly time period with the rows giving cumulative claims losses settled for claims incurred in each accident year. A minor adaptation to this format is the inclusion of incremental claims per year rather than cumulative values. Boland (2007) highlights the advantages of this representation which allows the consideration of factors such as inflation and the standardisation of payments.

In order to estimate future claim payments, the chain ladder method calculates development factors $d_{i|j}$, where each $d_{i|j}$ is the development factor from origin year i to development year j. These factors are averaged across all development years and used to predict future payments. Alternatively, and more simply, Boland (2007) also points out the possibility of using the development factors from a single year as the basis for predictions if it was strongly felt to reflect the pattern of development in future years.

2.2.2 Bornhuetter-Ferguson Method

As explained by Boland (2007), the basic operations of this method are similar to that of the chain ladder. In fact Boland's overview of the Bornheutter-Ferguson method describes it as an extension of the chain ladder whereby necessary reserves are estimated using a projection technique, while additionally incorporating information on loss ratios. Alternatively, Schmidt (2006) looks at the connection from a different perspective considering the chain ladder as a special case of the Bornheutter-Ferguson. Despite the varying viewpoints, both authors are clear in their explanation of the key assumption underlying this and all similar methods of loss reserving: that the development of the losses of every accident year follows a pattern which is common, or at least similar, to all accident years.

This method works by combining information on how claim amounts and numbers develop over time, and comparing them with changes in losses relative to insurance premiums collected (Boland 2007). While the basic loss ratio for a given time period is simply the ratio of incurred or paid claims to the amounts earned in premiums, information regarding expenses and investment returns can be incorporated into the ratio and consequently into the forecast. With choices to be made in how exactly the

development patterns and expected cumulative claims are estimated, Schmidt (2006) labels the Bornheutter-Ferguson a general framework for loss reserving.

2.3 Prediction of Claims

Research suggests that identifying groups of similar claims can help in the prediction of claim outcomes. Kluppleberg & Severin (2003) find that claims of a car insurance portfolio behaved differently depending on the number of payments throughout their life cycles. They conclude that the ability to estimate the final number of instalments of an open claim is an important factor in forecasting the costs of future payments. Jessen et al (2010) focus on the prediction of the total number of payments across all claims to more accurately calculate outstanding obligations for a given year. However, this approach only considers aggregated data and does not look to make predictions for individual claims.

Closely linked is research by Liu et al (2014) who recognise the value in partitioning income protection insurance claims according to their anticipated duration. They argue that predicting the duration class to which a claim belongs, identifies policyholders with similar risks thus giving actuaries a better understanding of the portfolios underwritten. Revealing high risk policies, i.e. those linked to claims of long duration, also allows management to optimise their resources accordingly (Liu et al, 2014).

Through analysing micro level automobile insurance records, Frees & Valdez (2008) find that total costs for different 'types' of claims are related. Thus they partition motor insurance claims into three types: claims for injury to third party; claims for property damage to a third party; and claims for damages to the insured, including property and injury. Claims are subsequently modelled according to possible combinations of each type, yielding more efficient prediction of automobile claims compared with traditional methods (Frees & Valdez, 2008). Plat & Antonio (2014) also find different behaviour in relation to claims development patterns for claims involving material damage or bodily injury claims and therefore analyse each type separately.

2.4 Data Mining

In order to effectively predict the life cycle of a claim using data mining and machine learning, a solid understanding of the relevant techniques, their workings and characteristics is essential. More specifically, for the purposes of this paper an examination of clustering and classification is imperative.

2.4.1 Clustering

Clustering, a type of descriptive analysis emerges as a sub problem within this research project as a means of identifying patterns of claim payment paths. By examining the proportion of a claim paid in a particular time interval, claims which are settled in a similar fashion can be grouped together. This idea is just one application of clustering or segmentation of data which as a process achieves the "grouping a set of physical or abstract objects into classes of similar objects" (Han & Kamber, 2001). Moreover, greater homogeneity within a cluster and greater difference between clusters generates better groupings (Tan et al, 2006).

As far back as 2001, Han & Kamber acknowledged that cluster analysis as a branch of statistics has been examined at length for many years, mainly focusing on distance-based cluster analysis. Yet in data mining, clustering is still considered one of the most promising techniques of data analysis (Ghorpade-Aher & Metre, 2014). At its core is the process of unsupervised learning where objects are labelled with class (cluster) labels which themselves are derived from the data and not learned from existing instances (Tan et al, 2006). Cluster analysis is a difficult problem and factors such as the selection of effective similarity measure, criterion functions, algorithms and initial conditions must be considered when devising a well-tuned technique for a specific clustering problem (Jafar & Sivakumar, 2010).

2.4.2 Clustering Methods

Han & Kumar (2001) discuss clustering techniques under a number of headings. Referring to a collection of clusters as a clustering, it is possible to distinguish between hierarchical versus partitional, exclusive versus overlapping versus fuzzy, and complete versus partial types of clusterings. Moreover, broad categories of clustering methods have been identified by examining the main characteristics of the various algorithms. As reported by Tan et al (2006), most algorithms can be

described as a partitioning, hierarchical, density-based, grid-based or model-based method. Meanwhile, some clustering techniques are hybrid models integrating ideas from various methods and so are more difficult to categorise.

Partitional Clustering

In an overview of partitioning methods, Han & Kamber (2001) outline the core steps of the process as follows:

- the data (consisting of n objects) is initially divided into k groups, $k \leq n$
- iterative relocation improves the partitioning by moving objects between groups

Heuristic partitioning algorithms such as k-means and k-medoids operate in this way resulting in clusters of the data which must contain at least one object. The requirement that all objects must belong to at least one cluster can be relaxed somewhat in fuzzy k-means techniques. In certain instances, fuzzy k-means has shown to improve on the quality of the partitioning such as in the clustering of image data.

Despite their popularity, k-means and k-medoids are sensitive to the random selection of initial cluster centroids and can fall into local optimal solution (Jafar & Sivakumar, 2014). Furthermore, k-means and other partitional clustering techniques struggle with some challenges where dimension is the core concern (Ghorpade-Aher & Metre, 2014). In such a high dimensional setting, where X denotes an $n \times p$ matrix with n observations and p features, Witten & Tibshirani (2010) address the possibility of the underlying clusters differing only with respect to a small fraction of features.

The result of their research is a framework for sparse clustering enforcing a lasso-type penalty on the weight of each p which serves to indicate the contribution of that feature to the resulting sparse clustering. Compared with the standard k-means, the Witten & Tibshirani (2010) found their method to yield superior results on their examples of sparse datasets. The algorithm is argued to be most useful however when the number of features is much larger than the number of observations.

In 2012, Yondo et al conducted further investigation into the technique of Witten & Tibshirani which they admit does work cleverly to exploit the underlying patterns in

data and retrieve a very good partition. However, they discovered that this strong performance was not consistent across all sparse datasets examined and in fact it was possible that a small proportion of atypical observations may adversely impact the solutions found by Wittern & Tibshirani's (2010) algorithm. Yondo et al (2012) devised a robust sparse k-means algorithm which, as shown through simulation studies on microarray data for cancer patients, performs better than similar competing methods.

Kaufman and Rousseeuw (1990) developed a k-medoids algorithm designed specifically to work well with large, high dimensional datasets. Clustering Large Applications (CLARA) overcomes the limitations of previous k-medoids algorithms which are often computationally expensive and have large storage requirements. Kaufman and Rousseeuw (1990) outline their algorithm as follows:

- A random sample of the dataset is clustered into k subsets using k-medoids.
- k values representing the medoids for each cluster are obtained.
- The remaining objects in the dataset are then assigned to their nearest k-medoid.
- The quality of the clustering is measured using the average distance between each data point and their corresponding k- medoid.

Once this process of sampling and clustering has been carried out a predefined number of times, the sample which produces the lowest average distance is selected.

In further developments, researchers have looked to fortify partitioning algorithms by incorporating other model-based methods which hypothesize a model for each of the clusters and find the best fit of the data to the given model (Jafar & Sivakumar, 2014). Kuo et al (2002) combined self-organising feature maps and the k-means algorithm in order to enhance market segmentation, and later improved on this with a modification involving a genetic k-means algorithm (Kue et al, 2006).

Swarm Intelligence

Addressing the concern of the standard partitioning algorithm getting stuck at a local optimal solution, researchers have successfully applied evolutionary algorithms such as Particle Swarm Optimisation (PSO) and Ant Colony Optimisation (ACO) to clustering tasks (Jafar & Sivakumar, 2010; 2014). Such algorithms exemplify the

tools of Swarm Intelligence (SI), an innovative and intelligent paradigm for solving optimisation problems arising from the study of colonies, or swarms of organisms (Jafar & Sivakumar, 2010). More specifically, these authors describe ACO as a model which imitates the way in which real ants find the shortest route between two points.

In their initial examination of applications of ACO to clustering, Jafar & Sivakumar (2010) found the benefits to be its ability to automatically ascertain the ideal number of clusters, its linear scaling against the dimensionality of data and its robustness in avoiding distortion due to outliers. Mary & Raja (2009) effectively applied ACO as a means of refining data initially partitioned by the k-means algorithm with the adaptation of selecting the initial seeds based on statistical modes. This variation on the choice of starting points, along with the post processing refinement was seen to generate clusters of improved quality. Using ACO has also been shown to enhance the k-means algorithm when applied on high dimensional data (Aparna & Nair, 2014).

PSO is described by Jafar & Sivakumar (2010) as a stochastic optimisation approach modelled on the social behaviour of animals and the way in which these animals, i.e. particles, are grouped into a swarm. Ghorpade-Aher & Metre (2014) provide a comprehensive overview of the recent, yet extensive, research into the applicability of PSO variants for clustering multidimensional data. One such example is the combination of PSO and k-means as studied by Singh & Singh (2013). In this instance, the evolutionary algorithm is applied first to give the optimal solution for seed selection after which k-means operates as normal. It was found to be more efficient and produce results with increased accuracy compared to the k-means algorithm while also eliminating the need to pre-determine k.

Proportional Clustering

Highly relevant to this research paper is the task of clustering a matrix of proportional values where the entries across each row sum to one. Little evidence of this has been located in literature to date with the only example being a research project by IBM. Authors Kashima et al (2009) adapt the standard k-means algorithm by introducing constraints which limit all elements and existing cluster centres to values greater than or equal to zero. They also require that all entries in a row as well

as cluster centres sum to one. In a given iteration, once the data points are assigned to a cluster, new centroids are estimated taking into consideration these constraints.

Essentially the problem becomes an optimisation one, as the algorithm works to minimise the sum of distances from each data point in a cluster to the centroid. Kashima et al (2009) justify their use of the L1 distance, i.e. Manhattan norm, as it is known to be robust to noise. They also identify potential challenges of using this distance calculation such as the fact that no closed solution exists when adopting the L1 distance on a problem with proportional constraints. Without constraints, the median becomes the closed form solution using the L1 distance and so they suggest alternatives that would act as approximations, e.g. to use sample means as cluster centroids or compute dimension-wise medians and normalise them to make them sum to one.

Kashima et al (2009) found the originally proposed method to outperform the two alternatives on experiments on their test data. The proposed method resulted in moderately sparse clusters, somewhere in between the extremes produced by the other two, thus leading to more interpretable cluster centroids.

Hierarchical Clustering

This family of clustering algorithms group data objects into a tree of clusters. Hierarchical clustering can take an agglomerative, bottom-up approach or conversely operate in a divisive, top-down manner (Han & Kamber, 2001). The BIRCH method formalised by Zhang et al (1996) was the first of its kind to handle 'noise' and significantly improved the effectiveness of clustering on large datasets. Awarded the 2006 SIGMOD Test of Time Award, this effective technique greatly impacted future research (ACM SIGMOD, 2015). This algorithm exemplifies a merging of two types of clustering by firstly applying an agglomerative hierarchical technique and then performing a selected type of iterative partitioning (Zhang et al, 1996; Han & Kamber, 2001).

2.4.3 Applications to the Insurance Industry

Within the motor insurance industry, hierarchical clustering techniques have been utilised in order to assess the perceived risk of a policy holder. Segmenting customers according to risk has traditionally been achieved using heuristic methods

involving factors such as territory, demographic variables and others, e.g. driving experience and record (Yeo et al, 2001).

These authors investigate the difference between a heuristic method and a data-driven approach based on hierarchical clustering to separate a portfolio of motor insurance policy holders into various risk groups. Clustering allows for more variables be taken into consideration when segmenting the dataset, a feature Yeo et al (2001) identify as a significant advantage. Their experiment discovers the clustering model to be more effective in distinguishing high-risk groups from the low-risk ones. This is the case because a lower, but more interpretable, number of clusters are identified.

2.4.4 Classification

Kotsiatntis (2007) defines supervised machine learning algorithms as those which reason from a training dataset to produce general hypotheses which can subsequently predict outcomes for testing or unseen datasets. Supervised machine learning differs from unsupervised learning in that the dataset used in training have known class labels. As such classification is a form of supervised machine learning.

Classification in machine learning is the process of identifying and assigning the most appropriate class to which an object belongs from several predefined classes. It is described by Tan et al (2006) as "the task of learning a target function f that maps each attribute set x to one of the predefined class labels y." In general, the approach to classification involves splitting the dataset into training, validation (optional) and test sets.

There are various families of classification techniques, each with distinguishing characteristics. Many components should be considered when deciding on the most appropriate technique to adopt. Any decision made should reflect the objectives of the problem and be suitable for the application at hand (Tan et al, 2006).

The main categories of classification techniques as outlined by Kotsiantis (2007) are *Logic based algorithms* (Decision Tree & Rule Based), *Perceptron based techniques* (Multilayer Perceptrons), *Statistical learning algorithms* (Naive Bayes & Bayesian Network), *Instance based learning* (*k*-Nearest Neighbour) and *Support Vector Machines*.

2.4.5 Classification Methods

Decision Trees

Decision trees are hierarchical structures consisting of nodes and directed edges (Tan et al 2006). Every node provides a test on a variable in the dataset with the corresponding edges depicting the possible outcomes. The end nodes or leaf nodes in the decision tree each represent a class label to which an unseen object is then assigned.

This method is widely used due to its understand-ability and the transparency of its process of assigning labels to unseen data. Tan et al (2006) make the point that decision trees require no assumptions to be made on the probability distributions of the data. However Kotsiantis (2007) notes an assumption must be made that objects with different class labels have different values for at least one variable.

Rule Based

Rule based classifiers construct a set of 'if-then' rules from test data which are applied to unseen data to predict their class label. Rules can be extracted from decision trees but can also be produced directly using rule based algorithms where each rule in the set is represented in a disjunctive normal form (Kotsiantis, 2007).

As with decision tree induction, transparency is one of the main reasons rule based classifiers are chosen. However, a disadvantage of using a rule based technique for multinomial classification is the possibility of obtaining contradictory rules which may assign an object to more than one class (Kotsiantis, 2007).

Multilayer Perceptrons

Jain et al (1996) describe Artificial Neural Networks (ANN) as weight directed graphs whose nodes are artificial neurons and whose directed edges connect the neurons output and neuron inputs. Multilayer Perceptrons are among the simplest and most popular class of multilayer feed-forward networks. These are ANNs with complex structures which only allow signals to travel from input to output as argued by Kotsiantis (2007). The same author explains the process of multilayer perceptron classification where the input-output mapping of the network is determined by training on a set of class-labelled data. New data is then classified by passing it through the developed network.

Tan et al (2006) highlight that although training can be time consuming with ANNs, classification is independent of the sample size of the training set. Therefore classifying test data or unseen data is fast. Conversely, Kotsiantis (2007) argues that ANNs are often not used due to their inability to effectively communicate the reasons behind their output i.e. their lack of transparency.

Statistical Learning Algorithms

In classification, statistical learning techniques explore the probability that an object will belong to a certain class label, rather than merely outputting a classification. The Naïve Bayes classifier estimates the conditional probability of each attribute from the training data, given the class label. Classification is then carried out by computing the probability of a given class label for particular instances of each of the attributes. It then predicts the class with the highest probability after taking into account all the relevant evidence (Fiedman et al, 1997). This is carried out under the assumption that the attributes are conditionally independent for a given class label (Tan et al, 2006).

Bayesian networks are graphical representations of the joint probability distributions among a set of random variables. This technique of classification relaxes the assumption of independence in not requiring all attributes to be conditionally independent. An advantage of statistical learning algorithms is their ability to capture prior knowledge of a particular problem. However Kotsiantis (2007) notes they are unsuitable for datasets with many attributes.

K-Nearest Neighbour

In K-Nearest Neighbour (*k*-NN) classification an object is assigned to a class based on the majority class of *k* nearest neighbours to the object. *K*-NN is described as a lazy learning algorithm due to induction being delayed until run-time, i.e. until classification of an object is required (Cunningham & Delany, 2007). As such the *k*-NN classifier does not require model building however classifying a test example can be expensive (Tan et al, 2006). Another disadvantage of *k*-NN is that its performance is affected by the user's choice of *k* (Kotsiantis, 2007).

Support Vector Machines

In Support Vector Machine (SVM) classification, training data points are plotted whereby points in different classes are separated by a clear gap. Linear SVMs find

the hyper-plane which maximises the margin between two classes. The solution is then represented as a linear combination of the points which lie on the hyperplane, known as support vector points (Kotsiantis, 2007). Nonlinear SVMs are applied to datasets which have nonlinear decision boundaries. In this case the data is first transformed into a higher dimensional space in order for a linear decision boundary to be used to separate classes (Tan et al, 2006).

SVM is a binary classification technique however methods can be applied to deal with multinomial class problems. One of the main advantages of using SVM classifiers is their ability to find the global minimum rather than a local minimum as could be the case with artificial neural networks which employ greedy based search strategies (Tan et al, 2006).

2.4.6 Applications to the Insurance Industry

Both Viaene et al (2002) and Liu et al (2014) apply different classification techniques to insurance claims then compare the outcomes. Specifically, claims are classified in relation to fraud detection (Viaene et al, 2002) and claim duration (Liu et al, 2014).

Fraud Detection

Fraud detection is a binary classification problem where claims are classed as either fraudulent or non-fraudulent. Because of their focus on fraud detection, some of the techniques compared by Viaene et al (2002) are specifically formatted to deal with this binary classification. However, strategies can be easily adapted in the event of multinomial classifiers.

