Faculty of Engineering and Information Technology
School of Software
University of Technology Sydney

Applying client churn prediction modelling on home-based care services industry

A thesis submitted in fulfillment
of the requirements for the degree of
Master of Analytics (Research)

by

Raul Manongdo

November 2017

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

To Maricel

for your love, understanding

and support

Acknowledgments

Foremost, I would like to express my deep appreciation to my supervisor, Professor Guandong Xu, for his professional guidance, persistent help and continuous support throughout my Masters study and research.

I would also like to thank Dr. Chunming Liu, Dr. Bin Fu and Stephan Curiskis for their scientific advice. Without their generous support, this thesis would not have been possible. Also to my co-workers at UTS Advance Analytics Institute, Xiao Zhu and Dr. Frank Jiang, whom I worked closely in this industry project and for their technical support for my research.

And most specially, to all the staffs at the anonymous company for providing the data and the domain knowledge on home care services industry.

Raul Manongdo
November 2017 @ UTS

This research is supported by an Australian Government Research Training Program Scholarship.

Contents

List of Figures

List of Tables

List of Publications

Papers Published

- **Manongdo Raul**, Xu Guandong (2016), Applying churn prediction modeling on home-based care services industry *in* '2016 International Conference on Behavioral, Economic and Socio-cultural Computing (**BESC2016**)', p.42, full paper accepted.

Abstract

Client churn prediction is widely acknowledged as a cost-effective way of realising customer life-time value especially for service-oriented industries and operating under a competitive business environment. Churn prediction model allows identification of clients as targets for retention campaigns. While there are for hospital-based care services, the author was unable to find application for home-based care services.

The objective of the study therefore is to develop an initial client churn prediction model in the context of home-based care services industry at Australia that can be adopted and subsequently enhanced. Real industry data as provided by a local and sizeable home-based care services provider was used in this study. For developing the model, various predictive models such as logistic regression, tree-based C5.0 and the ensemble Random Forest were tested. Feature selection techniques embedded in these models were integrated to identify significant and common variables in predicting a binary outcome of a client churning or not.

All model evaluations yielded overall prediction accuracies over 83%. The C5.0 model, however, was chosen as its prediction accuracy was marginally better and model results were easier to understand and adopt by the case company. It was discovered that in general, clients who are enrolled in the government's home assistance support program and with higher levels of home care needs (i.e. nursing) are more at-risk of churning. Clients enrolled in private and commercial programs are also at risk particularly those in the under-25 age group.

Chapter 1

Introduction

1.1 Introduction and Context of Study

Client churn prediction is a machine learning method performed in data mining to gain insights on the likelihood of a customer continuing or ceasing to subscribe to a company's products and services. (Luan 2002) stated that "data mining is a process of uncovering hidden trends and patterns using a combination of explicit knowledge base, sophisticated analytical skills and business domain knowledge using statistics and programming methods. In contrast to traditional analytical studies which are often in hindsight and aggregate in nature, data mining is forward looking and provides an ability to gain a deeper understanding of the patterns previously unseen". Machine learning, although it uses many of the same algorithms and techniques as data mining, is more focused on learning from known patterns and knowledge and apply that to other data. This study is more aptly categorised under data mining.

A churned client is an existing customer who discontinues enrolment to a company's services due to an unsatisfactory experience or unfavourable perception. Churn prediction provides an opportunity for an agent or company to plan and act before a client actually churns. In addition, churn models and analysis systems aid in developing marketing, sales and client retention

campaign strategies and also in improving operational customer services and processes.

The consequential benefit of client churn prediction modelling is direct and immediate: profitability and sustainability. Losing customers not only leads to opportunity costs because of reduced sales, but also leads to an increased need of attracting new customers. In telecommunication industry, (Huang, Zhu, Yuan, Deng, Yanhua, Ni, Dai, Yang & Zeng 2015) claimed that it costs 3 times more to acquire new customer than to retain existing ones. Another paper in (Heng-liang Wu n.d.) claimed a higher ratio of 5 to 6 times as much to sign in a new customer. It is no wonder then why data mining and machine learning were applied early to this application domain.

Churn analysis and modelling is an interdisciplinary research field covering operations research, marketing, business process management and information technology in the context of a single enterprise. (Nguyen 2011) described churn model as a major component of analytic systems that in general forms part of a company's Customer Relationship Management (CRM) system. CRM systems can include other models such as client retention and customer satisfaction that are usually integrated to client churn prediction models. In this thesis, customer satisfaction was also studied in relation to client churn for a specific home care agency.

1.2 The Problem

Various literature on successful client churn prediction projects had been published. The applications were mostly for well established companies like banks and telecommunications company with large and controlled data repositories captured and stored automatically. While there were studies for hospital-based health care services, the author however was unable to find an application for home-based care services.

It is generally acknowledged that home-based care is a growing industry globally. This is due to an aging population and prolonged life span as a result

of better health services and technology. With continued government subsidy for aged and disability care services, government regulation and controls to homeservice providers are expected to increase as market condition changes.

In (Brownlee 2015*a*), it was claimed that an ageing population is a fuelling demand with a large cohort of Australians aged 65 and older, strong growth in industry revenue is anticipated to continue. A research group (*i.e.* CoreData) found that consumers increasingly prefer to stay in their primary residence and delay moving to aged care facilities until declining health requires them to do so. Government funding arrangements also promoted delays in moving to nursing homes, as the value from the sale of the primary residence will be included in the future for government means testing.

In Australia starting this year, the market for home-based care services was deregulated. Home care clients can now choose their home care service provider (Health 2017) whereas previously, clients were referred by the government to a pool of age care service providers. With greater openness and transparency brought about by the recent introduction of the government web portal `http://MyAgedCare.gov.au`, consumers are now becoming more discerning in selecting a home-care service provider.

In view of deregulation, increasing competition, changing consumer preferences and growing consumer market, the local home-care service providers in general are now preparing and finding new ways to seek clients and retain existing ones. One way is by leveraging on existing company data assets, performing data analysis and modelling client behaviour, such as client churns. For the company under study, the client churn rate stands at 20% per year as shown in Figure 4.1 and the company will most likely benefit on having such a model.

1.3 Aim of this Study

The objective of the study therefore was to perform churn analysis and to create a churn prediction model in the context of an actual home-based care

service provider in Australia. As there are no prior prediction models done for this industry, the aim was to analyse client churns and to create an initial model that can subsequently be enhanced. The prediction model that exhibited high prediction accuracy and can readily be adopted by the service provider was selected. Insights from churn analysis and prediction modelling for a local company in this industry is to be presented. Real industry data as provided by a sizeable home-based care services company was used.

The research questions then were

- What are the main predictors for a client to churn from a company providing home-based care services in Australia?
- Given the limitations of home-based care service providers, what prediction models are suitable to use?
- What are the challenges in developing a churn model for this case company? Are the challenges applicable to other service providers in the local industry?

These questions was to be answered by developing a churn prediction model using industry data provided by an anonymous home-based care service provider in Australia. In this study, three candidate prediction models were considered: logistics regression (core team 2017) , Random Forest (Breiman 2001) and C5.0 (Kuhn, Weston, Coulter, Culp & Quinlan 2015) classification models.

1.4 Research Significance and Contribution

As mentioned, no client churn models had been developed for home-based care service provider both locally and overseas. The contribution of this research was the application of prediction modelling and churn analysis for this domain in the local setting. An initial model with acceptable prediction accuracy was built and was found to be suitable for use by a company in this

industry. Areas for further model improvement were suggested which were based on literature surveyed on associated home-care studies.

For home care clients, churn models can lead to better home-based care service and client satisfaction. The aged and disabled clients can get more quality care at a time in their lives when they are most vulnerable. For companies, churn models can help retain loyal customers and aid in acquiring new ones. The resulting churn analysis and business insights can guide company managers in devising customer marketing strategies and operational service improvements.