The study uses automobile insurance data and includes only variables that are known early in the claim life cycle. Methods of comparison centre on performance, i.e. how well claims are classified. The authors note that many other factors could be taken into account when choosing a technique including speed, user input, interpretability of the algorithm and also business requirements (Viaene et al, 2002). The classification methods compared cover some of the major types of supervised machine learning.

The authors find that, with the exception of the C4.5 decision tree whose performance is among the worst, all the algorithms show consistent performance in terms of predictive power with no one technique outperforming the rest. Focusing on

simplicity and efficiency, Viaene et al (2002) argue in favour of using the logit and linear kernel least-squares support vector machine methods, warning against the algorithms with high computational complexity, e.g. the TAN Bayes classifier.

Claim Duration

The comparison undertaken by Liu et al (2014) relates to this paper as it deals with multinomial classification techniques aiming to group claims according to their duration. It should be noted that income protection insurance data is used rather than automobile data. The techniques applied by Liu et al consist of Linear Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbour (*k*-NN), Log-logistic Regression and Ordered Log-logistic Regression. Variable selection methods or weighting methods are also used to highlight variables which have strong explanatory power in relation to claim durations. The backward elimination technique, principal components analysis and weighted combination models are applied to each classification technique and their results are compared.

Model performance is assessed by two loss functions; the first measuring misclassifications, the second also taking into account the severity of the misclassification. Liu et al (2014) find that in relation to the second loss function, the *k*-NN approach provides the best results across all variable selection methods, with the exception of the instance when all variables were used. In this case LDA is the only method which yields better results than *k*-NN.

Issues with the *k*-NN method include difficulties in interpreting the model and in assessing the importance of each variable when determining the class outcome. The authors suggest the application of principal components analysis or the weighted combination method to overcome these problems. In general, the performance of *k*-NN is sensitive to the dimensionality of the data so would not perform well for data with many attributes.

Overall, the results from Liu et al (2014) indicate the strong predictive ability of all the data mining methods applied. Their usefulness is highlighted in classifying claims ‘close’ to their true class, if not exactly in the correct class (for example in duration class 9 when in fact it should be duration class 8).

Significant Variables

In the research of Viaene et al (2002) predictors found to be relevant in the classification stage include accident, claimant, injury and age variables. Among others used are lags describing the beginning of the policy to the accident and from the accident to its reporting. Frees & Valdez (2008) find that when classifying claims according to their different types, significant determinants included vehicle characteristics, year and whether or not the vehicle was an automobile. For automobiles in particular, they found vehicle age was influential.

2.5 Summary

This chapter began by describing two actuarial methods commonly used to predict the overall claim costs an insurance company will be obliged to pay out in a given period of time: the Chain Ladder Method and the Bornhuetter-Ferguson Method. Previous research which has improved the prediction of claim outcomes through partitioning claims data was then briefly discussed. The outcome of such research provides the reasoning behind the clustering stage of the model, and justifies why we expect there to be clusters of claims in the data. After the basic concepts of both clustering and classification were outlined, a broad examination of the most prevalent techniques was presented for each. The aim here was to give an impartial description of how each method works and review their advantages and disadvantages in practice.

Chapter 3 - Data Exploration and Preparation

3.1 Overview

The purpose of this chapter is to describe the data provided to us by FTI Consulting Ltd and to outline the construction of the analytical base tables (ABTs) for the clustering and classification stages. Important findings from the initial exploratory data analysis are also highlighted.

The areas covered in this section are:

- Description of Data
- Data Cleaning & Pre-processing
- Data Summary

3.2 Description of Data

FTI Consulting Ltd provided two datasets made up of in-house data from a past client, a large insurance provider. The first dataset is made up of transactional data i.e. payments made or received in relation to an insurance claim. Each entry in this dataset represents a payment - a positive figure shows an outward payment and a negative figure describes an inward payment. Variables such as the date, time, amount and the claim ID which the payment relates to are included in this dataset. A full list describing the information available for each payment is shown in Appendix A.

The second dataset provides claim specific information. The rows correspond to individual claims and variables such as claim ID, date, the claim type as well as information regarding the policy holder is provided. A complete table describing the claim information variables is found in Appendix B.

According to FTI Consulting, the insurance company to which the data belongs began collected their data comprehensively in 2010. In 2014, new processes were put in place thus changing the structure of the data collected. For these reasons the data provided includes all claims which were reported and were subsequently settled within the period 2nd January 2010 and 3rd June 2014. Claims in the dataset are referred to as closed claims since they have been settled and are marked as 'finalised'.

The data consists of 28,387 individual claims with a total of 335,812 payments. As previously mentioned, all the data supplied comes from one insurance provider. All relate to motor insurance claims and 96% of claimants have comprehensive cover.

In some instances, there were inconsistencies in the data provided. Table 1 shows an example of a poor quality variable in the dataset. The insurer has documented in the data dictionary provided that all fire and theft claims should be labelled 'Not Applicable' for the 'Single Vehicle Accident' variable. In this instance however, 39.63% of 'Theft Recovered' claims and 33.04% of 'Theft Unrecovered' claims are mislabelled. It would be time consuming to completely eradicate all poor quality variables from the dataset and their presence reflects the reality of the data gathered by insurance providers. For these reasons, although certain outlying records were removed, the majority of claims with known errors were kept.

Class of Claim	**Single Vehicle Accident variable**			
	2/More vehicles involved	Don't Know	Not Applicable	Single Vehicle Accident
Theft recovered	43	8	760	448
Theft unrecovered	16	2	379	169

Table 1: Class of claim vs. single vehicle accident variables

3.3 Data Pre-Processing

3.3.1 Data Cleaning

Cleaning the data was necessary in order to deal with some of the noise occurring in the dataset. Initial data exploration highlighted a number of outliers and those which might affect the clustering in the proportional matrix were removed from both datasets.

Entries whose total claim cost amounted to zero (where the positive and negative transactions cancelled) were also removed as these would have created inconsistencies when constructing the proportional matrix.

Overall, 117 claims and their related payments were removed resulting in a workable dataset consisting of 28,270 claims.

3.3.2 Feature Construction

The new variables outlined in Table 2 were derived from the raw data in order to further describe each claim:

Variable Name	Variable Description
num_tp_involved	The number of third parties involved in a claim
num_tp_paid	The number of third parties who paid or received a payment in a claim
total_claim_cost	The total cost of a claim
month_reported	The month a claim was reported
claim_duration_full	The number of months the claim took to be settled
transaction_months	The number of months in which a transaction was made
num_days_loss_to_reported	The number of days between the incident occurring and being reported to the insurer
first_trans_date	The date of the first transaction in a claim
day_to_first_trans	The number of days between a claim being reported and the first transaction occurring
single_veh_dummy	A dummy variable indicating if a claim was an accident involving a single vehicle
multi_veh_dummy	A dummy variable indicating if a claim was an accident involving 2 or more vehicles
injury	A dummy vehicle indicating if a claim involved an injury (regardless of who was injured)

Table 2: Features derived from claim information dataset

3.3.3 Proportional Matrix Construction

In order to carry out the clustering stage, a matrix was created to describe the proportional payment life cycle of each claim. The construction of this matrix was carried out through Microsoft Excel. The sum of payments in each month was calculated for every claim, in addition to the total cost of the claim. So the claims may be comparable, months and years were then converted to monthly time intervals 1, 2, etc. where time interval 1 represents the total amount paid in the month the claim was reported, interval 2 the amount paid the month after and so on. The transaction amounts per time interval were then converted into proportions of the total claim cost.

The resulting matrix is as follows:

$$\begin{bmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & \ddots & \vdots \\ p_{N1} & \cdots & p_{NM} \end{bmatrix}$$

Rows represent a single claim while columns indicate monthly time intervals. Entries in the matrix are proportional values where each entry p_{ij}gives the proportion of claim i paid out in time interval j and each row sums to 1. The total number of rows N = 28,270 corresponds to the total number of claims. The total number of columns $M = 53$ shows that the longest claim in the dataset lasted for 53 months.

The claim IDs were inputted as row names so they would correspond with those in the claim information file. The matrix is a large $28270\ X\ 53$ sparse matrix, where roughly 96% of cells are made up of zeros. This reflects the generally short duration of motor insurance claims. Although a small number of claims are drawn out over many months or years, the average duration of claims in the dataset is 4.73 months, leading to a large number of zeros cells in the matrix.

Appendix C Shows a screenshot of the file as it appears when read into R. Packages designed to store such matrices are available in R, however most clustering and classification packages require the data to be entered as a data frame or data matrix. In order to maximise the number of machine learning techniques available for use, methods for storing sparse matrices were not employed.

3.4 Data Summary

Exploratory data analysis (EDA) was carried out on both the proportional data and the claim information data. This section discusses the key findings from the EDA.

3.4.1 Claim Duration

Comprehensive information regarding the motor insurance market in Ireland is limited thus gaining an overview of the industry as a whole proved difficult. According to figures from the Central Bank of Ireland (2012) between 2001 and 2006, roughly 74% of claim costs with comprehensive cover are settled within four years of the accident.

It could be said therefore that the claims in our dataset originating in 2010 – which could last up to 4½ years – represent approximately three quarters of the claims reported in that year. Thus for the purposes of this study the clustering and classification stages were carried out using only the 2010 data. Claims reported after 2010 were seen as an inaccurate and unbalanced reflection of the true life cycle of claims reported in a given year, for reasons discussed in this section.

Figure 1 plots each claim according to the date they were reported and their duration. Since all observations are closed claims, the maximum possible duration for those reported in May 2014 for example, is only two months, as the data only includes information up until June 2014. It is possible however for those reported in January 2010 to last up 54 months. Although Figure 2 shows the median claim duration for each of 2010, 2011, 2012 and 2013 is three months, modelling using the entire dataset would over represent shorter claims and under represent the longer claims.

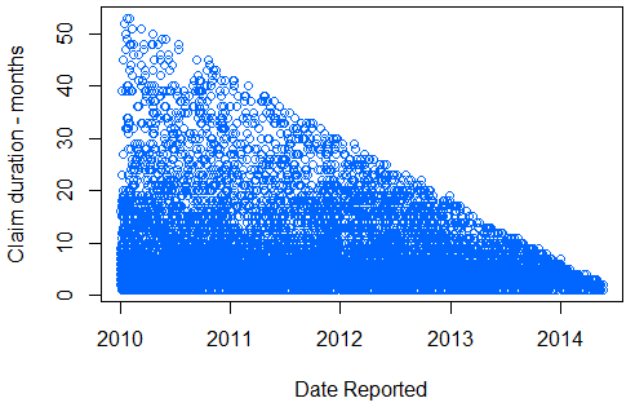

Figure 1: Plot of claim duration vs. date reported

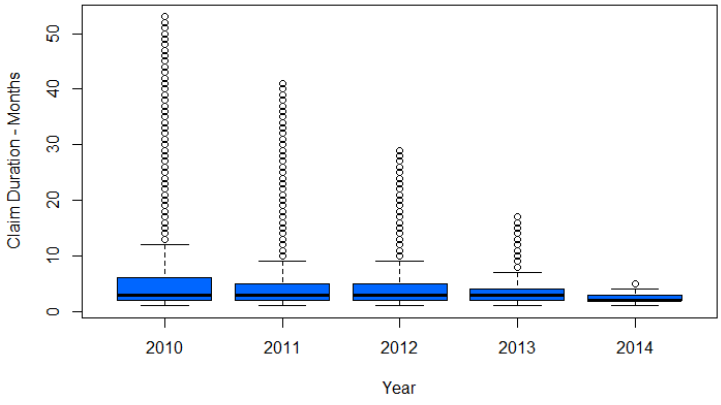

Figure 2: Boxplot of claim duration per year

3.4.2 Seasonality of Claims

Looking at the seasonality of claims year on year, it is apparent from Figure 3 that the winter months from November to January generally see a higher proportion reported and July to September sees the lowest proportion of claims reported. Claims reported in 2014 do not span the entire year thus were excluded from comparison. Using data only from 2010 runs the risk of capturing any abnormal behaviour present that year. Figure 3 however shows that monthly proportions reported in 2010 roughly reflect the average over four years in all months except October. The average proportion reported in this month however is skewed by the abnormally high number in October 2011. This was due to extreme rainfall and flooding particularly in Dublin which according to Met Éireann (Monthly Weather Bulletin, 2011) saw its highest October rainfall since 2002, almost half of which fell within a few hours.

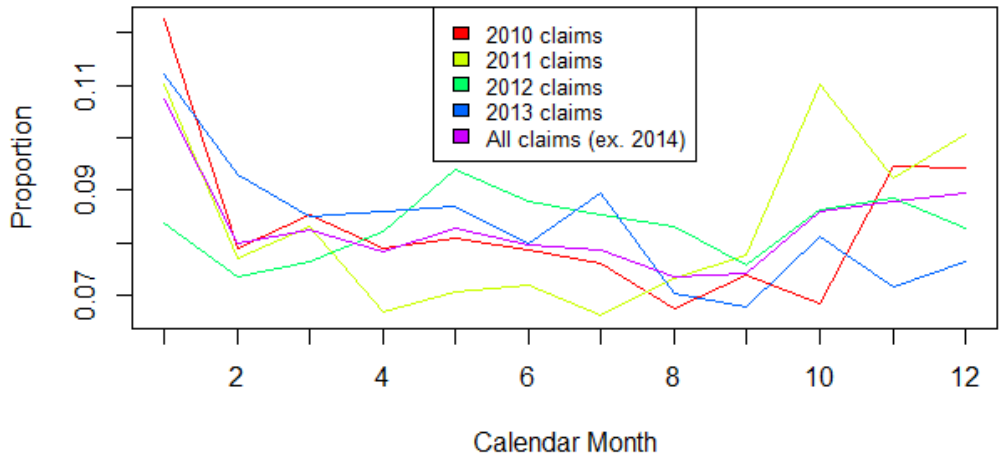

Figure 3: Proportion of claims reported per calendar month

3.4.3 Total Claim Cost

The Central Bank of Ireland report (2012) shows the average cost per claim with comprehensive cover between 2003 and 2009 fluctuated between €4,000 and €5,000. This is reflected in the data provided whose average total cost is €4,351. Figure 4 plots each claim according to the date reported and their total cost. Although 96% of claims are less that €20,000, data from 2010 captures high cost claims which are of great importance to the insurer and should be included in the modelling. As with the claim durations, using the entire dataset for modelling would over represent the average or low cost claims and underrepresent the high cost claims.

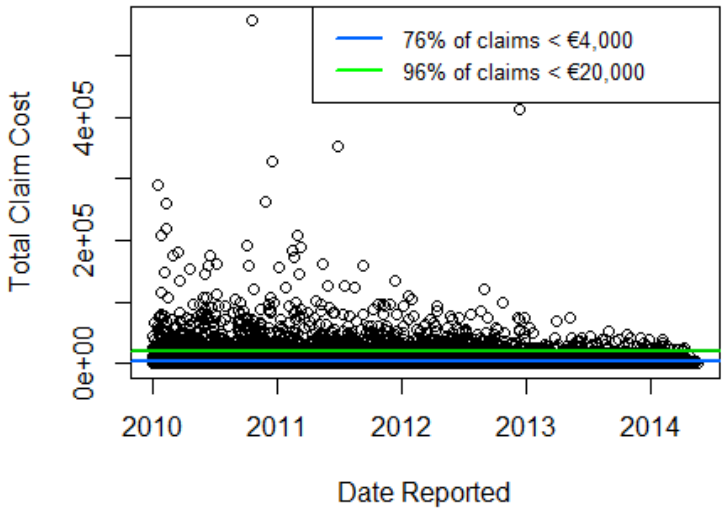

Figure 4: Plot of total claim cost vs. date reported for all observations

3.4.4 Proportions

Looking at the proportional matrix, according to Figure 5 on average around 43% of the total claim cost is paid out in the second month after the claim is reported. The distribution of the 2010 proportional payments closely follows that of the entire dataset.

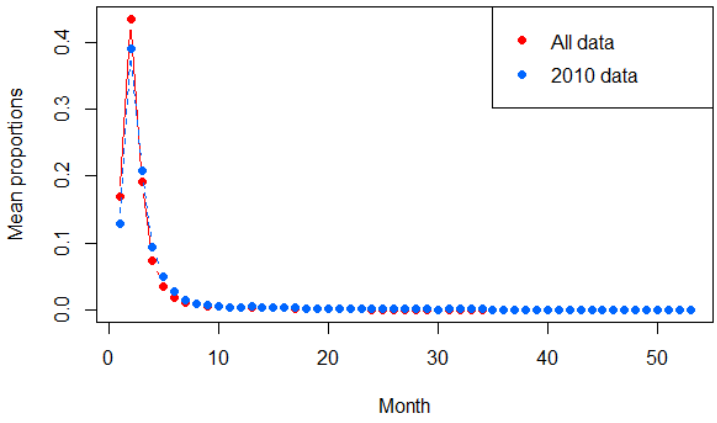

Figure 5: Average monthly proportional payments per claim

A plot of all of the 2010 claims in Figure 6 shows the majority of payments were made within the first six or seven months of a claim being reported with significantly less payments made as time goes on. A number of outliers are highlighted in this plot, those with proportions greater than 1 and less than 0. On further inspection, the corresponding payments were not erroneous thus we saw no reason to exclude them from the modelling stages. If a payment, x was made with regards to a claim in one month but a partial repayment, y was made in a subsequent month, the total claim

cost would be $x - y$. The first payment will therefore be greater than the total cost and thus its proportional value will be great than 1. Similarly, the repayment is presented as a negative value thus its proportion would also be negative.