1.5 Thesis Structure

The following chapters provide details on how this thesis is organised.

Chapter 2 starts with a description of home-based care services in the local setting and the business processes involved, from client program enrolment to the point of churning. It then describes client churn analysis and prediction models in general and gives an overview of the feature selection, modelling, testing and evaluation used in the study.

Chapter 3 covers the literature surveyed with emphasis on churn analysis and data mining under home-based care services, mostly coming from the USA. Prediction models and techniques in other industries are also presented. Together with the previous chapter, the review lays the conceptual and practical foundations for churn analysis and developing prediction models for home-based care services locally.

Chapter 4 describes the data used as obtained from the case company and includes data collection, cleansing, churn measures and exploratory analysis. This chapter also describes the data aggregation performed to apply churn analysis in various dimensions. Dimensions shown are client home state, services, age group, etc. as it relates to churns. Client satisfaction measure currently in use by the case company is also presented.

Chapter 5 covers the development and evaluation of candidate prediction

models considered in the study. The summary of the model development is presented in Table 5.1. The chapter includes a description of the selected model, the justification for its use in this type of industry and ways to improve its prediction accuracy. It ends with business insights deduced from various models developed.

Chapter 6 summarises the findings of the study and examines future enhancement of the prediction model given the literature surveyed.

To aid in understanding the terms used in this thesis, a vocabulary is shown in Appendix F,

Chapter 2

Background

2.1 Introduction

This chapter describes home care services industry in general and the case company under this industry in particular, that was used for this thesis. It proceeds with a technical description of churn analysis, prediction modelling and techniques used in general in this discipline with emphasis on those adopted for this project. The implementation details of chosen techniques are described in succeeding Chapter 4 on data descriptive analysis and Chapter 5 on]prediction modelling. The conclusion describes aspects presented in this chapter that influenced the conceptual and technical development of the churn analysis and model.

2.2 Home care services industry

In (Health 2017), it ws claimed that "there are over 2,000 aged-care service providers in Australia, supplying two basic types of aged care services; Home Assistance and Community Care (HACC) and residential care (*i.e.* retirement village). HACC provides four levels of care provided to older people living at home, ranging from basic to high-care needs. Residential care provides accommodation and support for those who choose to live in

residential aged-care facilities (*i.e.* retirement/ nursing homes). Both types employ home-based care services. All these aged-care services are supplied by a variety of for-profit, not-for-profit and church-based service providers... Majority of aged care services is supplied by not-for-profit service providers across all types of care, with the market share ranging from 52% in residential care to 74% in HACC... Most home care (51%) and residential care (57%) service providers in Australia operate solely in metropolitan areas. Regional providers of residential care incur higher costs owing to smaller facilities with a higher proportion of low-care low-revenue residents... The government makes a significant contribution to funding aged-care places across HACC and residential care services. In 2013-2014, it spent $4.1 billion on fees in residential care and $87 million on home care... the funding is determined by the level of service needed by the consumer and consumers contribute privately towards some age-care services".

2.2.1 Trends for Home Care Services

With a large cohort of Australians aged 65 and older, an ageing population and longer life expectancy , strong growth in industry revenue is anticipated to continue. In the last seven years, much of the growth in aged-care supply came from the not-for-profit providers and this trend will likely continue. (Health 2017)

(Brownlee 2015*a*) stated that "given the public sensitivity around aged-care and the increasing burden on the government, public funding and tighter regulation of the industry will continue. (Brownlee 2015*a*) Successful organisations will need to implement and adjust operations in line with changing regulatory environment and complex service standards which may increase the service providers' wage costs. Companies will continue to seek economies of scale to combat rising costs and competition. Further ageing-in-place preferences will require flexibility to meet the demand for services spanning the care continuum from independent living through to palliative care."

To ensure sustainability, the Australian Government recently introduced

a demand-driven model of service delivery that promotes 'Consumer Directed Care' . (Health 2017). Under such a model, demand for aged-care services is dependent on consumer needs. Consumers who wish to age-in-place (i.e. at home) but requires support services may access HACC. Consumers who have high-care needs requiring 24-hour care and accommodation may access to extra funding from residential care. To maximise benefit, customers are now free to choose their home-care service provider unlike before wherein the government grants HACC allocations to various providers. With greater openness and transparency as a result of the government web portal `http://MyAgedCare.gov.au`, consumers are becoming more discerning in selecting a home-care service provider.

2.2.2 Peculiarities of Home Care Services

It is claimed that home care services are different to acute-care, office-based health care, mental health-care and other consumer services in general. (Geron, Smith, Tennstedt, Jette, Chassler & Kasten 2000)

- Clients are mostly aged, frail and disabled wherein service expectations and issues may not be objectively and clearly communicated. For some cases (*e.g.* Alzheimer), the next-of-kin performs these on the clients behalf.

- In general, services performed are varied and in many possible combinations (*e.g.* nursing, home-maker, care management etc.) paid for from various sources (mixture of funding sources and government support entitlement)

- Services are performed at the consumer's daily living situation and are frequently for a long and extended duration. Clients have strong opinions and preferences on how services are to be delivered, having performed the same during earlier years when able-bodied.

- Most services are considered 'low- tech', often provided by personnel with limited training and without professionally derived standards of practice. (Mylod & Kaldenberg 2000)
- All services are delivered in-person usually by many attending home care worker (HCW). For the case company, more than three-quarter of its business operating expense is for manpower.

Home care providers are typically small to medium-size companies and the author is under the impression that they have limited capacity to develop business processes and automate as compared to large and established companies.

With the peculiar characteristics of home care services, churn analysis and modelling can be made slightly different. More variables pertaining to human client interactions need to be included. Data can potentially be sourced from multiple channels (*e.g.* blogs, social media), agents(*e.g.* client, next-of-kin, HCW) and formats (*e.g.* voice, text). Health and general well-being indicators need to be included and periodically monitored. Client satisfaction needs to be monitored more frequently and churn models evaluated on a shorter duration than normal. Typically, models are re-learnt after 5 to 6 months after initial deployment.

As service providers have lesser analytics capability, prediction models should be easy to understand and be used at the operational field level. The models to be developed should consider the mix of client enrolled services and associated risk, complexity and profitability.

2.3 Case company

The case company was chosen mainly due to the university's ongoing industry engagement at the time. In keeping with prior agreement, the company's identity is not disclosed in this thesis. The company had been operational for more than two decades in an industry of more than a thousand small

to medium-size service providers. It claims to offer the most comprehensive set of products and services and specialises only in home-based care services excluding retirement villages. It also claims to have a significantly large manpower and client base across all states, both regional and metropolitan. It has over 1,000 staffs as compared to the average size of 42.6 staff for home-care service providers in this industry (Brownlee 2015*a*). It readily adapts to new technology and presumably has a more mature data capture and data quality assurance systems. Recently, the company migrated to a popular cloud-based mobile application systems for use by HCWs on the field.

The business process agents for home care services in this case company is shown in Figure 2.1.