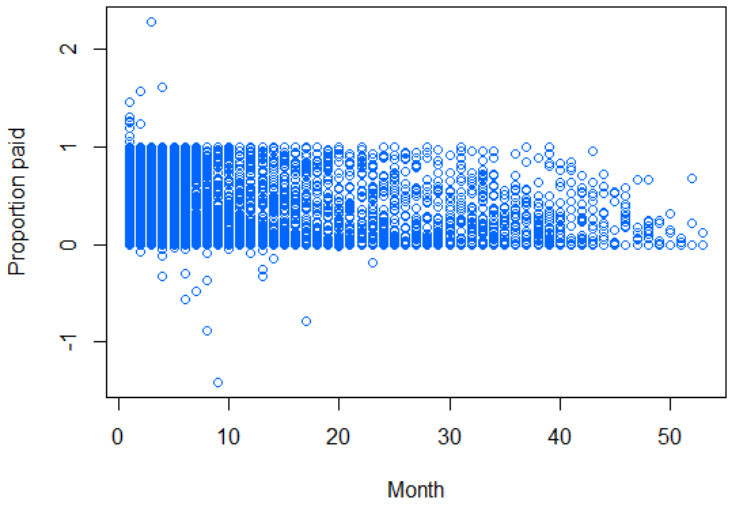

Figure 6: 2010 proportional data over time

No information was found in relation to claim-by-claim proportional payments in the Irish market or indeed any insurance market. There is no published evidence that research has been carried out on the proportion of individual claims paid at a given time interval. As a result, there was no benchmark on which to compare the proportional matrix derived from the data provided.

3.5 Summary

This chapter outlined the data provided by FTI and defined the cleaning and preparation processes which were required to produce a workable dataset for the subsequent modelling stages. A comprehensive description surrounding the construction of the proportional matrix was given so that is may be easily reproduced. The EDA in relation to claim durations, seasonality, total costs and proportional payments in the dataset was then documented. This provides an overview of the distribution of the underlying data.

Chapter 4 - Methodology

4.1 Overview

This chapter describes the steps taken in carrying out our investigation. With the overall aim being to establish a method of predicting the life cycle of a motor insurance claim, the work initially focuses on the payment patterns of closed claims as presented in proportional matrix form. Clustering techniques are applied to this matrix in order to uncover natural groupings in the data. These clusters are then analysed in detail with the purpose of profiling the clusters.

Most significantly, monthly proportional payments are examined for claims within each cluster in order to establish a predictor which describes the general behaviour of the cluster. Finally, a classification model is developed based on the original clusters but using FNOL information only. The final model is one which uses this classifier to assign a newly reported claim to a particular cluster. Consequently, predictions relating to the life cycle of this new claim can be made.

4.2 Software

All stages described in this chapter were conducted using R, "a language and environment for statistical computing and graphics" (R Core Team, 2015). The version of R installed on our machines was R 3.2.0 for 64bit Windows. R can be extended by importing *packages* and a number of these were utilised at various points of the data mining process in addition to those supplied within the standard R distribution. R functions within these packages will hereafter be referred to as *function* {package name} while packages mentioned in isolation are denoted {package name}. For a full list of the imported R packages utilised in the course of this research see Appendix D.

4.3 Clustering

Clustering techniques were applied in order to identify common paths taken by claims. Having selected a representative subset of the entire data from which to learn the clusters, i.e. 2010 data, the primary task of this phase was to determine the most effective clustering model.

4.3.1 Clustering model

Experimentation with a variety of clustering techniques was conducted. This range of techniques, some of which specifically dealt with sparse matrices, included partitioning (k-means, k-medoids), hierarchical (agglomerative, divisive), density based, model based and neural network methods (Self-Organising Maps (SOM), Self-Organising Trees (SOTA)).

From these we identified four methods that were performing to a satisfactory level:

- K-means found in the {stats} package
- K-Medoids using the CLARA algorithm for clustering large applications found in the {cluster} package
- Agglomerative hierarchical clustering, with use of the Ward aggregation method, found in the {stats} package and optimised by adding the packages {flashClust} and {fastcluster}
- SOTA a self-organising tree algorithm which combines hierarchical clustering and SOM found in the {clValid} package

4.3.2 Selection Measures

Further investigation was then conducted in order to determine both the optimal technique and number of clusters (k). This included calculation of the Davies-Bouldin Index, analysis of the silhouette coefficient and high level cluster profiling following the application of each of the four selected routines. We also set a lower bound on the cluster size deciding on a cut-off point of 10% of the total sample size. Basing this on the claims used for clustering, of which there are 7,985, the threshold was 798. This was not set as a hard bound on cluster size but rather an approximate heuristic measure to ensure sufficient data was available for the later classification stage.

Davies-Bouldin Index

This index quantifies the average ratio of within-cluster scatter to between-cluster separation (Charrad et al, 2014). When calculated for a range of k, that which minimises the Davies-Bouldin (DB) index is regarded as specifying the number of clusters.

It is calculated using the formula:

$$DB(q) = \frac{1}{q}\sum_{k=1}^{q} \max_{k \neq l}\left(\frac{\delta_k + \delta_l}{d_{kl}}\right),$$

Equation 1: Davies-Bouldin index

where

- $k, l = 1, \ldots, q =$ cluster number
- $d_{kl} =$ distance between centroids of clusters k and l
- $\delta_k =$ dispersion measure of a cluster k: standard deviation of the distance of observations in cluster k to the centroid of that cluster.

In calculating the Davies-Bouldin index for the clustering techniques under consideration, we applied the *NbClust* function in the package of the same name to the proportional matrix of 2010 claims. For all cases, the distance metric specified was Euclidean.

Silhouette

As a means of assessing clusters, Tan et al (2006) identify the use of silhouette coefficients as a popular measure as it combines the notions of cohesion and separation, i.e. the compactness and isolation of clusters. In the seminal paper on the topic, Rousseeuw (1987) highlights the usefulness of this metric in evaluating cluster validity and selecting an appropriate number of clusters.

The steps involved in the computation of the silhouette coefficient for an individual object i are as follows (Rosseeuw, 1987; Tan et al, 2006):

- Calculate $a(i)$ the average dissimilarity of i to all other objects in cluster A, the cluster to which it has been assigned (such a measure of dissimilarity would be the Minowski distance between objects)
- Consider any cluster C, different from A, and compute $d(i, C) =$ average dissimilarity of i to all objects in C
- Find the minimum $d(i, C)$ with respect to all clusters C calling this value $b(i)$, i.e. the distance between i and its closest point in a neighbouring cluster to which i does not belong

- For this i^{th} object, the silhouette coefficient is as follows:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

Equation 2: Silhouette coefficient

For each object, the silhouette coefficient takes a value within the range [-1, 1], the meaning of which is explained by Maechler & Rosseeuw et al (2015): higher values indicate observations that are well clustered, a small s_i (around 0) means that the observation lies between two clusters, while observations with a negative $s_i$are likely to have been placed in the incorrect cluster. By calculating the average silhouette width of all objects within each cluster, this method allows us to distinguish 'clear-cut' from 'weak' clusters. Effectively we can conclude that clusters with a larger average silhouette width are more pronounced (Rosseeuw, 1987).

For our research, we first transformed the original proportional payments data into a dissimilarity matrix based on Euclidean distance, employing the *dist* {stats} function to do so. We then calculated both the overall mean silhouette and the individual cluster averages for each of the four clustering methods listed in Section 4.3.1 with k varying from 4 to 9 inclusive. This upper limit of k was chosen because higher values would result in clusters that were too small. The lower limit of 4 was fixed since initial investigation with smaller values of k showed poor performance and did not merit further analysis.

Subsequently, we ranked each clustering method under consideration in order of its Davies-Bouldin index and mean overall silhouette. These rankings were summed to determine the quality of the algorithm from a broader perspective and hence give an initial indication of the most suitable clustering algorithm. The results of this ranking can be seen in Section 5.2.1, and the clusterings from the best performing techniques were examined. Any methods which violated the cluster size limit were dismissed.

4.3.3 Cluster Validation

Validation is a hugely significant step in cluster analysis as it ensures that patterns found are truly meaningful. By nature, clustering methods tend to generate clusterings even for quite homogeneous datasets (Hennig, 2007). Important aspects to consider are cluster consistency and stability.

Consistency

With the purpose of distinguishing between similarly performing techniques, we ran the same algorithms on the whole dataset of 27,290 observations and compared the clusters it assigned to the 2010 data (clustering X) with the labels given by the earlier clustering (clustering Y). This comparison was quantified using the Rand index, evaluated between 0 and 1 with 1 indicating a perfect match between clusterings. This measure considers how each pair of data points is assigned in each clustering, explains Rand (1971) after whom the index is named. Hennig (2007) also highlights the popularity of the adjusted Rand index which rescales the index and takes into account the occurrence of objects occupying the same clusters as down to random chance. We calculated the Rand and adjusted Rand values using the functions *rand.index*{fossil} and *adj.rand.index*{fossil} thus measuring the similarity between clusterings X and Y for the top techniques.

Stability

We assessed the stability by computing the Jaccard coefficient for each cluster, the calculation of which yields a value between 0 and 1. This cluster-wise measure of within-cluster similarity is based on set memberships and signifies, for two sets A and B, the number of observations in both sets divided by the number of observations in either, i.e.

$$\frac{A \cap B}{A \cup B}$$

A finite set, to which the clustering algorithm is then applied, is sampled from the existing clustered data and the Jaccard similarities of the original clusters to the most similar clusters in the sampled dataset are calculated (Hennig, 2007). The data is then resampled using a specified method, a replication of 50 is considered good. The mean similarity values are then computed to give the Jaccard coefficient for each cluster. To label a cluster as being “highly stable”, it should produce a Jaccard value of 0.85 or above, with 0.75 or more indicating a generally valid, stable cluster. Hennig (2014) goes on to explain that values between 0.6 and 0.75 can indicate patterns in the data, however below this clusters should not be trusted. To compute the Jaccard coefficients we implemented *clusterboot*{fpc} specifying bootstrap resampling with 100 replications in its operation.

Cluster Profiling

The best performing algorithm was then chosen and clusters were established. From this, we assigned a cluster label to each claim which formed the basis for subsequent phases of our investigation. We also undertook a period of analysis in order to profile the clusters and ascertain the characteristics of each. This involved the generation of summary statistics, visualisations and cluster exploration.

4.4 Classification

4.4.1 Overview

Having completed the clustering stage, different classification techniques were applied with the aim of predicting the correct cluster label for each claim. This section describes three stages involved in the classification process:

- Feature selection
- Implementation of different methods
- Classifier evaluation

4.4.2 Feature Selection

Feature selection is the process of selecting the most informative variables in a dataset which are then used in modelling. It is an essential procedure in building a classifier. For large datasets, feature selection has the ability to dramatically reduce the dimensionality of data thus increasing computational efficiency. The inclusion of irrelevant, redundant or noisy variables can also negatively affect a classifiers predictive ability (Moni-Sushma-Deep & Srinivasu, 2014).

Much research has been carried out regarding feature selection for classification resulting in a wide range of available methods. In this paper, both the Gini Index and the Boruta algorithm (Kursa & Rudnicki, 2010) were chosen as suitable measures to distinguish the most relevant variables in the dataset.

Boruta Feature Selection

The *Boruta*{Boruta} algorithm in R can be described as an all relevant feature selection process, whereby all strongly relevant and all weakly relevant variables are identified (Rudnicki et al, 2015). Kursa & Rudnicki (2010) argue that this kind of

selection is required to completely understand the mechanisms of the subject of interest rather than just the variables necessary to predict the target class.

With the Boruta algorithm, the dataset is extended by adding 'shadow' variables which are random by design. A random forest classification is then applied to the extended dataset and the importance of each attribute is evaluated by calculating a Z score. The maximum Z score among the shadow attributes is used as a benchmark and any variable with an importance considerably lower than this value is deemed unimportant. Conversely, an attribute with an importance higher than this Z score is considered relevant. To ensure statistically significant results, these steps are repeated until either an importance is assigned to all variables or the limit of random forest runs is reached (Kursa & Rudnicki, 2010).

The Boruta algorithm was executed on all variables in the 2010 claim information dataset which were unique and relevant to the incident. All default parameters were adopted.

Gini

The Gini Index was first used to describe the distribution of income in an economy. In that sense, a Gini of zero would indicate that income is evenly distributed throughout the population however a value of one would suggest only one person is receiving 100% of the country's income. According to Shang et al (2007) the Gini has more recently been applied to the area of feature selection and has been used extensively in decision tree learning as a method for selecting splitting attributes.

When interpreting the Gini Index in terms of feature selection, a variable with a high value indicates an inequality among the variable categories. This in turn suggests the variable may contribute significantly to the classification model.

The Gini Index was computed using *ineq*{ineq} for the unique and relevant variables in the 2010 claim information dataset.

The final set of attributes used in classification was chosen through a combination of iterative testing guided by the results from both the Boruta algorithm and the Gini Index. Specifically, each attribute was ranked according to how it performed in both measures and the most important were selected.

4.4.3 Implementation

Tan et al (2009) argue that a good classification model should both fit the training set used to build it as well as correctly predict the class of previously unseen data. To achieve this balance, the 2010 dataset was randomly split using the *sample*{base} function in R into a training set with 60% of observations and a test set with 40%. Claims reported in 2011 were also used as a second test set using the cluster labels resulting from clustering on the entire dataset. This was done in an attempt to measure the performance of a classifier on data which was previously unused in model development.

As reported in Section 5.4.1, clustering on the 2010 dataset produced an uneven number of observations in each cluster. This is known as the "Class Imbalance Problem" and is a common issue in data mining. According to Wang & Yao (2012), a skewed distribution can often make standard classifiers less effective, particularly when predicting minority classes. To address this problem, a balanced training set was constructed. This training set was created by oversampling the data using the *ddply*{plyr} function and consisted of an equal number of observations from each cluster.

With each training set, a number of classification methods were applied: Decision Trees, Random Forest, K-NN, ANN, SVM and Naïve Bayes. The four best performing techniques were analysed in greater detail and tested by changing available parameters. These four techniques were as follows:

- K-NN implemented using *kknn* in the {kknn} package
- ANN implemented using *nnet* in the {nnet} package
- SVM implemented using *svm* in the {e1071} package
- Naïve Bayes implanted using *naiveBayes* in the {e1071} package

4.4.4 Measurement

The results for each model were tabulated in a confusion matrix whereby every entry x_{ij} represents the number of observations in class i predicted to be in class j (Tan et al, 2006).

With multi-class classification, both an overall accuracy figure and an accuracy measure described by Sokolova & Lapalme (2009) were calculated from the confusion matrix. The first measure commonly used in classification and was calculated using Equation 3:

$$\frac{\sum_{i=1}^{l} x_{ii}}{\sum_{i,j=1}^{l} x_{ij}}$$

Equation 3: Overall accuracy

Where l denotes the cluster number

x_{ij} describes the number of observations actually in cluster i predicted to be in cluster j.

Yang et al (2015) argue however that this measure of overall accuracy becomes inappropriate or insufficient when there is a class imbalance issue, largely due to the dominating effect of larger classes. For this reason, the Sokolova & Lapalme (2009) multi-class accuracy which computes the average class accuracy was also used to evaluate the classifiers. Their accuracy measure removes the bias of large clusters and treats all classes equally. This method represents correctly classified observations for each class i and observations not in class i which were predicted not to be in class i, regardless of the class in which they were placed. The resulting value represents the average per-class effectiveness of a classifier (Sokolova & Lapalme, 2009).

$$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

Equation 4: Multi-class accuracy (Sokolova & Lapalme, 2009)

Where l denotes the cluster number

tp_i describes the number of observations correctly classed in cluster i

tn_i describes the number of observations not in class i which should not be in class i

fp_i describes the number of observations incorrectly classed in cluster i

fn_i describes the number of observations not in class i which should be in class i

A baseline accuracy figure was calculated which describes a scenario whereby every claim was classed into the largest cluster. This figure was then used as a benchmark from which to evaluate the performance of each classification technique.

4.5 Model Testing

Testing was carried out using claims data from 2012 and 2013 as it was entirely unseen by the model up to this point. In order to evaluate its overall predictive power, we applied the classifier to this test data. Once each claim was assigned to a cluster, the proportion of the total claim cost to be paid in each month was predicted. These predictions were made using the average monthly proportions of the appropriate cluster, which had been calculated using the clustering results from the 2010 dataset. We then measured the quality of the model by calculating the Pearson correlation coefficient between the true proportional payments and those predicted for the 2012-13 data. As a further indication of accurateness we measured the Euclidean distance between the matrices containing these values. The calculations were carried out using *cor*{stats} and *dist*{stats} respectively.

4.6 Summary

This chapter outlined the key stages of the project with reference to the tools and techniques employed. In describing the clustering phase, we detailed the algorithms considered and the measures of cluster quality and stability examined. We covered the classification stage of the research by outlining the factors involved in feature selection and the classification algorithms implemented thereafter. Finally, we referred to the manner in which the final model was constructed and its effectiveness tested.

Chapter 5 - Clustering Results

5.1 Overview

The chapter deals with the algorithms and performance measures examined during the clustering phase of our research. We report the Davies-Bouldin index and mean silhouette for the initial 24 clustering options under consideration. We also describe their contribution to the model selection process and that of additional validation measures, namely the Rand index and the Jaccard coefficient. The results of the chosen technique are detailed and the clusters formed for the 2010 dataset are profiled.

5.2 Model Selection

5.2.1 Selection Measures

The Davies-Bouldin index shows that overall CLARA performs consistently strongest for our range of k (Figure 7). The exact values for both the Davies-Bouldin index and overall mean silhouette coefficient, along with the associated rankings for each technique are given in Appendix E.

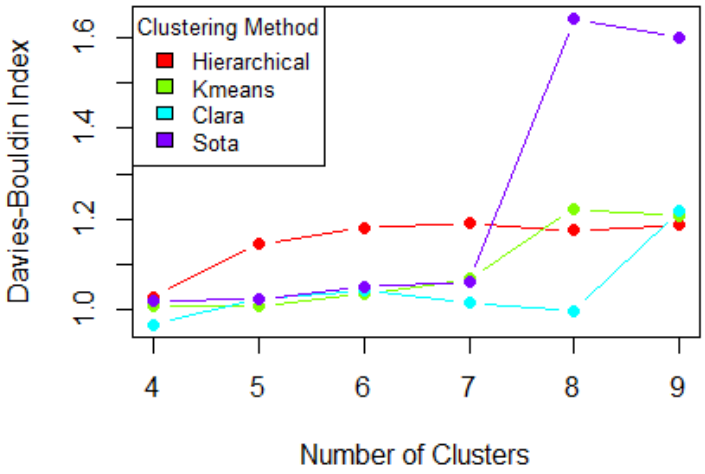

Figure 7: Plot of Davies-Bouldin results from clustering on 2010 data

5.2.2 Cluster Validation

To further supplement our decision-making process, the consistency and stability of the top 12 ranked methods were examined. These 12 represent 50% of the options under consideration; the remaining 50% were dismissed at this point. For each

method the Rand index and Jaccard coefficient was calculated. These quality measures, along with other factors, resulted in a number of the clustering methods being deemed of unacceptable quality, the reasons for which are summarised in Table 3. Recalling that Jaccard coefficients lower than 0.75 indicate increasingly poorer stability gives us an approximate on what is an appropriate requisite. We wish to avoid clusters that are too small, techniques that do not perform consistently well over different datasets and methods which result in negative mean silhouette values for individual clusters indicating there are many incorrectly placed observations.