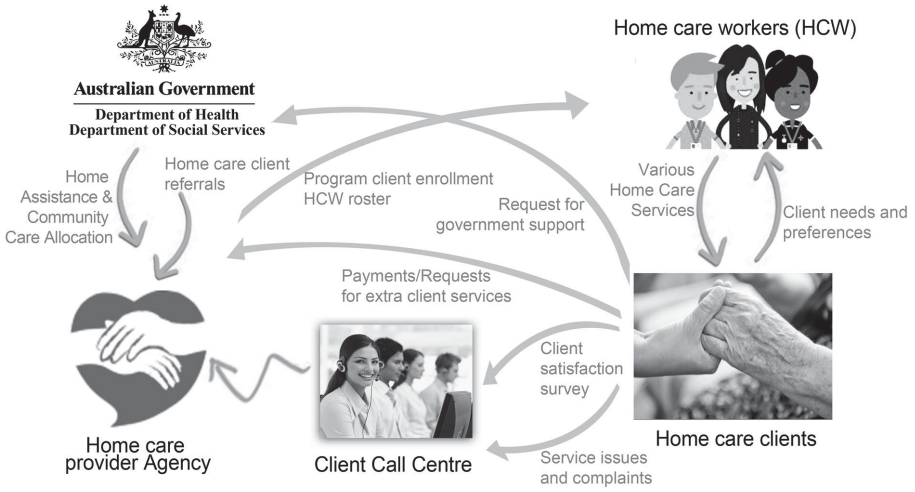


Figure 2.1: Home-based care services Business Process Agents

The case company offers a variety of home-based services to clients. Most services are related to health and ageing (*e.g.* personal care, nursing services, respite) while some are dedicated to home services (*e.g.* domestic cleaning services). Services are bundled into a program and a client can enrol into multiple programs at the same time. Services are delivered to clients by a home care worker (HCW) and some services require prior training and/or

license to practice(*e.g.* nurse, therapist).

The government allocates HACC unit entitlements to various home and aged care agencies in terms of head counts. As needs arise, the government refers specific clients to the company. The company also fills in its fixed unit HACC allocation through its own campaigns.

A potential client is initially assessed by a Program Manager with regards to eligibility for government and needed home care services. Clients receive subsidies from the government to pay for health and aged-care services. Other support services such as respite for their personal care-giver or transport are determined. Service delivery preferences are also obtained (*e.g.* service hours, attending HCWs gender, spoken language).

Upon enrolment, the required workforce to deliver agreed services is determined and a roster of HCWs with the appropriate skills is periodically created. When service appointment is due, a time sheet is given to an HCW to record service hours and kilometres travelled. An issue is that around 30% of rosters are revised due to errors in work scheduling, appointment cancellations and unavailability of skilled HCWs. Most HCWs are on casual on-call employment basis and replacements on short notice are not readily available.

Over time, historical information is captured and analysed. Home-based care performance metrics are monitored by the company. The number of distinct HCWs serving a client is of particular interest as clients as it is generally believed that clients prefer a few set of HCW who are familiar with his/her needs and had built a rapport. Other metrics are the count of service cancellations and time duration a client waits for an HCW appointment/service.

To monitor client satisfaction, a monthly customer survey is performed. The survey is separately performed at the levels of the national office and each state. The same customer survey instrument is used on both occasions.

A client's health condition may deteriorate requiring hospital admission leading to a suspension of enrolment or a change to a higher grade of home-care service. Service delivery issues arise on occasions and phone to a call-centre who relays the same to the client's program manager. Depending

on severity, Issues are escalated at predefined higher levels for appropriate action and resolution. If the client remains unhappy, he ceases enrolment in a program and is accordingly discharged.

2.4 Client Churn Prediction, Satisfaction and Retention

Within an enterprise, client churn analysis forms parts of a business intelligence framework and falls under customer relationship management(CRM) (Singh & Samalia 2014). Other CRM components closely associated are customer satisfaction and client retention.

The goal of these components are complimentary and basically the same. Churn prediction aims to maximise customer life-time value from limited resource for retention marketing campaigns (Sunil Gupta 2006). Churn model allows identification of high risk clients as target of retention campaigns. Client satisfaction is aimed at maintaining loyal and happy customers who are less likely to churn. Ideally, all components are integrated and an example automated system for a telecommunications company in the US is described in (William McCausland 1998). In (Kim Y. S. 2012), it is claimed that client retention strategies based on churn probability that considers expected yearly revenue makes the best out of many churn models.

A churn model's output is an input into a retention model. Clients with high probability of churning are prioritised for retention campaigns. Retention campaigns can satisfy disgruntled customers but can, however, be done regardless of churns such as enticing loyal customers to new products and services (*i.e.* cross-selling). Hence, CRM component models have variables in common and each component have other features that pertains only to its own sphere. For example, retention model considers special offers and a likelihood of acceptance for retention campaigns (Y. Kim & Johnson 2013). The frequency of execution of component models may also be different and need not be in sync. Satisfaction surveys and retention campaigns may occur

more frequently or independently of churn monitoring.

2.5 Churn Analysis and Prediction Modelling

Prediction model attempts to predict a target outcome based on a given dataset as input and an algorithm applied to it. The predicted value can be for a continuous or categorical (*i.e.* small finite set of values) variable. If the outcome falls into only two possible states (*i.e.* 0 or 1, non-churn and churn), this is referred to as binomial and the model as a binary classifier.

Many statistical models were successfully used as binary churn classifiers. This includes logistic regression, decision trees, boosted trees, gradient boosted decision trees, random forest, neural networks, evolutionary computation (*e.g.* genetic algorithm and ant colony optimisation), support vector machines and ensemble of hybrid methods (Huang et al. 2015).

In supervised learning, the target outcome of observations in a dataset is known. This allows discovery of patterns from relationships amongst the attributes and identification of suitable algorithms that fit the known outcome. Having a large set of observations allows generalisation of patterns that make it possible to predict the outcome for unseen data.

2.5.1 Feature Selection Techniques

Initial data gathered for data mining are normally huge with extreme high dimensionality (*i.e.* number of attributes). High dimension data tends to degenerate model performance due to noise (*i.e.* irrelevant) and redundant variables. Feature selection techniques aim to minimise redundancy and maximise relevance to the target label variable. (Tang, Alelyani & Liu 2015) described many feature selection techniques which are generally described in this section.

According to whether the training set is labelled or not, feature selection algorithms can be categorised into supervised, unsupervised and semi-supervised. Supervised feature selection assesses the relevance of features

guided by the label information attached to each observation instance while unsupervised feature selection works with unlabelled data. Clustering quality measures and other constraints are usually introduced in unsupervised techniques that potentially can eventuate into valid feature subsets.

It is common to have a data set with huge dimensionality but small labelled-sample size. High-dimensional data with small labelled samples permits too large a hypothesis space yet(unsupervised) with too few constraints (labelled instances). Under the assumption that labelled and unlabelled data are sampled from the same population generated by target concept, semi-supervised feature selection makes use of both labelled and unlabelled data to estimate feature relevance.

Supervised feature selection methods can further be broadly categorised into filter models, wrapper models and embedded models. The filter model separates feature selection from classifier learning so that the bias of a learning algorithm does not interact with the bias of a feature selection algorithm. It relies on measures of the general characteristics of the training data such as distance, consistency, dependency, entropy and correlation. Relief, Fisher score and Information Gain based methods are among the most representative. These technique maintain the physical meanings of the original features.

Another technique is feature extraction wherein variables are transformed into a new feature space with lower dimensions and the newly constructed features are usually combinations of original features. Examples of feature extraction techniques include Principle Component Analysis, Linear Discriminant Analysis and Canonical Correlation Analysis. With this method, it is difficult to link the transformed features to the original feature space as there is no physical meaning for the transformed features.

The wrapper model uses the predictive accuracy of a predetermined model to determine the quality of selected features. The feature selection algorithm is native to the model used and supposes that the optimal selection of features should consider the classifiers evaluation method.Wrapper technique needs to run a classifier many times to assess the quality of selected features which

can be a computationally expensive especially large number of features.

Embedded feature selection takes advantage of wrapper models that interact with the classifier and filter models that are far less computationally intensive than wrapper methods. There are three types of embedded methods; pruning methods, models with built-in mechanisms for feature selection and regularisation models which have as its objective minimise fitting errors such as Lazzo, elastic net, etc.