Overall Rank	**Clustering**	**k**	**Issues with Clustering**
1	CLARA	8	Small clusters : $n = 117, 224, 398$
2	SOTA	7	Computationally expensive Difficult to directly compare with other methods: • Could not find JC due to memory limitations • Only possible silhouette generated from best sub sample Small clusters : $n = 217, 397$
3	CLARA	7	1 cluster silhouette mean < 0 2 unstable clusters indicated by JC values 0.64 and 0.68 Extremely poor JC values when clustering on all data Small clusters: $n = 228, 393$
4	Kmeans	6	2 unstable clusters indicated by JC values 0.5 and 0.69 Small cluster: $n = 401$
5	Kmeans	5	2 unstable clusters indicated by JC values 0.3 and 0.58
6	Hierarchical	9	Small clusters: $n = 80, 103, 223, 418$
7	CLARA	6	2 unstable clusters indicated by JC values 0.29 and 0.53 Small cluster: $n = 404$
8	Kmeans	7	2 unstable clusters indicated by JC values 0.08 and 0.56 Small clusters: $n = 125, 397$
9	Hierarchical	8	Small clusters: $n =$103, 223, 418
10	CLARA	4	1 unstable cluster indicated by JC value 0.52
11	CLARA	5	Performed well (results given Section 5.3)
12	SOTA	6	As above in SOTA $k = 7$ Small cluster: $n = 397$
			$n =$ the number of observations in the cluster

Table 3: Summary of reasons for rejecting clustering techniques

5.3 CLARA: k-medoids clustering

5.3.1 CLARA Performance

CLARA splitting the data into five clusters proved the most effective and reliable clustering algorithm. Here we detail the results of the silhouette analysis and validation tests carried out which highlights this methods consistency and stability.

Silhouette

Cluster	1	2	3	4	5
Mean Silhouette Width	0.7327	0.8242	0.0785	0.7812	0.8059

Table 4: Mean silhouette width for each cluster

The overall mean silhouette for the entire dataset is 0.6743. When measuring the mean per-cluster silhouette values, Cluster 3 performs most poorly. At 0.0785 this cluster contains claims that are very close to another cluster or which have in fact been misclassified. Clusters 2 and 5 are the most isolated and cohesive while the remaining clusters are still regarded strong and pronounced. The silhouette plot in Figure 8 represents a random subset of claims used in the CLARA algorithm from which these cluster traits are visible.[1]

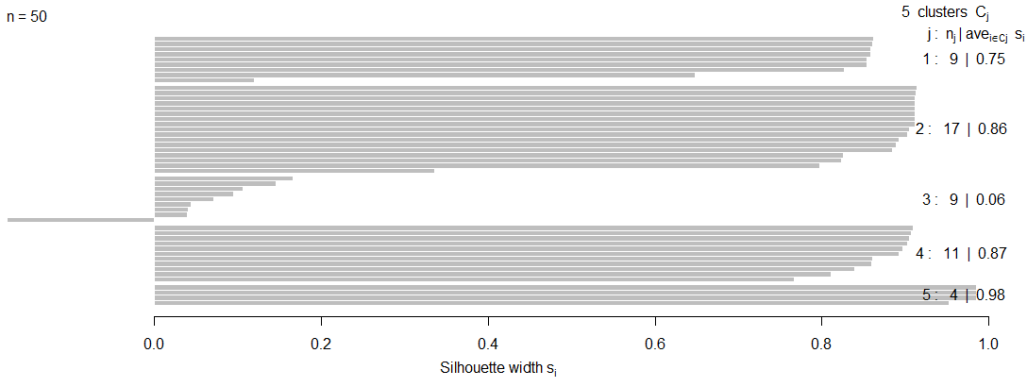

Figure 8: Silhouette plot of best claims in each cluster

Consistency

In comparing the cluster labels assigned to the 2010 claims from both clustering on these claims only and clustering on the whole dataset, the confusion matrix in Table

[1] This is done because plotting the entire dataset would become too large to be useful and require too much computational effort (Kaufman & Rousseeuw, 1990). Hence, cluster-wise silhouette averages on the plot differ to those shown in Table 4 whose computation used all observations.

5 shows the extremely high correspondence. Quantifying this similarity using the Rand index gives a value of 99.5% indicating an almost exact match. The adjusted Rand further verifies the similarity of the clusters as it remains high at 98.8%.

		Labels assigned from clustering on full dataset				
		1	**2**	**3**	**4**	**5**
Labels assigned from clustering on 2010 data	**1**	774	0	3	0	0
	2	0	3120	22	1	0
	3	0	0	1439	1	0
	4	0	0	5	1749	0
	5	0	1	4	0	866

Table 5: Confusion matrix of 2010 clustering vs. all clustering

Stability

Cluster stability proved to be a defining factor in choosing the best technique. Running *clusterboot*{fpc} on the clustered 2010 data gave the measures in Table 6:

Cluster	**1**	**2**	**3**	**4**	**5**
Jaccard Similarity	0.83477	0.94468	0.70706	0.96581	0.95493

Table 6: Jaccard similarity measures per cluster

Having a Jaccard similarity above 0.85 indicates that clusters labelled 2, 4, and 5 are highly stable. Despite falling below this threshold cluster 1 remains valid as it has a Jaccard coefficient of above 0.75. While cluster 3 has a measure just outside this range, we deem this to be acceptable having investigated further and acknowledged the diverse nature of observations within this cluster which contains a high proportion of the longer and more complex claims.

5.4 Cluster Profiling

5.4.1 Overview of Clusters

Table 7 and Figure 9 show the breakdown of the 2010 dataset into the five clusters. One cluster is somewhat larger than the others with the size ratio of all five being approximately 4:2:2:1:1.

Cluster	**1**	**2**	**3**	**4**	**5**	**Overall**
Size	777	3143	1440	1754	871	7985

Table 7: Number of observations in each cluster

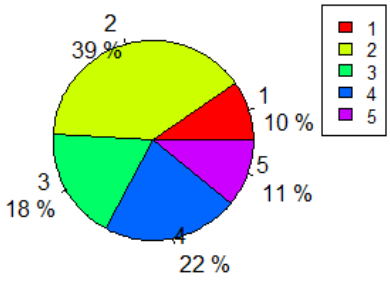

Figure 9: Cluster sizes

By honing in on proportional data from the first five months of all claims, Figure 10 further illustrates cluster behaviour. It not only reiterates the peak payment months for each cluster, but also affords a snapshot of how the algorithm is partitioning the claims for the first five time intervals; X1… X5. There is a clear distinction between clusters in most cases in terms of the proportional values of claims for the two components represented on each subplot.

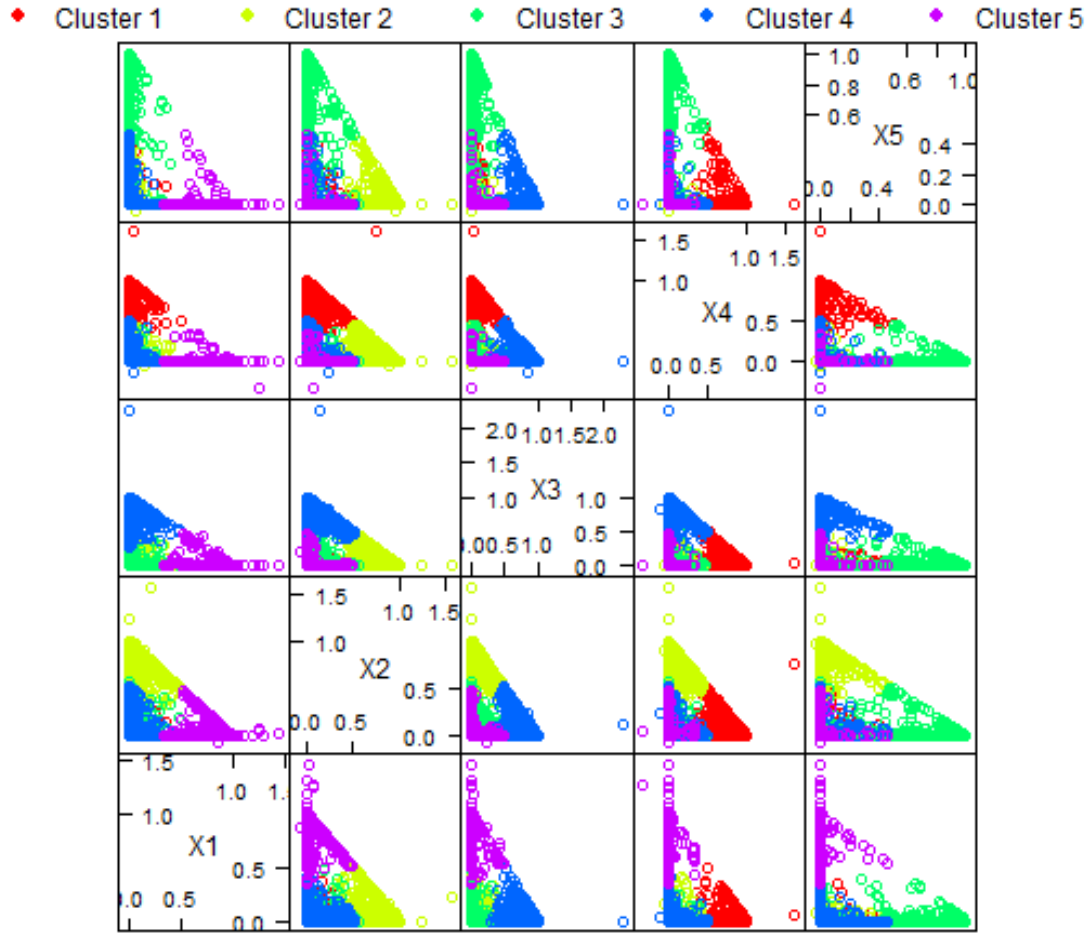

Figure 10: Scatter plot matrix showing proportional entry for time intervals 1-5

5.4.2 Proportional data

Table 8 reports the mean values for the first five monthly intervals calculated from the claims within each cluster, all of which are contained in Appendix F.[2] The proportion of claims which lie in clusters 5, 2, 4 and 1 peak significantly in months 1, 2, 3 and 4 respectively (Figure 11). The payment in this 'peak' month represents a minimum of an 84.3% settlement of the total claim cost as is the case for Cluster 1. In all of these clusters, the proportion of the total claim cost paid out in the first four months after a claim has been reported is at least 95.3%.

	Time Period (Month)				
	1	2	3	4	5
Cluster 1	0.022	0.059	0.029	*0.843*	0.013
Cluster 2	0.042	*0.907*	0.02	0.012	0.004
Cluster 3	0.019	0.051	0.024	0.015	*0.249*
Cluster 4	0.02	0.067	*0.869*	0.016	0.009
Cluster 5	*0.937*	0.029	0.013	0.005	0.004

Table 8: Average proportion per cluster for months 1-5

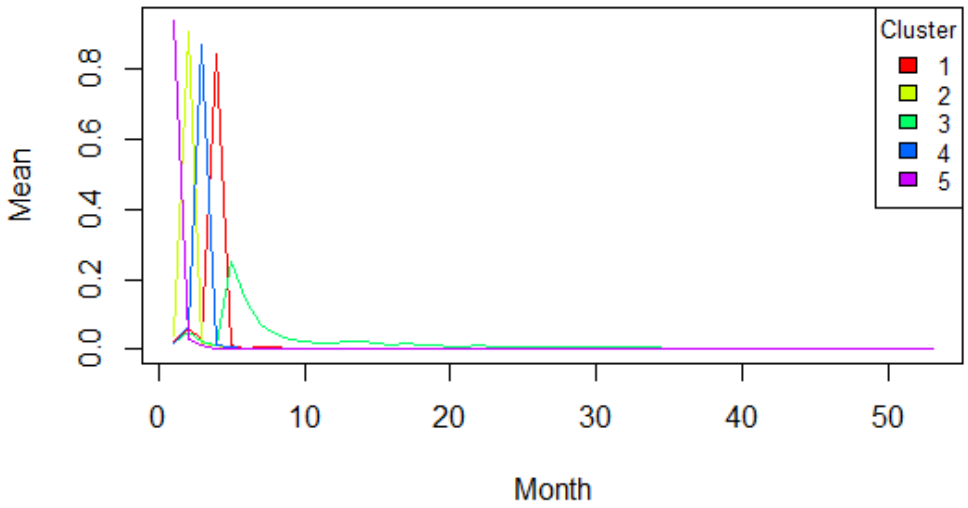

Figure 11: Plot of mean monthly proportions per cluster

Cluster 3 on the other hand does not experience such peaks. Month 5 sees the highest payment rate which is on average only 24.9%. Beyond this month however, while other clusters see little or no transactions, Cluster 3 has an above-zero average in all except four of the monthly time periods. Cluster 3 contains claims where payments

[2] These are the proportional values used in the predictive model to estimate the life cycle of a claim, after its cluster is predicted.

are generally made over a longer time period than those in other clusters. The variety of claims in this cluster is dramatically emphasised in Figure 12 which graphs the proportion paid out during all time intervals for each individual claim. The majority of the total cost is seen to be paid in a vast range of monthly intervals with no particular pattern emerging.

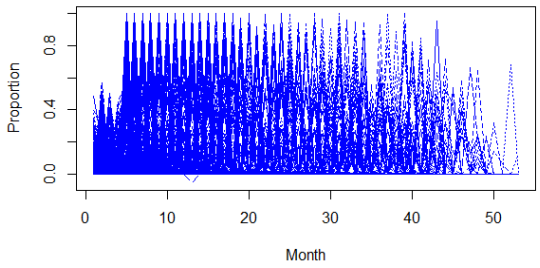

Figure 12: Life cycle of all observations in Cluster 3

On the other hand, Figure 13 demonstrates trends which underlie the remaining four clusters. While the proportion for some claims is not as high as the monthly mean in each cluster, we observe that in the aforementioned 'peak' months there is always a payment of some magnitude. Outlier claims also become apparent when plotted in this way, in particular those with negative proportions.[3]

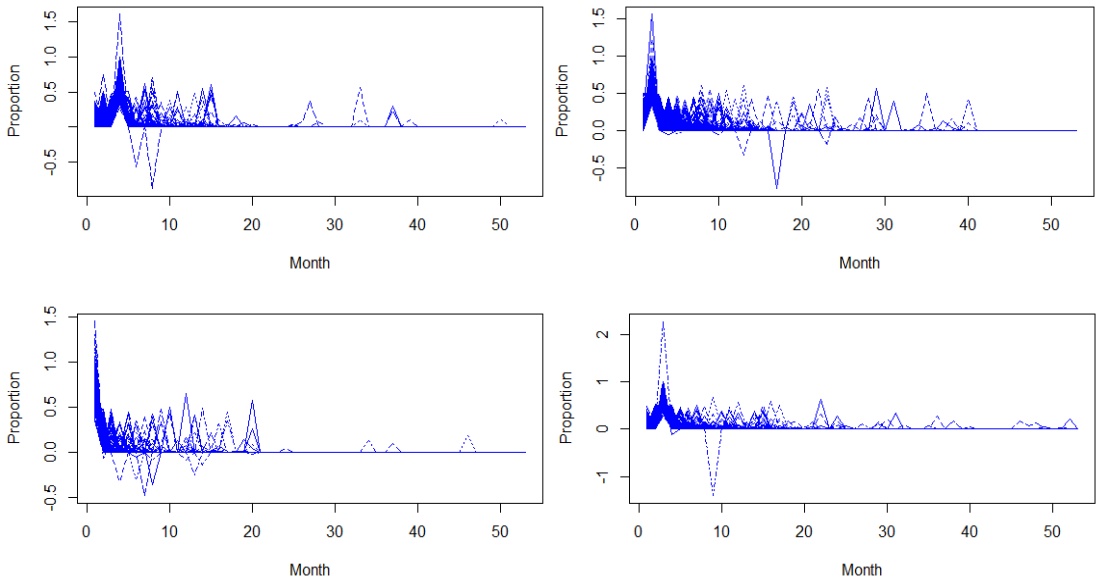

Figure 13: Life cycle of all observations in Clusters 1, 2, 4 and 5 (appearing clockwise from top-left)

[3] These were noted during the data pre-processing phase as unusual but yet reflective of true claims pathways which involve a return of funds to the insurance company and so remained in the dataset.

The variability of monthly averages is greater for Cluster 3 than any other cluster while it also remains high across many months (Figure 14). Furthermore, we discover that the 'peak' months for each cluster are also those with the highest standard deviation.

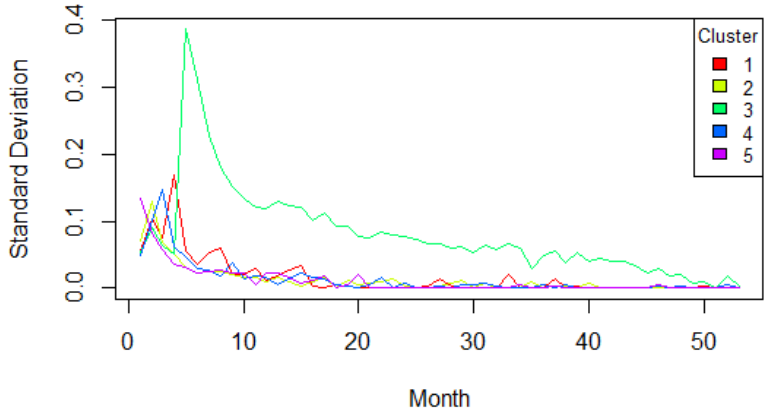

Figure 14: Average monthly standard deviation of proportions per cluster

5.4.3 Claim Features

This section details the characteristics that distinguish a motor insurance claim on a cluster-by-cluster basis. We split this examination into two parts; the first of which looks at information known at the time of the accident and is recorded when a report is filed, while the second highlights additional results which are of particular interest in the business context.