Correlation analysis can also be used to identify redundant attributes. There are two main types of correlation coefficients: Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient. Pearson's coefficient is used when both variables being studied are normally distributed. Spearman's coefficient is appropriate when one or both variables are skewed or ordinal and is robust when extreme values are present (National Institute of Health Sep 2012)

2.5.2 Regression and Classification

Regression relates the response variable (dependent) to a linear combination of predictor (independent) variables by calculating coefficients. The distinction between regression and classification models is on the type of response variable; regression predicts a numerical or quantitative value and classification predicts a value from a finite (though still possibly large) set of classes or categories. Generalised linear models(GLM) allow categorical response variables (or some transformation) to be processed in a manner like modelling for numeric responses. Logistic regression is an example of a GLM wherein it uses a function, logit, to relate response variable to predictors variables (Group June 2017).

Regression uses maximum likelihood estimation that does several iterations in finding solutions until it gets the smallest possible deviance or best fit. A deviance is a measure of distance between an observed and predicted outcome value and the distribution of the deviance residuals is used for calculating coefficients and model accuracy.

The iteration is referred as Fisher scores and is the count of model fitting cycles in changing the coefficient estimates towards a better model fit. Incrementally, different estimates are used (Newton-Raphson algorithm by default) and the model refitted. The algorithm stops when re-estimating and moving further would not yield much additional improvement (Erhardt 2017). The Akaike information criterion (AIC) is used to measure the relative quality of models generated using the same maximum likelihood estimation The one with the minimum AIC value is the preferred model (Wikipedia June 2017).

2.5.3 Decision Trees and Ensemble methods

Predictions can be determined using decision-tree based models. The topmost node in a tree is the root node, the leaf nodes are terminal nodes that hold the classification label and non-leaf nodes that are internal nodes containing a rule condition. A dataset is recursively partitioned into smaller subsets as the tree is being built. These algorithms adopt a greedy or non-backtracking approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner (Han & Kamber 2006).

In partitioning, an attribution selection measure is used to split a tree node. Attribute selection is a heuristic for selecting the splitting criterion that best separates a data partition given observations labelled into individual classes. At each split, a measure for tree construction is computed; information gain or Entropy, Gain Ratio and Gini Index (Han & Kamber 2006).

Ensemble methods on decision trees generate many classifiers and aggregates model results to achieve better prediction accuracy. For classification, two well-known methods are boosting and bagging (Liaw & Wiener 2002).

In boosting, decision trees are learnt in sequence giving extra weight to previous incorrect predictions which are fed into he next classifier iteration. This process continues for a pre-determined number of iterations and stops if the most recent classifier are either extremely accurate or inaccurate. Trials over numerous datasets show that 10-classifier boosting reduces the error

rate by an average of 25% (Research March 2017).

In bagging, successive trees do not depend on earlier iterations. Each is independently constructed using a bootstrap sample of the dataset. A bootstrap is a sample with replacement of the population wherein an observation can be re-selected or be missed out in sampling. In RF, each node is split using the best among a subset of predictors randomly selected in constructing tree nodes. Standard trees consider all variable to split tree nodes. This counter-intuitive strategy is claimed to perform very well compared to other classifiers and is robust against over-fitting (Breiman 2001).

In both ensemble methods, final classification of an observation instance is determined by majority voting.

2.5.4 Support Vector Machine

SVMs try to find a linear optimal hyperplane so that the margin of separation between the positive and the negative examples is maximised However, in practice, the data is often not linearly separable. To enhance the feasibility of linear separation, one may transform the input space via a non-linear mapping into a higher dimensional feature space. This transformation is done by using a kernel function. In SVM, there are only two free parameters to be chosen; namely the upper bound and the kernel parameter. The solution of SVM is unique, optimal and global since the training of an SVM is done by solving a linearly constrained quadratic problem wherein the data points closest to the optimal hyperplane, play a crucial role (*i.e.* support vectors). SVMs are based on the structural risk minimisation principle, which means that this type of classifier minimises the upper bound on the actual risk, compared to other classifiers which minimise the empirical risk. There are only a few published implementations of SVMs in a customer churn environment. (Coussement & den Poel 2008)

2.5.5 Artificial Neural Net

Artificial neural networks (ANN) attempt to simulate biological neural systems which learn by changing the strength of the synaptic connection between neurons upon repeated stimulations by the same impulse. Connection weights express the relative importance of each input to a processing element called a neuron, and a network learns through repeated adjustments of weights.

A summation function computes the weighted sums of all the input elements entering each processing element. A transformation function (*e.g.* sigmoid) combines the inputs from several neurons and then produces an output based on the transfer function before the output reaches the next processing element. ANN is composed of processing elements termed as neurons grouped in layers to form the networks structure. Three layers of a neural network are input, intermediate (called the hidden layer), and output. Each input corresponds to a single attribute. A hidden layer is a layer of neurons that takes input from the previous layer and converts those inputs into outputs for further processing. The outputs of a network contain the solution to a problem. (Tsai & Lu 2009)

2.5.6 Ant Colony Optimisation

Ant Colony Optimisation (ACO) employs artificial ants that cooperate in a similar manner to their biological counterparts in finding solutions for discrete optimisation problems (W. Verbeke 2011). Ants iteratively construct solutions by adding pheromone to the paths corresponding to solutions. Path selection is a stochastic procedure based on not only a history-dependent pheromone value but also a problem-dependent heuristic value. The pheromone value gives an indication of the number of ants that chose the trail recently, while the heuristic value is a problem dependent quality measure. When an ant reaches a decision point, it is more likely to choose the trail with the higher pheromone and heuristic values This recent and

new model had been applied in the data mining field, addressing both the clustering and classification. AntMiner+ is one algorithm in ACO and is a classification technique that induces rules

2.6 Model Bias, Variance and Imbalance Data

Prediction errors can be decomposed into two main sub-components; bias and variance. Bias refers to the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. The variance is how much the predictions for a given point vary between different realisations of the model.

With a small training sample, the model bias will increase while variance decreases. As training sample size grows towards infinity, model's bias will fall to 0 and variance will be no worse than any other potential model.

When more parameters are added to a model, errors due to bias decreases and errors due to variance increases because of the extra model complexity. The optimal spot for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance. If our model complexity exceeds the optimal spot, we are in effect over-fitting our model while if our complexity falls, we are under-fitting the model. In practical and general terms, however, only the overall error is mitigated, not the specific decomposition.

There are common techniques to handle errors from bias and variance. A key to this process is the selection of an accurate error measure (*e.g.* Cohen Kappa or F-score). In addition, re-sampling based measures such as cross-validation should be preferred over theoretical measures such as Aikake's Information Criteria. Bagging and other re-sampling techniques can be used to reduce the variance in model predictions.

The problem of model bias and variance can be viewed in a different perspective by penalising misclassifications. By introducing misclassification cost, a model can pay more attention to the minority class. Often the han-

dling of class penalties or weights are specialised to the learning algorithm(*i.e.* penalised-SVM and penalised-LDA) but there are generic frameworks (*e.g.* Cost-Sensitive Classifier in WEKA). (Fortmann-Roe June 2012).

Lastly, another challenge is imbalanced data where class memberships of observations are significantly skewed. A way to offset is to create synthetic samples from the minority class to create more data of different distribution. The most popular of such algorithms is SMOTE (Synthetic Minority Over-sampling Technique).(Brownlee August 2015*b*).

Another way is to change the class membership of the instances close to the border line that separates classes. Prediction probability is rounded to nearest class membership and in binary classification, 50% is the cut-off value. This cut-off is lowered allowing more predictions for the minority class. (Liaw & Wiener 2002).

2.7 Model Performance Measures

Precision is the percentage of correctly predicted churns overall predictions ($TP/(TP+FP)$) while Recall measures predicted churns out of all observed churns ($TP/(TP+FN)$). An ideal model is one that scores high on both precision and recall. A classic accuracy measure is over-all prediction accuracy which is the ratio of correct classifications ($TN+TN$) over the population ($TP+TN+FP+FN$).