5.4.4 First Notice of Loss (FNOL) information

Here we describe some of the variables which contain incident-specific FNOL information. As this amounts to almost 80 variables, including those derived in our preparatory phase, it is not possible or of interest to report each individually. Those included in this section are variables which offer insight and aid in the description of clusters.

Injury

Claims that fall into Cluster 3 involve a much higher proportion of injuries than those in any other category. Of all 2010 claims in the database, 47.5% of those in Cluster 3 report an injury as compared to 8.7% overall. This equates to 73% of those which report an injury belonging to Cluster 3 while the distribution of 'No Injury' claims closely matches to cluster size (Figure 15).

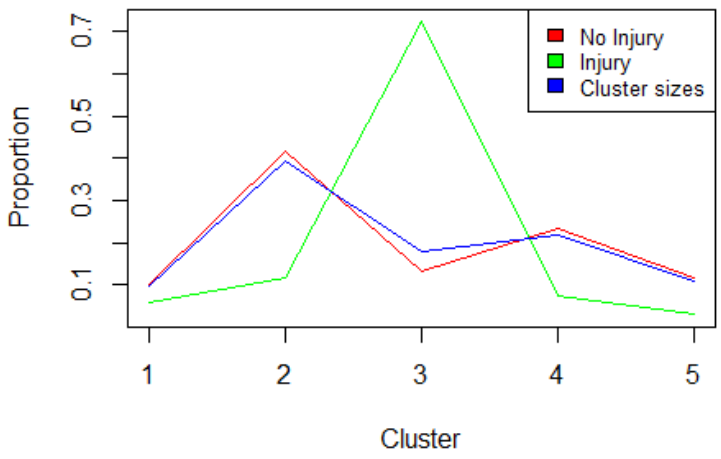

Figure 15: Cluster-wise information on the injury variable

Number of Third Parties Involved

Of the claims in which there is no third party involvement, their dispersion into clusters is in proportions almost identical to the actual cluster sizes. However this ratio changes radically with Cluster 3 featuring more predominantly as the number of third parties rise (Figure 16). At least 50% of claims which involve two or more third parties lie within Cluster 3. Exactly 50% of incidents where the number of third parties is 6 belong to Cluster 3. This percentage rises to 92% and 100% in cases where 4 and 8 external bodies are involved.

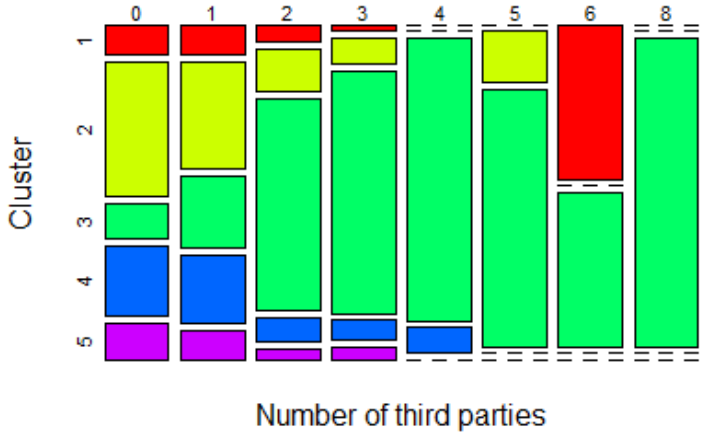

Figure 16: Cluster-wise information on the number of third parties involved in the incident

Further Comments

Other features which were examined in the possibility that they may show marked differences between clusters included;$driver\_gender$, $marital\_status$, class of use

of the car ($class\_of\_use$), insurance cover type ($cover\_type\_desc$). However, in the case of all the variables listed above, the proportion of claims from each cluster falling into each of the potential categories of response was in line with the overall ratio of cluster sizes.

5.4.5 Additional Results

The results of analysing claim costs and duration are reported in this section. Although this information can only be deduced once a claim is closed, they are included due to their business contributions. Figure 17 illustrates the cluster-wise averages of both these characteristics and their relationship.

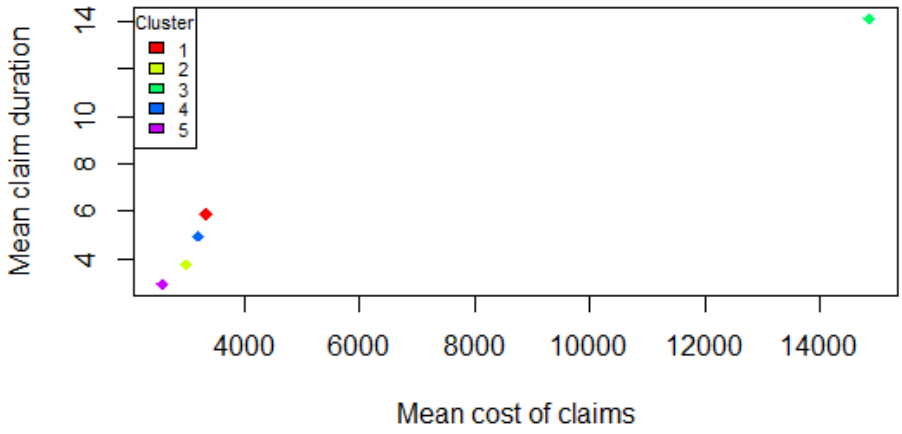

Figure 17: Scatterplot of mean total claim cost and mean claim duration per cluster

Total Claim Cost

Claims in Cluster 3 are more costly than those in other clusters with the mean for this cluster 4.5 times greater than that of the others. This measure of centrality is most skewed for Cluster 3 indicating the heavier presence of outliers here than in other clusters. Yet a comparison of the median values still finds the average for Cluster 3 to be 34% higher than that with the next highest median, Cluster 4.

Claim Duration

Similar to the trend shown in total claim costs, claims in Cluster 3 extend over a longer period of time than those in other clusters (Figure 18). With this variable however, the skew and kurtosis for clusters 1, 2, 4 and 5 is higher than that for Cluster 3. In this context it tells us that the duration of claims within these four

clusters is quite similar, with the exception of outliers, while those in Cluster 3 are generally more varied and longer overall.

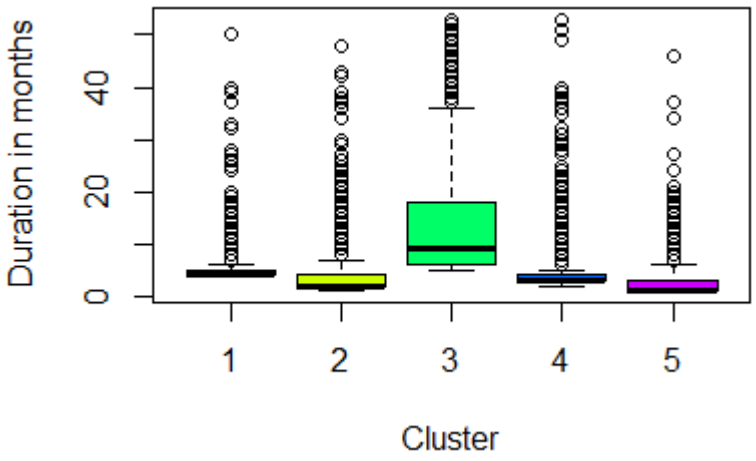

Figure 18: Boxplot showing the spread of claim durations in each cluster

5.5 Summary

In this chapter we reported on the results of the clustering phase of the research. Specifically, we detailed the performance of four clustering algorithms when applied to the proportional matrix containing the 7,985 claims originating in 2010. We provided the Davies-Bouldin index and the mean silhouette for the 24 options under consideration before eventually selecting the k-medoids algorithm, CLARA, to partition the data in five clusters. This choice is further defended by the stability and consistency of the clustering as quantified using the Rand index and the Jaccard coefficients. The clusters attained were then profiled and the results of this analysis described.

Chapter 6 - Classification Results & Analysis

6.1 Overview

This chapter begins by describing the results for the feature selection process with the objective of identifying the variables most relevent and useful for the classification stage. The second part of the chapter outlines and compares the results for each of the classification methods that were applied to the data, using the selected variables. The aim is to find a classificaiton technique which correctly classifies new and unseen claims into the appropriate cluster so that accurate predictions may be made regarding their payments life cycle.

6.2 Feature Selection

Figure 19 and the corresponding table in Appendix G describe the feature selection results from *Boruta*{Boruta}. A total of 26 variables were confirmed as being relevant for the classification of claims, twelve variables were rejected and two were identified as tentative, meaning there were insufficient iterations to draw conclusive results.

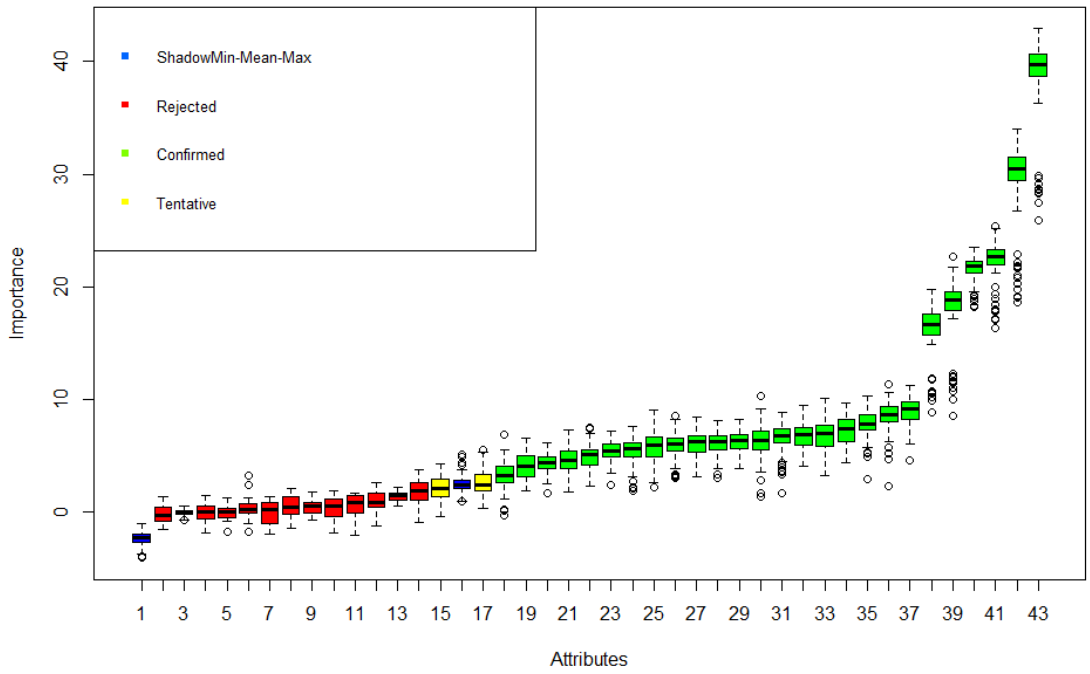

Figure 19: Boruta feature selection results when applied to 2010 data

The variable *claims_teamcode* produced the highest values and was therefore the most relevant variable in the selection process. The next five variables in order of importance were *num_tp_involved*, *class_of_claim*, *injury*, *date_reported* and *month_reported*. The variable with the highest Gini Index was *special_inv_flag*. The next five highest scoring variables were *injury*, *single_veh_dummy*, *multi_veh_dummy*, *num_days_loss_to_reported* and *num_tp_involved*. Variables ranked below 15 in both tests were included in the final formula (Figure 20). Appendix H reports the exact results for each variable and their corresponding ranking in both tests.

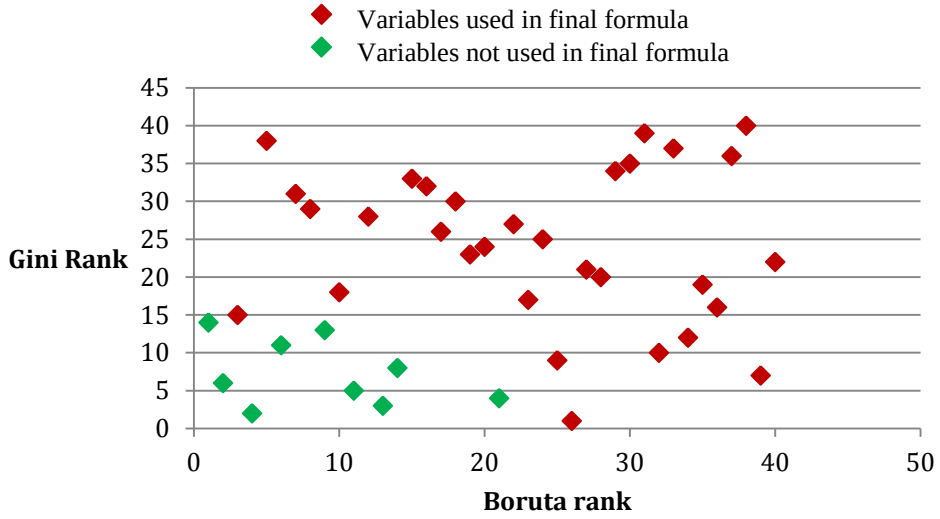

Figure 20: Boruta ranking vs. Gini ranking

The *class_of_claim* variable ranked 3rd in the Boruta results but 15th in the Gini Index. Since the *injury* variable was derived from the *class_of_claim* variable, and it ranked higher in both Boruta and the Gini Index, it was decided not to include the *class_of_claim* variable for the classification stage.

It was decided to include the variable, *multi_veh_dummy* despite its Boruta rank of 21. This was due to its low rank in the Gini Index, and also since it completed the information partly captured by *single_veh_dummy*.

The higest ranking variable in the Gini Index, *special_inv_flag* is a fraud indicator. It was excluded from the classification due to it's irrelevance in task at hand as well as its low rank in the Boruta test.

The final formula used in classification was:

$$formula.true = Cluster2010 \sim injury + single\_veh\_dummy + multi\_veh\_dummy + num\_days\_loss\_to\_reported + num\_tp\_involved + vehicle\_value + month\_reported + license\_type + claims\_teamcode$$

$Cluster2010$ was a factor variable ranging from 1 – 5 which described the cluster to which a claim was assigned during the clustering process. This was the target variable used in the classification stage.

6.3 Classification Results

Table 9: Multi-class classification accuracy results documents the performance measures as described in Section 4.4.4 for the multi-class classification task, having used the best performing parameters for each technique.

	Results using original training set			**Results using oversampled training set**	
	Training Set 2010	Test Set 2010	Test Set 2011	Training Set 2010 Oversampled	Test Set 2010 (as before)
Accuracy					
Baseline	0.3926111	0.3951158	0.4423625	0.2	*0.3926111*
K-NN	*0.4619077*	0.4211021	0.4440114	*0.9768*	0.2917971
ANN	0.4408265	*0.4483406*	0.487633	0.2388	0.151221
SVM	0.4416614	0.4480276	*0.4925798*	0.3438	0.3061991
Bayes	0.3926111	0.3951158	0.4423625	0.2	0.09298685
Average Class Accuracy					
Baseline	0.6860781	0.7580463	0.776945	0.68	*0.7580463*
K-NN	*0.7847631*	0.7684408	0.7776046	*0.98936*	0.7145898
ANN	0.7763306	*0.7793362*	0.7950532	0.53552	0.6604884
SVM	0.7770403	0.779211	*0.7983511*	0.73752	0.7224797
Naïve Bayes	0.6860781	0.7580463	0.776945	0.68	0.6371947

Table 9: Multi-class classification accuracy results

6.3.1 Original Training Set

When built using the original training set the SVM classifier was marginally better than the others on the 2011 test set, and second only to the ANN classifier on the 2010 test set. With regards to the training set, K-NN consistently produced the most accurate results, i.e. had the lowest training error.

All classifiers performed better than the overall accuracy baseline except the Naïve Bayes classifier, which predicted Cluster 2 for all claims. No classifier managed to produce an overall accuracy measure of above 50% in any of the training, 2010 test or 2011 test sets. The best overall accuracy of 49.26% was found with the SVM classifier using a polynomial kernel.

This classifier also produced the best average class accuracy of 0.7984. The figures for this accuracy measure are higher as they remove the bias of large clusters, in this case Cluster 2. Again all classifiers average class accuracy was equal to or higher than the baseline, though only very marginally.

Table 10 reports the confusion matrix for the SVM 2011 test set results as per *CrossTable*{gmodels}. Almost 93% of claims were predicted to be in Cluster 2 although only 44.2% of the 2011 test set is made up of claims in this cluster. As a consequence, only 47% of claims predicted to be in Cluster 2 were in Cluster 2. In contrast, the corresponding value for Cluster 3 is 79% indicating the predictive power for this cluster is more reliable.

Cell Contents						
N N / Row Total N / Column Total	N = the number of claims in Cluster **i** that were predicted in Cluster **j** w here $\boldsymbol{i, j} \in \{\mathbf{1, 2, 3, 4, 5}\}$					
Actual Cluster *i*	**Predicted Cluster *j***					
	1	**2**	**3**	**4**	**5**	**Row Total**
1	0 0.000 0.000	514 0.957 0.083	21 0.039 0.044	2 0.004 0.154	0 0.000 0.000	537 0.080
2	0 0.000 0.000	2907 0.985 0.470	42 0.014 0.088	2 0.001 0.154	0 0.000 0.000	2951 0.442
3	2 0.002 1.000	574 0.599 0.093	376 0.392 0.790	6 0.006 0.462	0 0.000 0.000	958 0.144
4	0 0.000 0.000	1377 0.979 0.223	26 0.018 0.055	3 0.002 0.231	0 0.000 0.000	1406 0.211
5	0 0.000 0.000	808 0.987 0.131	11 0.013 0.023	0 0.000 0.000	0 0.000 0.000	819 0.123
Column Totals	2 0.000	6180 0.926	476 0.071	13 0.002	0 0.000	**6671**

Table 10: SVM polynomial confusion matrix 2011 test set

6.3.2 Oversampled Training Set

The results in the previous section show that Cluster 2 was over predicted by all classifiers. This is due to the class imbalance issue previously discussed. The accuracy results did not improve however when the classifiers were built using the oversampled training data. For every classifier, both accuracy measures on the 2010 test set were below the baseline. Due to the poor results, further investigation of this method was not merited.