Some prediction models allow you to nominate more weight to misclassified churns (Type II error) over misclassified non-churners (Type I error). This misclassification costs in effect penalise one type of prediction misclassification over another (precision versus recall, FN versus FP). Depending on the business intent of modelling, one prediction accuracy measure may be preferred over others. In the context of this study for the case company, this will be discussed in Chapter 5. In cross-validation, available data is randomly partitioned into k mutually exclusive subsets or folds of equal size. Training and testing are performed k times. In iteration i, partition D_i is reserved as

the test set, and the remaining partitions are collectively used to train the model. A common setting for the number of folds is 10 that creates 10 pairs of training and test datasets (P., Tang & L. 2009).

Receiver Operating Characteristic (ROC) curve is a useful visual tool for comparing classification models. ROC visually shows the trade-off between TP and FP as its discrimination threshold is varied. The area under ROC (AUC) is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative example; the higher the AUC value, the better its prediction accuracy (Fawcett January 7 2003).

2.8 General Methodology and tools used

The general data mining methodology used in this study was CRISP-DM which is Cross Industry Standard Process for Data Mining (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer & Wirth 2000). In terms of the tool used, SQL Server 2012 was used for the initial phase of business understanding, data capture and preparation. R Language was the primary tool for data preparation, modelling and evaluation. To visualise results, Microsoft Excel tool and WEKA(Waikato Environment for Knowledge Analysis)(Bouckaert, Frank, Hall, Kirkby, Reutemann, Seewald & Scuse 2016) were used.

2.9 Conclusion

This chapters identified applicable and important aspects for consideration when developing the churn prediction model. The aspects presented came from 2 perspectives; home care services and prediction modelling.

As home care services are human interaction-intensive by its business nature, variables on customer interactions and feedback through various client channels are to be obtained whenever possible. The variables to be used are likely to change significantly in as much as the local home care industry is undergoing major changes affecting the data that can be captured and

considered in the future.

For feature selection, it is preferred to maintain the original meaning of variables. Doing so promotes understanding and interpretability of churn analysis and models unlike in feature extraction wherein new variables are created from existing ones and the original meaning is lost. As the stated aim of the study is only for an initial model and analysis, readability and interpretability are more important than prediction accuracy.

Furthermore, a combination of filter and embedded feature selection techniques will be employed to eliminate irrelevant/ redundant data and minimise data dimensions.The count of features and predictors are to be reduced to a manageable level for better interpretability in churn analysis.

For prediction modelling, the literature recommends the use of k fold cross-validation and ensemble techniques such as bagging and boosting. In addition, as client churn prediction typically uses imbalanced data, a strategy to adequately learn from the minority class will be considered. Sophisticated machine learners like Neural Net and SVM are deferred after the initial model had been built.

In the absence of comparative service provider information, it can arguably be claimed that the case company is a good representative for its industry to perform churn analysis and prediction modelling. Distinction along home service lines is blurred as services increasingly vary as providers cater to the needs of the aged across their lifespan.

Chapter 3

Literature Review

3.1 Introduction

Client churn prediction models successfully implemented and published are numerous, but the author did not find any model as applied to home-based care overseas or in Australia.

This chapter describes the binary churn prediction models applied for other industries and other studies on home-based care services closely related to client churn. Related studies are about home-care client health improvement, satisfaction and service utilisation. A churn analysis section follows that dwell on critical variables and churn features. The conclusion describes how the literature review influenced the development of churn analysis and model.

3.2 Applied Churn Prediction Model

Prediction models had been applied in many applications such as weather forecasting, fraud detection, risk mitigation etc. and widely used for large and well-established companies. Various statistical models had been used as churn classifiers in many industries, and the reader is referred to the previous chapter for background on the models described herein.

In an application for a telecommunications company described in (Guo-en & Wei-dong 2008), the prediction models considered were logistic regression, C4.5 decision tree, Support Vector Machine (SVM), Naive Bayes and Artificial Neural Net (ANN). The industry was characterised as collecting data from various business and operational support systems. Data collected was described as typically large scale and with high dimensionality, non-linearity, non-normality, time series, and has rare minority class on client churns. Factor analysis was used to grouped and reduced variables to half and results exhibited wide variances in variables employed. As model parameter used for SVM, several kernels were attempted and with Radial basis as kernel function, prediction accuracy was at 90%. SVM was selected as it claimed to be suitable for this type of data, has simple classification plane, strong generation ability and good fitting precision.

In another churn prediction modelling under this industry described in (Huang et al. 2015), Random Forest (RF) model, Gradient Boosted Model (GBM) and variants of linear regression were considered. With the company experiencing a churn rate of 9.2%, Big Data technologies were leveraged to collect and prepare a large dataset.RF model was chosen using AUC (area under ROC curve) as evaluation criteria. In examining class imbalance, several strategies were attempted including up-sampling, down-sampling and weighing strategies. Weighted-instance method, wherein higher weights are assigned to churners compared to non-churners, outperformed others.

In another telecommunications application described in (Tsai & Chen 2010), a churn prediction model for a specific product type: multimedia-on-demand (MOD) was developed. ANN was used for prediction modelling and C5.0 association rules for reducing data dimensions. Only the top 20% C5.0 rules based on importance (*i.e.* support, confidence) were extracted. The study compared prediction accuracies obtained with the C5.0 filter technique against the original unfiltered dataset as baseline. The result showed that using the hybrid models in series (*i.e.* C5.0 decision trees and ANN) yielded better prediction accuracy. In a further study by the same author described

in (Tsai & Lu 2009), ANN was again used but with another feature selection techniques, SOM (*i.e.* self-organising maps. It found that using the hybrid model of ANN for feature selection and another ANN for model building yielded best results. A single ANN model was found to be better than SOM-ANN hybrid.

In a different industry setting for a print media subscription service in Europe described in (Coussement & den Poel 2009), a model was built to predict if a client will renew his/her newspaper subscription. Subscription cycle was in 3 months interval, and the company was experiencing 18% churn rate. Other than customer socio-demographics internally obtained, emails were monitored for two concecutive cycles. The dataset was divided into 70% training and 30%testing. Model evaluation measures used were percentage correctly classified, area under curve (AUC) and top decile lifts. RF was chosen over Logistic Regression and SVM at 75% prediction accuracy. In a separate paper by the same author in (Coussement & den Poel 2008), SVM was further explored. The study reaffirmed RF as the better model but qualified that in some cases, SVM outperforms RF. When a model is trained using a dataset with a balanced distribution and with the right kernel parameter, SVM is better classifier. Both studies concluded that client/company interactions and emotion-related variables contribute more to model performance than monetary-related features.

In another setting described in (Ruiz-Gazen & Villa 2007), a prediction model of cloud formations leading to thunderstorms/lightning was presented. The focus of the study is on handing imbalance data, as weather events of this nature occur infrequently, and using data of dubious quality. Usual model performance measures were found to be unsuitable, and a new prediction accuracy was introduced; 'missed alarms' and 'threat score' which are conceptually close to precision and recall. The new measures can similarly be derived from the confusion matrix, and by adjusting the prediction probability threshold used in classifying an instance, a usual ROC graph was shown. In this study, it was claimed that more recent and popular models

like Random Forest, has no advantage over simpler models when data collected is highly imbalanced and of low quality. Under this scenario, Logistic Regression was selected for simplicity and interpretability in classifying cloud formations into convective or non-convective.

More state-of-the-art and recent rule induction techniques are discussed in (W. Verbeke 2011). One method, Ant Colony Optimization, mimics the behaviour of biological ant colonies. The other way extracts rules from SVM which is characterised as a black-box. These techniques, however, were attempted on publicly available data and not on a real industry project.