6.3.3 Binary Classification

A clear distinction between Cluster 3 and the remaining clusters became apparent through the cluster profiling analysis. Therefore, due to the unsatisfactory results from the multi-class classification, the problem was transformed into a binary one. Clusters 1, 2, 4 and 5 were represented by a 0 label and Cluster 3 was represented by a 1 label. It was established with our business sponsor that the loss in information resulting from this method would be outweighed if more accurate results were produced.

Table 11 describes the scores for the binary classification where the overall accuracy was the only performance measure computed.

	Results using original training set			**Results using oversampled training set**	
	Training Set 2010	Test Set 2010	Test Set 2011	Training Set 2010 Oversampled	Test Set 2010 (as before)
Accuracy					
Baseline	0.8280109	0.8071384	0.8563933	0.5	0.8071384
K-NN	0.8707994	0.8519098	0.8947684	0.99425	0.6912962
ANN	0.8701732	0.8550407	0.8931195	0.68525	0.780526
SVM	0.870382	0.8534753	0.8991156	0.6535	0.8456481
Naïve Bayes	0.8280109	0.8071384	0.8563933	0.5	0.8071384

Table 11: Binary classification accuracy results

The accuracy level of all classifiers appears to have improved dramatically with the highest accuracy of 89.91% again coming from the polynomial SVM classifier.

On further inspection however, only 38% of claims in Cluster 3 were predicted correctly (Table 12). This is 1% less accurate than the comparative result from the multi-class classifier. The binary classification therefore does not improve the

accuracy of predictions for Cluster 3. With the naïve Bayes binary classifier, all claims were classed as label 0 and it therefore performed as well as the baseline scenario. The other three classifiers were marginally more accurate than the baseline when the original training set was used.

Actual Label *i*	**Predicted Label *j***		
	0	**1**	**Row Total**
0	5634 0.986 0.905	79 0.014 0.178	5713 0.856
1	594 0.620 0.095	364 0.380 0.822	958 0.144
Column Total	6228 0.934	443 0.066	6671

Table 12: Binary SVM polynomial confusion matrix for 2011 test set

6.4 Model Performance

In this section we document the results of the model application. Applying the best classifier to the 2012-13 dataset resulted in the shown allocations to each cluster (Table 13):

Cluster	**1**	**2**	**3**	**4**	**5**
No. of claims	23	5140	1492	5751	6
Proportion of total	0.0019	0.4141	0.1202	0.4633	0.0005

Table 13: Cluster distribution of 2012-13 claims when classified using SVM classifier

Claims within each cluster were predicted to follow that cluster's life cycle as detailed in Section 5.4.2. Comparing the actual pathway of proportional payments for each claim with the predicted proportions for all 2012-13 claims showed a correlation of 0.36. As a benchmark, we report the Pearson correlation coefficient between the true proportions and those if all claims were predicted to be in Cluster 2. This is in keeping with the reference point used in testing the classifier and gives a correlation value here of 0.5. While aligning the monthly averages for both the true

and predicted values shows a similar trend (Figure 21), the statistics documented here uncover the discrepancies between the underlying values.

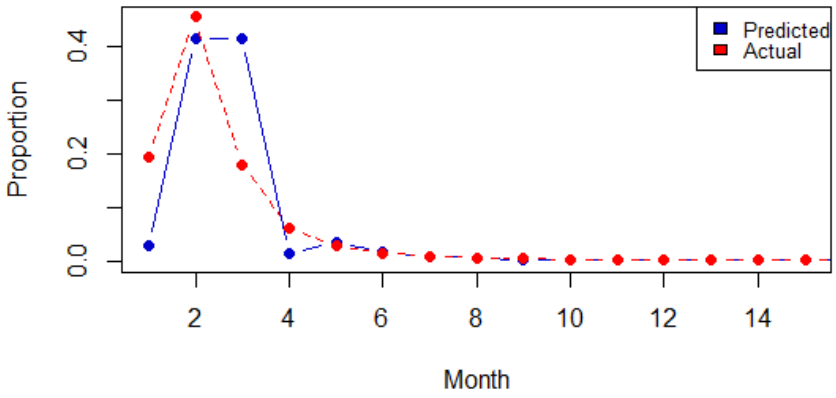

Figure 21: Mean monthly proportions comparing actual values and model predictions on 2012-13 data

6.5 Summary

This chapter highlighted the key results from the classification stage. After the outcomes of the Boruta and Gini index were presented, the decision process behind the final feature selection used to build the classification model was described. The accuracy results were then documented for each of the classification methods applied, along with a contingency table illustrating the outcome of the most accurate method, the SVM with a polynomial kernel. Overall however, the results of all classification techniques, multi-class, oversampling and binary, were disappointing. The evaluation of the model performance as a whole was then described and in keeping with the poor performance of even the best classifier, the predictions were less than satisfactory.

Chapter 7 - Discussion

7.1 Overview

In this chapter we examine the level of success achieved in the areas of clustering, classification and overall model performance. We discuss the key findings from each of these phases and address topics such as insights gained, contributions to business and academic novelty. Certain elements of the process produced better results than others. In light of this, we inspect issues which perhaps give rise to these problems and suggest potential adaptations.

Within the analysis of the clustering stage, we discuss the significance of the main observations generated by profiling of each cluster. In particular, we argue the validity of partitioning into five segments of claims. We also highlight issues which indicate that difficulties may be encountered at a later point. Analysing the outcome of the classification stage, we discuss factors involved in its poor performance and additions or changes which may aid its effectiveness.

Finally, we consider the success of our model as a whole and its ability to predict the payment paths, as expressed in proportions, of claims from 2012-13. We recognise that due to the weakness of the classifier, it is difficult to assess the quality of other aspects of the predictive process and so delve into an alternative means of evaluation. We conclude by discussing our approach to claim life cycle predictions and its potential to translate into an estimator of costs, thus competing with existing actuarial methods of loss reserving.

7.2 Clustering

From our research we have discovered that partitioning on claim payments data, as expressed in proportional form, does indeed result in the formation of stable clusters. A review of existing literature did not reveal any similar investigations or findings regarding the behaviour of insurance claims.

These groupings appear to be heavily based on the month in which the majority payment is made, with a clear ‘peak’ payment month identifiable in all five cases. One may argue that a manual allocation of claims to groups based on this ‘majority’ month would then suffice, eliminating the need for advanced machine learning

techniques. However, upon further examination of claims within each cluster, it seems that 'zero' payments as well as the proportional amounts in other months do impact on the cluster assignment. Hence, the application of clustering techniques is justified.

Such insight into the varying payment patterns amongst claims is beneficial to insurance traders as it offers a breakdown of the different categories of claims that exist. In conjunction with a predictor of the total cost of a claim, knowledge of cluster membership and the expected monthly percentage payment can effectively improve a company's loss reserving capabilities.

While natural groupings of proportion-based claim life cycles are apparent, one must recall the analysis of other features of each cluster. The lack of distinction between clusters 1, 2, 4 and 5 is noteworthy and perhaps hints at potential difficulties with classification. It also lends itself to the possibility of splitting the dataset into two groups, namely claims who fit the profile of Cluster 3 and those who don't.

The exercise of cluster investigation was highly beneficial. Although certain characteristics were quite similar across all clusters, it exposed the 'peak' payment months for each. This in turn supported the specification of five groupings when running the chosen clustering algorithm. Such results can be advantageous for a business once an effective classifier is developed. Furthermore, upon turning the attention to attributes other than the proportions paid out in each month, we can gain substantial understanding of the type of claim each cluster represents.

7.3 Classification

After iteratively training and testing using a range of different classification techniques, parameters and attributes, even the best performing classifier still produced remarkably poor predictions. From a business perspective, although clustering yields insightful and interesting results, an accurate predictive model is necessary in order to convert these findings into actionable knowledge. This section analyses the results from the classification, giving possible reasons for its failure, and discusses possible improvements.

The results show that, while it is possible to successfully partition the transactional data, the classification of claims into such clusters is a significantly more complex

process. Instinctively, one would assume that claims which follow similar payment pathways would also share other common characteristics. For all features used in classification, the correlation between the cluster label and the variable value is weak (Figure 22). *license_type* has the highest correlation of 0.053 and *single_veh_dummy* gives the lowest at -0.024. This in itself exemplifies the complex relationship between the target and explanatory variables which further reinforces the need for machine learning techniques in this process.

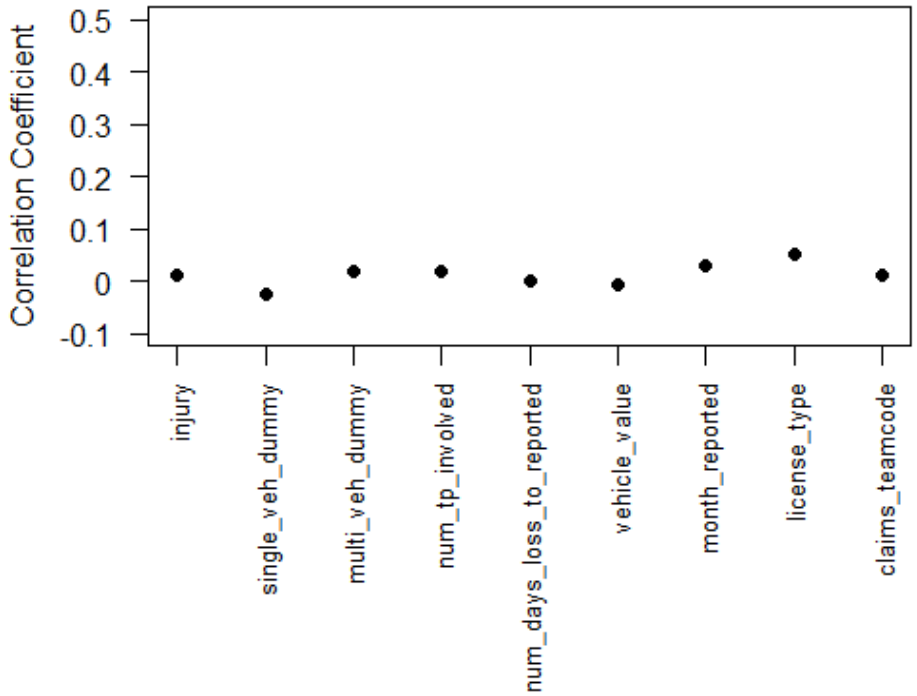

Figure 22: Plot of correlation between target and explanatory variables

Wang & Yao (2012) argue that, when the issue of class imbalance is present, random oversampling often causes the classifier to suffer from overfitting. This was found most notably with the K-NN method. Going forward, it may be of interest to implement other techniques to overcome this issue. Wang & Yao (2012) for example propose an ensemble method combined with oversampling which they argue has the ability to balance performance across several classes. This could be applied to both the multi-class and binary classification problems.

Cieslak & Chawla (2009) cite covariance shift as one potential reason for the failure of a classification model. Covariance shift occurs when the distribution of the training set differs from that of the test set. It is possible that the 2010 data used for

training the classifiers was distributed differently from that of the 2011 test data. After further discussion with our business sponsor, it became apparent that over the period the data was collected a number of changes occurred regarding the underlying claim reporting processes. Even within 2010 therefore, it is possible that procedural changes may have taken place which would result in inconsistencies across the data. This could perhaps explain the high training error present across all classifiers, as well as poor generalisation in the test sets.

The predictive power of the classifiers might also be improved with the addition of certain variables. Although the claim information dataset is comprehensive, there are a number of features not available but which we suggest might affect the payment life cycle of a claim. An example of such is information surrounding the location of the incident. Certain areas may have varying structures which affect the length of a claim or the speed at which claims payments are made, e.g. motor assessors in rural areas may be less accessible if required to cover a wider geographical area. This could increase the processing time for an incident assessment and therefore lengthen the claim duration.

7.4 Model Performance

Applying the model to claims from 2012 and 2013 gave very poor results when predicted proportions for all observations were compared with the true payment pathway of the claim. The fact that the benchmark correlation, where all claims were placed in Cluster 2, was better than that which arose from the predictor is disappointing. This bodes the question: what is to blame for the unsatisfactory performance of the model?

7.4.1 Contributory Factors

As we have established, the clustering of the dataset was successful. Therefore, we must assume that the poor predictive power of the model is caused by either the classifier, the values used for prediction, or a combination of both. As ascertained previously, all classification techniques trialled on the data were highly inaccurate. Despite using the best of these in our model, predictions made by the classifier are not reliable. Errors occurring at this phase of the process are then magnified when subsequent predictions are made.

Additionally, noted in Section 5.4.2 is the presence of higher variability of mean proportions for the 'peak' months in each cluster. When mean values are used in the prediction of proportional payments for new claims, this phenomenon is one which must be considered.

7.4.2 Alternative Assessment

To determine which of the above factors most contributed to the unsatisfactory level of prediction, we considered the following: bypassing the classifier, we used the labels which arose from clustering on the whole dataset to segment the 2012-13 claims (Table 14).

Cluster	**1**	**2**	**3**	**4**	**5**
Size	817	5712	1387	2426	2070

Table 14: Cluster sizes for 2012-13 data

This eliminated the misclassification element from the process and revealed a correlation between the true and predicted value matrices of 0.936, a striking improvement from the original correlation of 0.36. Furthermore, in calculating the Euclidean distance between these matrices as an alternative measure of similarity, we get a value of 35.25. This is comparable to a distance of 109.38 when computed based on the classifier and 100.19 from the benchmark scenario. These results are remarkably more promising and indicate that the quality of the classifier is dominant in hindering the success of the model.

7.4.3 Actuarial Methods

As outlined in Chapter 2, actuarial methods are currently the norm for predicting claim costs. They generally work by forecasting aggregated payouts over a set period of time thus enabling insurance companies reserve adequate funds. Though they do not make estimations on a claim-by-claim basis, their performance could be measured against a claim-specific model whose individual claim predictions are summed across the required time intervals.

However, our model predicts proportions rather than payments in monetary terms. While it could estimate the actual amount of a claim to be paid out in a given month, this is reliant on having an approximation of the total claim cost. With this in mind,

we considered the possibility of using the initial estimate of each claim as this approximation value. Such an initial estimate is recorded in the claim payments dataset provided having been assigned upon FNOL. With further investigation, it was found that the correlation between this original estimate of the total and the actual final claim cost was just 0.236. We considered this value too inaccurate to justify its use in predicting monthly claim costs. Thus, in its current form, our proposed technique is incomparable with methods such as the Chain Ladder or the Bornhuetter-Ferguson.

7.5 Summary

Within this chapter the results of the research process were discussed and analysed. We examined the successes and failures of the various stages of the project and outlined their meaning in academic and practical contexts. We also included a thorough analysis of the classification process and suggested possible improvements to the model. Finally, we discussed an alternative approach to assessing the predictor values of the model and the potential to extend the model to cost prediction.

Chapter 8 - Conclusions

After briefly summarising the paper, this chapter describes both the academic and practical contributions resulting from our work. The main limitations of our approach are addressed and future recommendations are then suggested to advance the task of predicting claim life cycles.

8.1 Summary

The main objective of this paper was to investigate the feasibility and precision of using machine learning techniques to predict the life cycle of motor insurance claim costs. It was clearly demonstrated through the clustering results and analysis that it is possible for motor insurance claims to be successfully grouped according to the life cycle of their payments. K-medoids with five clusters was the most effective and stable technique in this case.

The classification of claims into one of these clusters proved to be a more complex problem. None of the techniques investigated produced results which were reliable enough to be of use to the business sponsor. Possible reasons and ways to overcome the poor classification results have been discussed in Section 7.3.

Finally, the model as a whole was tested against the true life cycle of claims in the data provided. The actual average monthly proportions were compared to two sets of estimates. The first set was the aggregated monthly predictions from the classification stage and the second set was the monthly predictions if all claims had been predicted to be in the correct cluster. In making comparisons with these estimates we illustrated that using cluster-wise monthly averages to predict monthly proportions on a claim-by-claim basis can produce accurate results. Efforts to improve the classification of these claims are therefore justified.

8.2 Contributions

The approach taken in this paper is novel in a number of ways. To the best of our knowledge no research has been published to date which explores the behaviour of claim payments in terms of proportions. As such, the analysis carried out in this paper offers a unique perspective on claim costs.

The successful clustering of claims by their monthly proportional payments extends the understanding of claim cost life cycles. The findings in this section contribute to previous literature on claim costs and durations and could be used as a platform to future work.

With regards to the business contributions, we have succeeded in increasing the project sponsors knowledge surrounding claim payment behaviours. Although with its present performance the classification model does not perform well enough to be adopted by FTI, there were a number of results in the clustering stage from which valuable information can be drawn for example:

- In the data provided there exists five stable and valid groups of claims which follow similar cost life cycles. Each cluster is predominantly defined by their 'peak' payment month.
- The clusters are generally less distinguishable in terms of FNOL information. However the claims in one particular group have features which clearly differentiate them from those in the other four clusters. These characteristics are descriptive in the real world and are useful from a business perspective for example the number of third parties involved, the total cost of the claim, and whether an injury was involved.
- This cluster will have a minimum duration of five months whereas the other four are expected to have the majority of the claim paid out within the first four months of it being reported.

Having shown strong evidence that distinct pathways exist among claim payments, there is potential for the business to incorporate this into future predictive work.

8.3 Limitations & Future Work

The main limitations of the model outlined in this paper relate to the classification stage. Various recommendations are made in Section 7.3 which may improve the performance of the classifier. In terms of using additional variables however, we have restricted the information used to that which is known at FNOL. This was stipulated by our business sponsor as they wished to develop a model capable of immediate predictions as soon as a claim is reported. This prevents the inclusion of

variables such as the solicitor employed, healthcare professionals attended or other potentially important factors.

It is possible that the classification failed to produce accurate results because the relationship between claim payments and the incident and policy holder information is too weak. In this instance, the central assumption of our approach would be violated and predicting the cost life cycle of claims using data available at FNOL cannot be achieved. Intuitively however this outcome seems unlikely. On further investigation and through discussions with our business sponsor, it became evident that industry experiences suggest this is not the case. We therefore suggest that it would be beneficial for other pathways to be explored, with the aim of increasing classification accuracy levels.

Assuming the issues encountered during classification are rectified and an accurate predictive model built, there are a number of directions which future research could take. We now outline some recommendations for prospective work. These are aimed at extending the findings of this paper to further benefit both the business sponsor and the insurance industry as a whole.

- Rather than using average monthly proportions per cluster as the predicted values for future claims, a more advanced predictive function could be established, e.g. by using regression.
- Using the initial total estimate of a claim, the predictive function could then be used to convert the expected monthly proportions into cost values. This would subsequently allow for cost predictions on a more granular level, and has the potential to advance an insurance company's loss reserving methods.
- The analysis throughout this paper was carried out using motor insurance data. Many of the variables used to build the classifier are therefore unique to this type of insurance. A potential area of future work could be to apply the approach to other types of non-life insurance. Although the attributes used in the classification stage would need to be appropriate to the type of insurance, we see every possibility for the successful generalisation of our approach in this way.

- Modifying the model to incorporate process mining could also enhance the solution. Doing so would allow for any company specific processes to be taken into account and therefore has the potential to further improve the accuracy of predictions.
- Finally, as an alternative method to tackling the prediction of claim cost life cycles, the data used in the clustering and classification stages could be reversed. That is, the information and policy holder dataset could be used initially to find clusters of claims with similar characteristics. From this, techniques could be applied to each group to find any common themes with regards to their payment life cycles.

Appendices

Appendix A: Data Dictionary for Claim Payments spreadsheet

Variable	Description
claim_key	Claim ID
third_party_sequence • **Number 0 – 10**	Third Party sequence • 0 = System update – not transaction • 1= The insured person
policy_premium_class 1. **DMA** 2. **DMF** 3. **DMT** 4. **PMA** 5. **PMF** 6. **PMT**	Premium Class Policy • Unknown
claim_status 1. **F** 2. **O** 3. **R**	Status of Claim 1. Finalised 2. Open 3. Reopened
trans_id_date • **Date ranging from 2/1/2010 – 3/6/2014**	Transaction Date • Date of the transaction
trans_id_time	Time of Transaction • Time the transaction occurred
transaction_amount • **Number**	Transaction Amount
current_estimate • **Number**	Current Estimate
payment_code 1. **1** 2. **2** 3. **3** 4. **4** 5. **5** 6. **6** 7. **7** 8. **8** 9. **9** 10. **10** 11. **11** 12. **12** 13. **13** 14. **14** 15. **15** 16. **16** 17. **17** 18. **18** 19. **19** 21. **21** 22. **22** 23. **23** 24. **24** 25. **25** 26. **26** 27. **27** 28. **28** 29. **29** 30. **30** 31. **32** 32. **33** 33. **34** 34. **35** 35. **36** 36. **37** 37. **38** 38. **299** 39. **310** 40. **330** 41. **A02** 42. **B03** 43. **B04** 44. **C00** 45. **E02** 46. **E03** 47. **E04** 48. **F01** 49. **F02** 50. **F03** 51. **F04** 52. **F05** 53. **H00** 54. **I00** 55. **J00** 56. **K00** 57. **L05** 58. **NULL** 59. **X01**	Payment Code 1. Settlement-Insd.(Prop) Claim 2. Adjusters/Investigators/Engin. 3. Motor Assessors 4. Windscreen/Glass Breakage only 5. A.D. 6. A.D. (W/OFF) 7. Fire Damage 8. Fire (W/OFF) 9. Theft Damage 10. Theft (W/OFF) 11. Personal Effects 12. T.P. Vehicle Damage 13. Personal Accident Benefits 14. AD Car Hire Only 15. Medical Report Fee 16. Opponents Legal Costs 17. Own Solicitors Costs 18. T.P. Property(Non Vehicle) 19. Medical Exam Attendance Xs. 20. Costs paid on account 21. RTA Emergency Treatment Fees 22. Translators Fees 23. Trailer 24. Bank Charges 25. Share. Agreement 26. VAT 27. AD New Veh 28. PIAB 29. TP Car Hire Only 40. Ex Gratia T/L Insd.Veh Ex(F&T) 41. Brain (Temp,Fractures etc) 42. Sight (Injury) 43. Taste (Injury) 44. Facial (Scarring/Teeth/Etc) 45. Spine Neck (Fractures) 46. Whiplash 47. Back Injuries (Push/Pull/Lift) 48. Arm,Hand,Wrist, Fingers 49. Leg,Knee, Ankle,Foot,Toes 50. Shoulder 51. Pelvis,Hip, Sternum 52. Other Limb Injury 53. Other Diseases 54. Cosmetic Injury

20. 20	30. In House Legal Expenses 31. Witness Expense - non expert 32. Cost Accountant's fee 33. Towage Only 34. TP Claimant Expenses 35. Salvage Disposal Fee 36. Cash In Lieu Of Repairs AD 37. Cash In Lieu Of Repairs TP 38. Ex Gratia Settlement Insd.Prop 39. Ex Gratia IV Damage Ex(F&T/TL) 55. Fatal Injury 56. Minor Trivial Injury 57. Stress Related Injury 58. NULL 59. Private Investigator
payment_desc • **As Above**	Description of Payment Code • As Above
batch_year • **2010 - 2014**	Batch Year • The year that the record was last changed
total_paid	Total Paid
recovery_estimate	Recovery Estimate
current_gross_os	Current Gross Outstanding

Appendix B: Data Dictionary for Claim Information spreadsheet

Variable & Descriptors	Description
claim_key	Claim ID • The unique claim number
policy_number	Policy Number of Claimant • Policy against which claim is being made
acc_driver	
driver_name	Name of Driver
name_sim	
driver_gender **1. F** **2. M** **3. NULL**	Gender of Driver 1. Female 2. Male 3. NULL – not available
driver_date_of_birth **1. NULL**	Date of Birth of Driver 1. NULL – not available
employment_status **1. COMP** **2. CONV** **3. EMPL** **4. NULL** **5. RETI** **6. SELF** **7. UNEM**	Employment status 1. Company 2. Conversion code for integration 3. Employed 4. NULL 5. Retired 6. Self Employed 7. Unemployed
employment_desc • **As above**	Description of Employment • As above
driver_occupation	Driver Occupation
driver_occ_desc	Description of Occupation
license_type **1. 1** **2. 2** **3. 3** **4. 4** **5. 5** **6. 6** **7. 7** **8. NULL**	License Type 1. Full Irish License 2. Provisional Irish License 3. Full EU License 4. Full UK License 5. Full Australian License 6. Full Non-EU License 7. Full US & Canada License 8. NULL – Not Available
license_desc • **As Above**	Description of License Type • As Above
marital_status **1. M** **2. NULL** **3. O** **4. S**	Marital Status 1. Married 2. NULL 3. Other 4. Single
marital_desc • **As Above**	Description of Marital Status • As Above
class_of_use **1. 1** **2. 0**	Class of Use 1. Class 1 : Private Use 2. Social Domestic & Pleasure Use
class_of_use_desc • **As Above**	Description of Class of Use • As Above
certificate_driving • **13 levels made up of numbers**	Driving Certificate • Unknown
cert_driving_desc • **As Above**	Description of Driving Certificate • As Above

years_ncb • **Integers from 0 – 52**	Years of No Claims Bonus
driving_experience • **43 levels made up of a letter and number e.g. D5**	Driving Experience Code • Unknown
type_of_cover **1. A1** **2. A2** **3. A3**	Type of Cover 1. Third Party Only 2. Third Party Fire and Theft 3. Comprehensive
cover_type_desc • **As Above**	Description of Type of Cover • As Above
inception_date • **Date ranging from 14/3/1963 – 7/4/2014**	Inception Date • Original date of policy inception
expiry_date • **Date ranging from 9/1/2010 – 15/5/2015**	Policy Expiry Date • Date the policy is due for renewal
date_of_loss • **Date ranging from 1/1/2010 – 21/5/2014**	Date of Loss • Date the incident occurred
date_reported • **Date ranging from 2/1/2010 – 22/5/2014**	Date Loss Reported • Date the incident was reported to the insurance company
claim_narrative • **Manually inputted text**	Description of Claim • Short description of the claim
class_of_claim **1. 01** **2. 11** **3. M1** **4. M2** **5. M3** **6. M4** **7. M5** **8. M6** **9. M7** **10. M8** **11. M9** **12. MR** **13. MU**	Class of Claim 1. First notification no details 2. Medical Defence Union 3. Accidental damage only 4. Accidental damage & third party injury 5. Accidental damage & third party damage 6. Third party injury only 7. Third party damage only 8. Third party injury & third party damage 9. Invalid theft 10. Fire 11. Accidental damage, third party damage & third party injury 12. Theft recovered 13. Theft unrecovered
Class_of_claim_desc • **As Above**	Description of Class of Claim • As Above
claims_teamcode **1. 1** **2. 2** **3. 3** **4. 5** **5. 7** **6. 8** **7. 10** **8. 11** **9. 12** **10. 16** **11. 21** **12. LGL**	Claim Handler's Team Code 1. Motor Damage Claims Team 2. Own Damage Claims Team 3. RAPID RESPONSE CLAIMS TEAM 3 4. Motor Injury Team 5. PROPERTY CLAIMS TEAM 7 6. Specialist Claims Team 7. Claims Recoveries Team 8. Commercial Liability Team 9. MERRION SOLICITORS 10. CLAIMS OPERATIONS MANAGER 11. Property Damage Claims Team (& 1 case Miller Farrell 12. Legal Department

teamcode_desc • **As Above**	Description of Team Code • As Above
claims_handler • **100 levels made up of handlers initials**	Claims Handler • Initials of the claim handler
claims_handler_desc • **As Above**	Description of Claims Handler • As Above
special_inv_flag 1. **1** 2. **0**	Special Investigation • A flag which indicates whether a claim was referred to SIU - Special Investigation Unit
open_driving_claim 1. **Y** 2. **N**	Open Driving Claim 1. Yes – the diver was a named driver 2. No - the driver was not a named driver
gardai_attended 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Gardaí Attended Incident 1. Don't Know 2. Not Applicable 3. Gardaí did not attend incident 4. Garda attended incident
gardai_attended_desc • **As Above**	Description of Gardaí Attendance • As Above
car_stolen_without_keys 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Car Stolen Without Keys 1. Don't Know 2. Not Applicable 3. Car Stolen with the keys 4. Car Stolen without the keys
car_stolen_without_keys_desc • **As Above**	Description of Car Stolen without Keys • As Above
only_1_key_for_car 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Only One Key for Car 1. Don't Know 2. Not Applicable 3. 2 or more keys available 4. Only 1 key for the vehicle
only_1_key_for_car_desc • **As Above**	Description of Only 1 Key for Car • As Above
total_loss_frm_firemotor 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Total Loss from Motor Fire 1. Don't Know 2. Not Applicable 3. Not a total loss from fire 4. Total Loss from Fire
total_loss_frm_firemotor_desc • **As Above**	Description of Total Loss from Motor Fire • As Above
finance_on_the_car 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Outstanding Finance on the Car 1. Don't Know 2. Not Applicable 3. No Finance outstanding 4. Finance Outstanding
finance_on_the_car_desc • **As Above**	Description of Finance on the Car • As Above
single_vehicle_accident 1. **DKW**	Single Vehicle Accident 1. Don't Know

2. **NA** 3. **NO** 4. **YES**	2. Not Applicable 3. 2 / More vehicles involved 4. Single Vehicle Accident
single_vehicle_accident_desc • **As Above**	Description of Single Vehicle Accident • As Above
time_of_accident_motor 1. **1** 2. **2** 3. **DKW** 4. **NA**	Time of Motor Accident 1. Between 22.00 - 07.00 Hrs 2. Between 07.00 - 22.00 Hrs 3. Don't Know 4. Not Applicable
time_of_accident_motor_desc • **As Above**	Description of Time of Motor Accident • As Above
any_witnesses1 1. **DKW** 2. **NA** 3. **NO** 4. **YES**	Any Witnesses 1. Don't Know 2. Not Applicable 3. No Witness to the accident 4. Witness to the accident
any_witnesses_desc1 • **As Above**	Description of any Witnesses • As Above
single_vehicle_multiple_occs 1. **0** 2. **1** 3. **2** 4. **3** 5. **4** 6. **DKW** 7. **NA**	Single Vehicle Multiple Occupants 1. None 2. 1 3. 2 4. 3 5. 4 6. Don't Know 7. Not Applicable
age_of_acc_vehicle 1. **DKW** 2. **G10** 3. **L10** 4. **NA**	Age of Accident Vehicle 1. Don't Know 2. Greater than 10 years 3. Less than 10 years 4. Not Applicable
fire_brigade_attended 1. **DKW** 2. **NA** 5. **L10** 6. **NA**	Fire Brigade Attended Incident 1. Don't Know 2. Not Applicable 3. Did not attend 4. Did attend
key_location 1. **1** 2. **2** 3. **3** 4. **4** 5. **5** 6. **DKW** 7. **NA**	Location of Car Key • Unknown
tp_vehicle_multiple_occs 1. **0** 2. **1** 3. **2** 4. **3** 5. **4** 6. **DKW**	Third Party Vehicle Multiple Occupants 1. None 2. 1 3. 2 4. 3 5. 4 6. Don't Know
miab_claim_type 1. **F** 2. **I** 3. **M** 4. **N** 5. **O** 6. **P** 7. **T** 8. **W**	Motor Insurance Advisory Board Claim Type • Unknown
main_driver_indicator 1. **N** 2. **NULL** 3. **Y**	Main Driver Indicator 1. Does not normally drive the car 2. Null 3. Does normally driver the car
abi_vehicle_code • **Number Ranging from 7045 – 100000000**	Association of British Insurers Vehicle Code • Code per car make/model according to ABI
vehicle_make_and_model_descrip	Description of Vehicle Make and Model

vehicle_registration	Vehicle Registration
vehicle_value **Number from 200 – 300000**	Value of Vehicle in €
abi_year_of_manf • **Year ranging from 1984 - 2014**	Association of British Insurers Year of Manufacture
driver_type_of_cover **1. A1** **2. A2** **3. A3** **4. NULL**	Type of Driver Cover 1. Third Party Only 2. Third Party Fire and Theft 3. Comprehensive 4. No given
driver_cover_desc • **As Above**	Description of Driver Cover • As Above
relationship_to_proposer **1. BR** **2. DA** **3. EM** **4. FA** **5. ME** **6. NULL** **7. OT** **8. PR** **9. PT** **10. SI** **11. SO** **12. SP**	Relation of Insured to Proposer 1. Brother 2. Daughter 3. Employee 4. Family 5. Self/Proposer 6. Not given 7. Other 8. Parent 9. Partner 10. Sister 11. Son 12. Spouse
relationship_desc • **As Above**	Description of Relationship of Insured to Proposer • As Above
excluded_driver_flag **1. N** **2. NULL** **3. Y**	Excluded Driver Flag • Unknown
conviction_code **1. A1** **2. B1** **3. C1** **4. D1** **5. D2** **6. E1** **7. G1** **8. L1** **9. NO** **10. NULL**	Conviction Code 1. Speeding Under 5 yrs ago 2. Without due care U 5yrs ago 3. Careless Driving Under 5 yrs 4. Dangerous Driving Under 5 yrs 5. Dangerous Driving Over 5 yrs 6. Drunken Driving Under 5 yrs 7. No Insurance Under 5 yrs 8. Other Convictions 9. No Conviction(s) 10. NULL
conviction_desc • **As Above**	Description of Conviction Code • As Above
drivers_refused_cover **1. N** **2. Y**	Drivers Refused Cover 1. No 2. Yes
suspension_of_cover **1. F&A** **2. F&T** **3. FUL**	Suspension of Cover • Unknown
no_of_vehicles **1. 1**	Number of Insured Vehicles 1. All observations =1
batch_type **1. AR** **2. EN** **3. NBA** **4. NB** **5. NULL** **6. RA** **7. RN**	Batch Types 1. Auto Renewals 2. Endorsement (MTA) 3. New Business Adjustment 4. New Business 5. NULL 6. Renewal Adjustment 7. Manual Renewal
batch_type_desc • **As Above**	Description of Batch Type • As Above
driver_age • **Integer from 17 - 97**	Age of Driver
row_rank • **1**	Row Rank • Unknown All observations = 1

Appendix C: Proportional Matrix extract from R

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
462330	0.213204495	0.062497516	0.000000000	0.72429799	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462334	0.043215212	0.884461810	0.000000000	0.00000000	0.000000000	0.07232298	0.00000000	0.000000000	0.000000000	0.000000000
462336	0.000000000	0.960729860	0.021184111	0.01808603	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462337	0.000000000	1.000000000	0.000000000	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462338	0.070223379	0.929776621	0.000000000	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462341	0.000000000	0.000000000	0.000000000	1.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462345	0.005308838	0.000000000	0.000000000	0.04135826	0.023034216	0.00000000	0.00000000	0.016984992	0.913313689	0.000000000
462348	0.030780972	0.000000000	0.969219028	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462349	0.000000000	0.046823726	0.953176274	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462350	0.181636227	0.748424955	0.000000000	0.00000000	0.069938818	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462355	0.046078562	0.793998362	0.158971039	0.00000000	0.000952036	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462359	0.906012060	0.053216639	0.040771301	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000
462371	0.000000000	1.000000000	0.000000000	0.00000000	0.000000000	0.00000000	0.00000000	0.000000000	0.000000000	0.000000000