The churn prediction model found closest to this thesis is described in (Golmohammadi & Radnia 2016) about patient readmission to a hospital in the US. Like this thesis, the model was also a binary prediction as applied to home-care services. It predicts if a patient will be readmitted back into a hospital for the same or related condition within one year of being discharged. The patient is admitted from and returned to a home-based care agency for follow-up care and monitoring and in this sense, is closest to the target application domain. Patient-level attributes about laboratory, drug, surgical procedures, insurance claims, etc. were summarised and used. For modelling, ANN, Classification and Regression(C&R) and Chi-squared Automatic Interaction Detection (CHAID) models were used for analysis and prediction. In searching for recurring patterns of unnecessary hospital readmissions, a C5.0 algorithm was used. The best over-all prediction accuracy of 84.2% was obtained with ANN model.

Table 3.1: Client Churn Prediction Models reviewed

Title of Paper	Models and Techniques
Model of Customer Churn Prediction on Support Vector Machine	Factor Analysis, SVM
Telco Churn Prediction with Big Data	Logit, C4.5, SVM Naive Bayes, ANN
Churn prediction in subscription services: An application of SVM while comparing two parameter-selection techniques	SVM, Logit, RF
Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers	SVM, Logit, RF
Customer churn prediction by hybrid neural networks	SOM Cluster, ANN
Variable selection by association rules for customer churn prediction of multimedia on demand	C5.0, ANN
Storm Prediction: Logistic Regression vs RFfor Unbalanced Data	Logit, RF
Prediction modelling and pattern recognition for patient readmission	CART, CHAID, C5.0 , ANN
Building comprehensible customer churn prediction models with advance rule induction techniques	AntMiner (Ant Colony Optimization) ALBA (SVM)

3.3 Churn associated studies on home care services

As mentioned, the author was unable to find churn prediction models as applied to home care services locally or internationally. There are however related studies in client satisfaction and health improvement that can be considered for churn analysis and modelling in this target industry. It can reasonably be assumed that satisfied customers are happier customers least likely to churn but the author is unable to present an empirical study to support assumption.

In (Madigan & Curet 2006), a study on home care services was associated with a client's health improvement. Indicators used for health improvement were his/her reason for discharge and length of stay under an agency's home care. A positive outcome was associated with favourable discharge reasons (*i.e.* goals met, services no longer required) and others (*i.e.* died, transferred to nursing home, hospitalised) considered a negative. Variables used were activities of daily living (ADL) and instrumental activities of daily living

(IADL) the patient receives as indicated in his/her response to a survey. Examples of ADL are bathing, toilet uses and light housework and telephone use for IADL.

The patients were divided into three focus groups according to the prior history of specific conditions; chronic obstructive pulmonary disease (COPD), hip replacement, and heart failure. The data mining approach used was CART (Classification and Regression Trees). A separate CART model and analysis was applied done for each focus group. The highest prediction accuracies obtained were 66% for COPD patients, 71% for heart failure patients and 50% for hip replacement patients.

In (Mylod & Kaldenberg 2000), client satisfaction survey results were analysed to identify and explain home care patient satisfaction. The survey form consisted of 35 items with each item representing a specific home care experience. Two types of analysis were used; top-box and lower-box analysis. Top box analysis revealed dimensions of care that are winning client satisfaction and bottom box analysis identified those requiring attention and with the most significant potential for client satisfaction improvement. The two techniques are straight-forward and employ simple statistical function to categorised survey responses. Survey respondents were asked to rate his/her experience on every item through a Likert-type scale ranging from 1 (Poor) to 5 (Very Good). Top box considered the highest rating while bottom box the lowest. In examining variance in survey responses, patients were clustered into segments to discriminate by patient needs and preferences. For simple segmentation analysis, independent t-tests of dichotomous variables (*e.g.* male/female, yes/no) was performed. For variables with more than two outcomes, analysis of variance (ANOVA) was performed. For complex segmentation, CHAID was used for analysis using combinations of patient characteristics.

In (Geron et al. 2000), the design and development of a home care survey instrument was described. In formulating survey questions, common factor and correlation analysis were used to reveal dimensions of client sat-

isfaction. To test the validity of the survey instrument, regression analysis was performed wherein overall satisfaction scores were associated with home services score. In assessing test-retest reliability, correlation between client satisfaction dimension and service scores were done with Pearson correlation method.

Table 3.2 summarises prior studies on home care services.

Table 3.2: Churn associated studies on Home-based Care Services

Title of Paper	Models Used
A data mining approach in home healthcare: outcomes and service use	CART(Classification and Regression Tree)
Data mining techniques for patient satisfaction data in home care	Box Analysis, Segmentation, CHAID and ANOVA
The home care satisfaction measure: a self-centred approach to assessing the satisfaction of frail older adults with home care services	Correlation and Common Factor Analysis

The studies on home care services in the US, however, had limitations. All the studies were industry-wide that compared home care providers and not in the context of a single company. The subject of their studies were focused groups; by ethnicity, physical ailments, income levels and geographical region. The scope of services covered mostly nursing care services which is a subset of the broader home care services predominantly offered in Australia. And most importantly, the studies aimed to analyse and identify features towards health improvement, client satisfaction and home service utilisation.

3.4 Client Churn Analysis

It is worth noting the variables and features identified in churn prediction modelling covered in the previous section. Identified key variables fell mostly under a client service profile, historical events and time-related aggregations. For telecommunications, services with favourable discounts that customer registered in, were found critical. Historical events such as suspension of

service and time-related features like regency of complaints were predictors. Clients using a mobile service for an extended duration and under considerable customer care are likely to churn. These features can be categorised in general to recency, frequency and monetary value and reaffirmed the RFM framework prevalently used in marketing studies and extended to include emotionality related variables (eRFM). (Coussement & den Poel 2009). The same study emphasised that client interactions and emotions expressed by clients had more effect than age and socio-demographic variables. There is a strong correlation between emotionally-related words for both positive and negative emotions. The churn prediction models surveyed however are not for home care services.

For home care services, the data mining studies goal were to measure client satisfaction, health improvement and service utility performance. It was claimed, amongst other things, that improved client well-being and customer satisfaction have positive impacts on the duration of client relationship and leads to loyal customers not likely to churn(Madigan & Curet 2006). The home care studies thus identified key variables that can be considered as candidate predictors for churn analysis and predictive modelling and are hence discussed next.

The studies in (Madigan & Curet 2006) and (Golmohammadi & Radnia 2016) were with regards to health improvement using home-care discharge reason and hospital readmission as indicators. In general, patients aged 85 and above constituted the age group likely to remain longer under home care due to ill health. The exact age at cut-off depended on the health condition (*i.e.* COPD, heart failure, hip replacement) of the study group. The type of payment and ethnicity also had impacts. There were differential effects related to agency type (*i.e.* proprietary, not-for-profits, others) by condition although the length of stay was lower for hospital-based agencies. In predicting hospital readmission for home care clients, the factors identified were age, gender, number of previous medical prescriptions, and duration of the prior stay under hospital care. Crucial contributors were also the place

of service (*i.e.* home care agency, ambulance, hospital), the sum of the previous laboratory test, number of medical claims and number of unique medical providers.

The studies in (Mylod & Kaldenberg 2000) and in (Geron et al. 2000) were with regards to client satisfaction. The analysis revealed that nursing issues are ranked highest in patient satisfaction (*e.g.* friendliness of nurse, the technical skill of nurses, nurses' concern for privacy and comfort). Lowest marks were on billing /cost issues, responsiveness to a client request, the degree of family involvement in planning services. Duration under home-care and self-reported health ratings had the greatest ability to segment the client population. Older patients were found to be more satisfied, but the age at which the split occurs between older and younger patients differed depending on the segment being analysed. The results claimed that clients in the 65 79 age group, living with someone else, experienced one to three months of home-care service and self-rated their health favourably were more satisfied.

In (Geron et al. 2000), physical disability was negatively associated with client satisfaction. The study found that gender, age or ethnicity were not related to home care satisfaction. Responses to a survey were used in the study and were found to be biased due to significant social desirability effects. Social desirability refers to an inclination by clients to be accommodating and friendly to service providers that leads to over-reporting of client satisfaction.

3.5 Conclusion

The churn prediction models surveyed were applied to large and well-established companies that had large data repository of acceptable quality. In general, home-based care service providers are small to medium-size establishments with limited resources for data collection. Other studies covered in this chapter were for home-based care but not directly about churn analysis.

The studies however identified important variables in the context of their

respective problem domains and industries. Where possible, the defined variables were collected for this thesis, used for churn analysis and prediction modelling. In particular, client age, service profile, health condition etc. were included and analysed as shown in succeeding chapters.

The literature also identified prediction models, like Random Forest, C5.0 and techniques such as Pearson correlation, which were all employed in this study. In (Neslin, Gupta, Kamakura, Lu & Mason 2006), it is noted that Logistic Regression and tree approaches were proven to perform well and claimed to be good techniques to begin with by companies starting up a predictive modelling function. As the intent is for an initial model only, the author is inclined towards comprehensibility and interpretability in an initial model.

SVM is claimed to be superior as it captures non-linearity in the data and offers good generalisations but regarded as an incomprehensible black-box models (W. Verbeke 2011). The same holds true for ANN. Using hybrid techniques can unravel these black boxes but further adds complexity and is best considered for subsequent model enhancement.

In (Huang et al. 2015), it was claimed that prediction models is not important as features. Given a variety of features, most classifiers can achieve the same accuracy. Adding a new feature may enhance the predictive modelling more than changing to a better classifier.

Chapter 4

Data Description and Churn Analysis

4.1 Introduction

This chapter describes the data and churn definition used in this thesis. The definition is used to label clients as churn or non-churn, to compute the churn rate accordingly and was conceived only for the purpose of the study. The chapter proceeds with analysing client churns on various dimensions such as age, services and client satisfaction and concludes by summarising churn insights.

4.2 Churn Definition and Measure

Definition 4.1 (Churned Client) *Churned client is a previously active (existing) customer who discontinued enrollment to subscribed programs and services due to an unsatisfactory customer experience or unfavourable perception. The reason given upon discharge is used as indicator for the customer experience or perception.*

$$Churn\,label \Leftarrow function\ (Client\,Program\,Enrollment,\ Discharge\,Reason)$$

$$\Leftarrow \begin{cases} 1\ (\textbf{Churn})\ Discharge\,Reason \in R\ and \\ \quad Client\,Program\,Enrollment\,type \in P\ and \\ \quad Client\,Program(s)\,Enrollment = 0 \\ 0\ (\textbf{non Churn}) \end{cases}$$

$$where\ R\,is\ set\,of Discharge\ Reasons\ r_1, r_2 ... r_n,$$

$$P\,is\ a\ set\,of\ Programs\ tagged\ as\ 'Core'\ p_1, p_2 ... p_n\ and$$

$$Client\ Program\ Enrollment(s)\ is\ the\ remaining\ count\ after\ discharge \tag{4.1}$$

Equation 4.1 refers to a set of client discharge reasons. Not all discharge reasons are churn related. For example, *Dissatisfied with company* is considered while *Change in living arrangement* is not. Figure 4.4 shows the valid reasons considered for churns with applicable churn reasons emphasized.

Some client enrolled programs are not considered for monitoring. The program may have a fixed expiration term or is a one-off incident in business nature. For example, the government assisted transitional care from hospital is fixed for a duration of 12 weeks and will terminate regardless. *Ambulance service* on its own is on-call basis, and the program ends once service is performed. To differentiate, programs to include in churn labelling, a program is tagged as **Core** or **Non-core**. The 'core' is for core business services that are monitored for client churns and are predetermined. Figure 4.6 lists client programs and with the exception of *'TransPac'* (transitional care package), all are tagged as 'Core'.

A client can simultaneously enrol into multiple programs which can be a combination of core and non-core programs. Core programs take precedence when labelling a client as churned or exited for non-churn discharge reason. A program constitutes one or many home-based care services served by home

care workers (HCW) to a client's home address. The same distinct home service can be used by various programs. Figure 4.7 shows services affected by clients who churned.

In Strouse [1999], churn rate (a.k.a. attrition) is defined as the annual turn-over of the market base. Clients need to be segregated into active and non-active and periodically monitored over time in order to determine the churn rate. Yearly, the client count is shown in Equation 4.2. The churn rate formula is as shown in Equation 4.3.

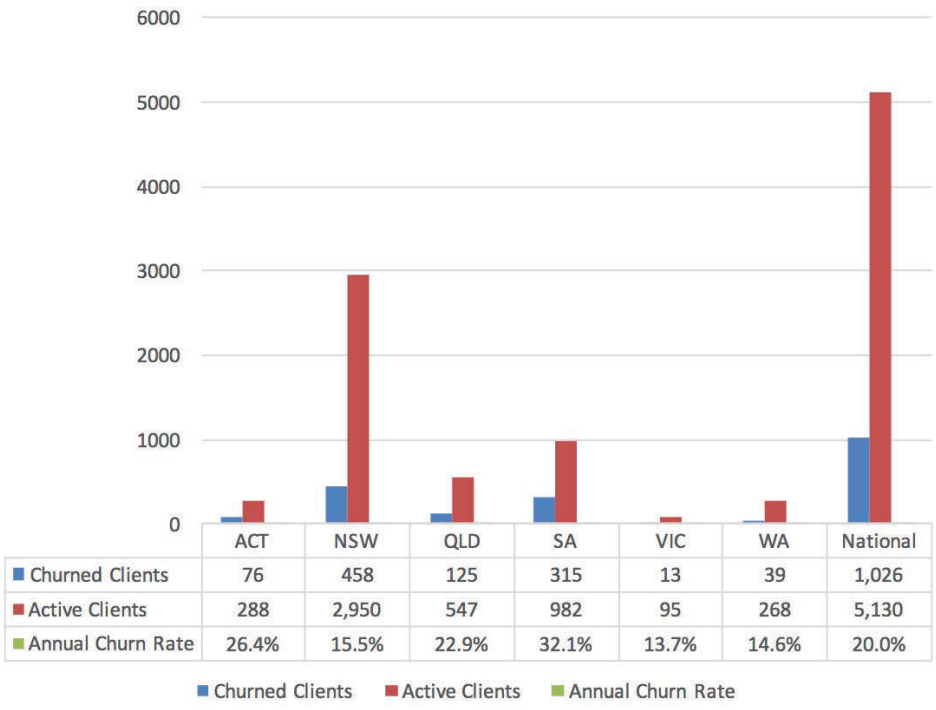
$$\begin{aligned} Active\ Clients_{(at\ end\, year)} = \sum Active\ Clients_{(at\ begin\, year)} \\ + \sum new\ Clients_{(enrolled\ within\ year)} \\ - \sum Churned\ Clients_{(discharged\ during\ year)} \\ - \sum Non-churn\ Client\ exits_{(discharged\ during\ year)} \end{aligned} \tag{4.2}$$

Definition 4.2 (Client Churn Rate) *Churn Rate is a percentage of active clients who churned over a predefined time period covering a company's operational region(s).*

$$\begin{aligned} Churn\ Rate\ \% = \sum_{s=1}^{S} \frac{Churned\ Client_s}{Active\ Clients_s} \\ where\ S\ is\ set\ of States\ s_1, s_2, ..., s_n \end{aligned} \tag{4.3}$$

With the definition of churned client and the formula applied, the churn rate for the case company is shown in Figure 4.1. At the national level, 20% of active clients churned between November 2014 to October 2015. The aggregation level is at each state as state managers are organisationally responsible for customer relationship and service delivery functions.