Appendix D: List of R packages used

Package name	Purpose	Citation
cluster	K-medoids clustering using CLARA	Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2015, "cluster: Cluster Analysis Basics and Extensions. R package version 2.0.1"
clValid	SOTA clustering	Brock. G, Pihur. V, Datta, .S., Datta, S. 2008, "clValid: An R Package for Cluster Validation. Journal of Statistical Software, vol. 25, no.4, pp. 1-22, " Available at: http://www.jstatsoft.org/v25/i04/
e1071	Support Vector Machine classification	Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. 2015, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-6. " Available at: http://CRAN.R-project.org/package=e1071
fastcluster	Hierarchical clustering	Müllner, D. 2013, "Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python", Journal of Statistical Software, vol. 53, no. 9, pp. 1-18
flashClust	Hierarchical clustering	Langfelder, P. & Horvath, S. 2012, "Fast R functions for robust correlations and hierarchical clustering", Journal of Statistical Software, vol. 46, no. 11, pp. 1-17
fossil	Calculating Rand Index: a measure of cluster similarity	Vavrek, M.J. 2011, "fossil: Palaeoecological and palaeogeographical analysis tools", Palaeontologia Electronica, vol. 14, no. 1
fpc	Calculating the Jaccard Coefficient: a measure of cluster stability	Hennig, C. 2014, "fpc: Flexible procedures for clustering. R package version 2.1-9". Available at: http://CRAN.R-project.org/package=fpc
gmodels	Generating Contingency Table	Warnes, G.R., Bolker, B., Lumley, T., Johnson, R.C. 2015, "gmodels: Various R Programming Tools for Model Fitting. R package version 2.16.2". Available at: http://CRAN.R-project.org/package=gmodels
kknn	K-Nearest Neighbour classification	Schliep, K. & Hechenbichler, K. 2015, "kknn: Weighted k-Nearest Neighbors. R package version 1.3.0". Available at: http://CRAN.R-project.org/package=kknn
NbClust	Davies-Bouldin index for evaluation of clusterings	Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. 2014, "Nbclust: An R package for determining the relevant number of clusters in a dataset", Journal of Statistical Software, vol. 61, no. 6, pp. 1-36

nnet	Artificial Neural Network classification	Venables, W.N. & Ripley, B.D. 2002, Modern applied statistics with S, 4th edn, Springer, New York
plyr	Balancing cluster sizes within training sample	Wickham, H. 2011, "The split-apply-combine strategy for data analysis", Journal of Statistical Software, vol. 40, no. 1, pp. 1-29
psych	Descriptive statistics	Revelle, W. 2015, "psych: Procedures for Personality and Psychological Research", Northwestern University, Evanston, Illinois, USA. Available at: http://CRAN.R-project.org/package=psych Version = 1.5.4

Appendix E: Clustering algorithms ranked by Davies-Bouldin index and overall average silhouette

Clustering	K	DB Index	DB Rank	Silhouette	Silhouette Rank	Sum of Ratings
CLARA	8	0.9977014	2	0.6971	11	13
SOTA	7	1.06115	13	0.7129	3	16
CLARA	7	1.016873	5	0.6955	12	17
Kmeans	6	1.034771	10	0.7013	8	18
Kmeans	5	1.009678	4	0.6738	15	19
Hierarchical	9	1.1883	18	0.7171	1	19
CLARA	6	1.04234	11	0.7004	9	20
Kmeans	7	1.069978	14	0.7089	6	20
Hierarchical	8	1.1752	16	0.7126	4	20
CLARA	4	0.9693397	1	0.6233	20	21
CLARA	5	1.02439	8	0.6743	14	22
SOTA	6	1.052342	12	0.6974	10	22
SOTA	5	1.02411	7	0.6707	16	23
Kmeans	4	1.007712	3	0.6087	22	25
Hierarchical	7	1.1905	19	0.7055	7	26
SOTA	8	1.64042	24	0.7137	2	26
SOTA	9	1.599893	23	0.7117	5	28
SOTA	4	1.018973	6	0.6068	23	29
Hierarchical	4	1.0292	9	0.615	21	30
Hierarchical	6	1.1825	17	0.6896	13	30
Hierarchical	5	1.1464	15	0.6625	17	32
Kmeans	9	1.209129	20	0.6562	18	38
Kmeans	8	1.222053	22	0.6525	19	41
CLARA	9	1.218803	21	0.5211	24	45

Appendix F: Monthly cluster mean proportions

Column headings X1, X2, etc. relate to monthly time intervals.

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	0.022	0.059	0.029	0.843	0.013	0.004	0.008	0.005	0.002	0.002	0.003	0.001	0.001	0.002
2	0.042	0.907	0.02	0.012	0.004	0.003	0.003	0.002	0.002	0.001	0.001	0	0.001	0
3	0.019	0.051	0.024	0.015	0.249	0.141	0.071	0.045	0.031	0.025	0.021	0.021	0.025	0.022
4	0.02	0.067	0.869	0.016	0.009	0.004	0.003	0.002	0	0.001	0.001	0.001	0	0.001
5	0.937	0.029	0.013	0.005	0.004	0.001	0.001	0.001	0.002	0.001	0	0.001	0.001	0

Cluster	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27
1	0.002	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.02	0.016	0.018	0.013	0.013	0.01	0.009	0.011	0.01	0.009	0.009	0.007	0.007
4	0.002	0.001	0.001	0	0	0	0	0.001	0	0	0	0	0
5	0	0.001	0.001	0	0	0.001	0	0	0	0	0	0	0

Cluster	X28	X29	X30	X31	X32	X33	X34	X35	X36	X37	X38	X39	X40
1	0	0	0	0	0	0.001	0	0	0	0.001	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.006	0.007	0.006	0.007	0.006	0.008	0.006	0.003	0.005	0.005	0.004	0.005	0.003
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0

Cluster	X41	X42	X43	X44	X45	X46	X47	X48	X49	X50	X51	X52	X53
1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.003	0.003	0.003	0.002	0.001	0.002	0.001	0.001	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0

Appendix G: Boruta Results Table

Variable	Boruta X-axis label
shadowMin	1
driving_experience	2
shadowMean	3
class_of_use	4
total_loss_frm_firemotor	5
only_1_key_for_car	6
certificate_driving	7
marital_status	8
years_ncb	9
car_stolen_without_keys	10
abi_vehicle_code	11
fire_brigade_attended	12
type_of_cover	13
conviction_code	14
excluded_driver_flag	15
shadowMax	16
fince_on_the_car	17
special_inv_flag	18
batch_type	19
tp_vehicle_multiple_occs	20
driver_gender	21
gardai_attended	22
multi_veh_dummy	23
time_of_accident_motor	24
abi_year_of_manf	25
single_vehicle_multiple_occs	26
any_witnesses1	27
main_driver_indicator	28
driver_type_of_cover	29
vehicle_value	30
single_veh_dummy	31
employment_status	32
num_days_loss_to_reported	33
miab_claim_type	34
license_type	35
single_vehicle_accident	36
relationship_to_proposer	37
month_reported	38
date_reported	39
injury	40
class_of_claim	41
num_tp_involved	42
claims_teamcode	43

Appendix H: Feature Selection Ranking Results from Gini and Boruta metrics

Variable	BorutaMedianZ	BorutaRank	GiniIndex	GiniRank
claims_teamcode	39.69500175	1	0.2552176	14
num_tp_involved	30.51473639	2	0.6469958	6
class_of_claim	22.73920997	3	0.2481149	15
injury	21.81775079	4	0.9195992	2
date_reported	18.87948983	5	0.004340092	38
month_year_reported	16.64534127	6	0.3309097	11
relationship_to_proposer	9.16646543	7	0.06097014	31
single_vehicle_accident	8.68057214	8	0.07254773	29
license_type	7.86049592	9	0.2933923	13
miab_claim_type	7.36663181	10	0.1859634	18
num_days_loss_to_reported	7.00348235	11	0.7845273	5
employment_status	6.84518849	12	0.08117891	28
single_veh_dummy	6.79816446	13	0.8963056	3
vehicle_value	6.38417551	14	0.433664	8
driver_type_of_cover	6.38181978	15	0.0241063	33
main_driver_indicator	6.30218618	16	0.05181901	32
any_witnesses1	6.25655997	17	0.08974607	26
single_vehicle_multiple_occs	6.11225239	18	0.07254773	30
abi_year_of_manf	5.9318757	19	0.1096234	23
time_of_accident_motor	5.69305864	20	0.1087662	24
multi_veh_dummy	5.45720007	21	0.866124	4
gardai_attended	5.07877649	22	0.08265399	27
driver_gender	4.61562597	23	0.193669	17
tp_vehicle_multiple_occs	4.43502422	24	0.0972235	25
batch_type	4.1134987	25	0.3943683	9
special_inv_flag	3.2876671	26	0.9412649	1
fince_on_the_car	2.42545918	27	0.1241488	21
excluded_driver_flag	2.13167023	28	0.1427775	20
conviction_code	1.94775883	29	0.01965465	34
type_of_cover	1.44624075	30	0.01118765	35
fire_brigade_attended	0.90553869	31	0.003332625	39
abi_vehicle_code	0.84929784	32	0.3545068	10
car_stolen_without_keys	0.50786348	33	0.004865471	37
years_ncb	0.50737356	34	0.3236738	12
marital_status	0.42423388	35	0.147899	19
certificate_driving	0.28451758	36	0.2032828	16
only_1_key_for_car	0.26975352	37	0.006489956	36
total_loss_frm_firemotor	0.05584174	38	0.003207875	40
class_of_use	0.0414303	39	0.5474014	7
driving_experience	-0.32757541	40	0.1220288	22

References

ACM SIGMOD. 2015, ACM SIGMOD. Available at: http://www.sigmod.org/sigmod-awards/citations/2006-sigmod-test-of-time-award-1 [Accessed 20/06/2015]

Aftab, S., Abbas, W., Bilal, M.M., Hussain, T., Shoaib, M. & Mehmood, S.H. 2013, "Data mining in insurance claims two-way mining for extreme values", IEEE, pp. 1

Aparna, K., Nair, M.K. 2014 "Enhancement of K-Means algorithm using ACO as an optimization technique on high dimensional data," Electronics and Communication Systems (ICECS), 2014 International Conference on, pp.1-5

Aviva plc. 2014, *Annual report and accounts 2014*. London: Aviva plc. Available at: http://www.aviva.com/reports/2014ar/ [Accessed 25/08/2015]

Boland, P.J. 2007, *Statistical and probabilistic methods in actuarial science*, Boca Raton, FL: Chapman & Hall/CRC

Central Bank of Ireland. 2012, *Private Motor Insurance Statistics*. Dublin: Central Bank of Ireland. Available at: http://www.centralbank.ie/polstats/stats/motorins/Pages/releases.aspx [Accessed 08/06/2015]

Central Bank of Ireland. 2014, *Central Bank Insurance Statistics 2013*. Dublin: Central Bank of Ireland. Available at: https://www.centralbank.ie/publications/Documents/Insurance%20Statistics%202013.pdf [Accessed 25/08/2015]

Central Bank of Ireland. 2015, *Solvency II – Introduction*. Available at: http://www.centralbank.ie/regulation/industry-sectors/insurance-companies/solvency2/Pages/default.aspx. [Accessed 08/06/2015]

Cieslak, D.A. & Chawla, N.V. 2009, "A framework for monitoring classifiers' performance: when and why failure occurs? ", Knowledge and Information Systems, vol. 18, no. 1, pp. 83 – 108.

Cunningham, P., & Delany, S. J., 2007, "k-Nearest neighbour classifiers", Technical Report UCD-CSI-2007-4, 1-17

Devale, A.B. & Kulkarni, R.V. 2012, "Applications of Data Mining Techniques in Life Insurance", International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 4, pp. 31-40

England, P.D. & Verrall, R.J. 2002, "Stochastic Claims Reserving in General Insurance", British Actuarial Journal, vol. 8, no. 3, pp. 443-518

Friedman, N., Geiger, D. & Goldszmidt, M. 1997, "Bayesian Network Classifiers", Machine Learning, vol. 29, no. 2, pp. 131-163

Ghorpade-Aher, J. & Meter, V.A. 2014, "PSO based Multidimensional Data Clustering: A Survey", International Journal of Computer Applications, vol. 87, no. 16, pp. 41-48

Grize, Y.L. 2015, "Applications of Statistics in the Field of General Insurance: An Overview", International Statistical Review, vol. 83, no. 1, pp. 135-159.

Frees, E.W. & Valdez, E.A. 2008, "Hierarchical insurance claims modelling", Journal of the American Statistical Association, vol. 103, no. 484, pp. 1457-1469

Han, J. & Kamber, M. 2001, *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, California; London

Hennig, C. 2007, "Cluster-wise assessment of cluster stability", Computational Statistics and Data Analysis, vol. 52, no.1, pp. 258-271

Hennig, C. 2008, "Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods", Journal of Multivariate Analysis, vol. 99, no. 6, pp. 1154-1176

Hennig, C. 2014, "fpc: Flexible procedures for clustering. R package version 2.1-9". Available at: http://CRAN.R-project.org/package=fpc. [Accessed 15/06/2015]

Jafar, O.A.M. & Sivakumar, R. 2010, "Ant-based Clustering Algorithms: A Brief Survey", International Journal of Computer Theory and Engineering, vol. 2, no. 5, pp. 787-796

Jafar, O.A.M. & Sivakumar, R. 2014, "Distance Based Hybrid Approach for Cluster Analysis Using Variants of K-means and Evolutionary Algorithm", Research Journal of Applied Sciences, vol. 8, no. 11, pp. 1355-1362

Jain, A.K., Mao, J. & Mohiuddin, K.M. 1996, "Artificial neural networks: a tutorial", Computer, vol. 29, no. 3, pp. 31-44

Jessen, A., Samorodnitsky, G. & Mikosch, T. 2010, "Prediction of outstanding payments in a Poisson cluster model", Scandinavian Actuarial Journal, vol. 2011, no. 3, pp. 214-24

Kashima, H., Hu, J., Ray, B. & Singh, M. 2008, "K-means clustering of proportional data using L1 distance", 19th International Conference on Pattern Recognition. Tampa, FL 8-11 Dec 2008. pp.1-4

Kaufman, L. & Rousseeuw, P.J. 1990, *Finding groups in data: an introduction to cluster analysis*, New York: Wiley, Chichester

Klüppelberg, C. & Severin. M., 2003, "Prediction of Outstanding Insurance Claims"

Kotsiantis, S.B. 2007, "Supervised machine learning: a review of classification techniques", Informatica, vol. 31, no. 3, pp. 249

Kuo, R.J., Ho, L.M. & Hu, C.M. 2002, "Integration of self-organizing feature map and K-means algorithm for market segmentation", Computers and Operations Research, vol. 29, no. 11, pp. 1475-1493

Kuo, R.J., An, Y.L., Wang, H.S. & Chung, W.J. 2006, "Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation", Expert Systems with Applications, vol. 30, no. 2, pp. 313-324

Kursa, M.B. & Rudnicki, W.R. 2010, "Feature Selection with the Boruta Package", Journal of Statistical Software, vol. 36, no. 11, pp.1-13

Liu, Q., Pitt, D. & Wu, X. 2014, "On the prediction of claim duration for income protection insurance policyholders", Annals of actuarial science, vol. 8, no. 1, pp. 42-62

Mack, T. 1994, "Which stochastic model is underlying the chain ladder method? ", Insurance: Mathematics and Economics, vol. 15, pp. 133-138

Mane, S.U. & Gaikwad, P.G. 2014, "Hybrid Particle Swarm Optimization (HPSO) for Data Clustering", International Journal of Computer Applications, vol. 97, no. 19

Martínez-Miranda, M.D., Nielsen, J.P. & Verrall, R. 2012, "Double Chain Ladder", ASTIN Bulletin – The Journal of the IAA, vol. 42, no. 1, pp. 59-76

Mary, C. & Raja, S.V. 2009, "Refinement of Clusters from K-Means with Ant Colony Optimization", Journal of Theoretical and Applied Information Technology, vol. 9, no. 2, pp. 28-32

Met Éireann. 2011, *Monthly Weather Bulletin No. 306.* Dublin: Available at: http://www.met.ie/climate/irish-climate-monthly-summary.asp [Accessed 10/06/2015]

Moni-Sushma-Deep, K. & Srinivasu, P. 2014, "The Importance of Feature Selection in Classification", International Journal on Computer Science and Engineering, vol. 6, no. 1, pp. 63-68

Plat, R. & Antonio, K. 2014, "Micro-level stochastic loss reserving for general insurance", Scandinavian Actuarial Journal, vol. 2014, no. 7, pp. 649-669

Rand, W.M. 1971, "Objective Criteria for the Evaluation of Clustering Methods", Journal of the American Statistical Association, vol. 66, no. 336, pp. 846-850

R Core Team, R Foundation for Statistical Computing, Vienna, Austria. 2015, R: A language and environment for statistical computing. Available at: http://www.R-project.org/ [Accessed June 2015]

Rudnicki, W.R., Wrzesien, M. & Wieslaw, P. 2015, "All Relevant Feature Selection Methods and Applications", *Feature Selection for Data and Pattern Recognition.* Heidelberg: Springer Berlin. pp.11-28

Schmidt, K. D. 2006, "Methods and Models of Loss Reserving Based on Run-Off Triangles–A Unifying Survey", Casualty Actuarial Society Forum, Fall 2006, pp. 269-317

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. & Wang, Z. 2007, "A novel feature selection algorithm for text categorization", Expert Systems with Applications, vol. 33, no. 1, pp. 1-5

Singh N. & Singh, D. 2013, "The Improved K-Means with Particle Swarm Optimization", Journal of Information Engineering and Applications, vol. 3, no. 11, pp. 1-7

Sokolova, M. & Lapalme, G. 2009, "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, no. 4, pp. 427-437

Tan, P., Steinbach, M. & Kumar, V. 2006, *Introduction to Data Mining*. First Edition. Boston: Addison-Wesley Longman Publishing Co., Inc,.

Viaene, S., Derrig, R.A., Baesens, B. & Dedene, G. 2002, "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection", The Journal of risk and insurance, vol. 69, no. 3, pp. 373-421

Wang, S. & Yao, X. 2012, "Multiclass Imbalance Problems: Analysis and Potential Solutions", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130

Wickham, H. 2011, "The Split-Apply-Combine Strategy for Data Analysis", Journal of Statistical Software, vol. 40, no. 1, pp. 1-29

Witten, D. M., & Tibshirani, R. 2010, "A framework for feature selection in clustering", Journal of the American Statistical Association, vol. 105, no. 490, pp. 713–726

Wüthrich, M.V. & Merz, M. 2008, *Stochastic claims reserving methods in insurance*, Chichester: John Wiley

Yang, J., Wang, Y., Qiao, Y., Zhao, X., Liu, F. & Cheng, G. 2015, "On Evaluating Multi-class Network Traffic Classifiers Based on AUC", Wireless Personal Communications, vol. 83, no. 3, pp. 1731-1750

Yeo, A.C., Smith, K.A., Willis, R.J. & Brooks, M. 2001, "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry", International Journal of Intelligent Systems in Accounting, Finance and Management, vol. 10, no. 1, pp. 39

Zhang, T., Ramakrishnan, R. & Livny, M. 1996, "BIRCH: an efficient data clustering method for very large databases", ACM SIGMOD Record, vol.25, no. 2, pp.103-114