Notice that 60% of the clients reside in NSW where the company began and grew its home-based care operation. Note too that this is not a sample but the entire population of clients and about a fifth of which churned within the period.



	ACT	NSW	QLD	SA	VIC	WA	National
Churned Clients	76	458	125	315	13	39	1,026
Active Clients	288	2,950	547	982	95	268	5,130
Annual Churn Rate	26.4%	15.5%	22.9%	32.1%	13.7%	14.6%	20.0%

Figure 4.1: Annual Client Churn Rate

4.3 Data Collection and the Dataset

The data model of the attributes considered in this study is shown in Figure 4.2. The data was sourced from business support (*e.g.* client surveys) and operational support systems (*e.g.* data mart). We selected attributes about client demographics, client-expressed home care needs and service preferences. From operational customer data, we obtained a history of program enrollment, discharges, time sheets, complaints and interactions.. Appendix A lists all attributes, definitions and categories included in this study.

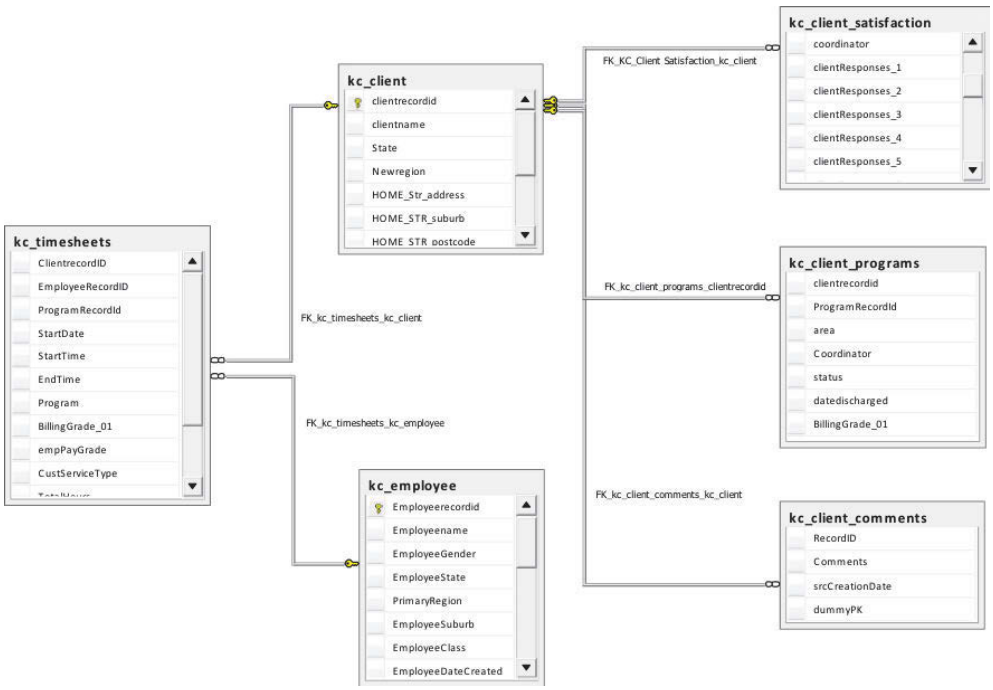


Figure 4.2: Source data Entity Relationship Diagram

The E.R.D. model in Figure 4.2 is useful in understanding the relationships amongst attributes especially in deriving new ones. In obtaining client service usage/interaction history, new attributes were introduced such as the count of *Issues_ raised* and *HCW_ratio* (*i.e.* a count of unique home care worker attending to the client). These attributes were aggregated at the client level for the period starting November 2014 to October 2015 termed as

the observation window as shown in Figure 5.1 described in the next chapter on modelling. Appendix A also includes derived new attributes.

Note that the dataset used in this study is the same as turned-over to the case company on completion of the industry project.

4.4 Data Cleansing

Prior to describing the data and calculating churn metrics shown in Figure 4.1, data collected needed to be cleansed. In the absence of a pre-existing definition and monitoring of client churns, each distinct client needed to be labelled as churned or not-churned.

- De-duplication of client IDs. It was found that some clients were identified by multiple client IDs due to the inconsistent recording of client names. In establishing uniqueness, a client's date of birth and full name were used to identify uniqueness. Client naming and recording conventions were standardised for all clients (*i.e.* all in the upper case recorded as last name, first name, and title). In so doing, the client accounts base was reduced by 18%.

- Define and formulate client churns. The definitions formulated and agreed are as described in Section 4.2 of this chapter.

- Label all unique clients as churned or non-churned at the start of the observation window, November 2014 according to the definition.

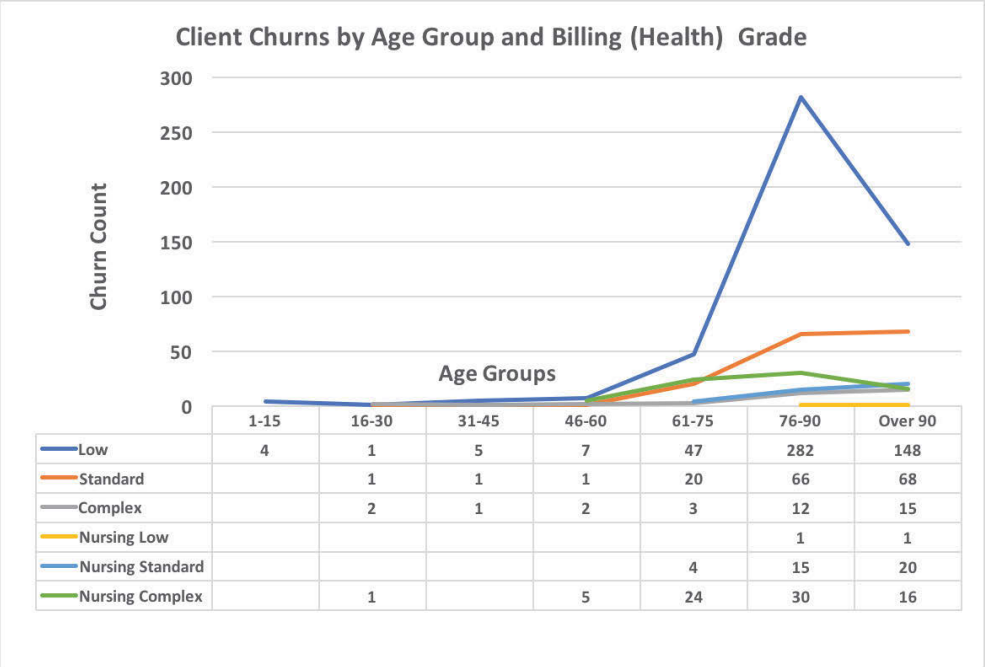
- Identify and aggregate attributes at client level for the entire period.

- Exclude inactive clients outside the observation window (*i.e.* Nov 2014). A client is considered inactive if he/she was not served during the window (*i.e.* no HCW timesheet record).

After cleansing, the descriptive analysis in terms of statistical measures such as average, variance, distribution, etc. is shown In Appendix B and Appendix C.

4.5 Churn Analysis in various dimensions

This section shows client churn in a various dimensions. Churns are viewed by client age groups, health grades (*i.e. MostUsedBillingGrade*), reasons for client discharge *discharge reasons*, programs/services affected and client satisfaction survey results. These dimensions were selected as literature surveyed described in Chapter 3 covered these attributes and subsequently identified as critical in prediction modelling described at next chapter. The data covered for churn analysis is the same for computing churn rates shown in Figure 4.1. The period covered again was November 2014 to October 2015 which included 5,130 unique clients across states.



	1-15	16-30	31-45	46-60	61-75	76-90	Over 90
Low	4	1	5	7	47	282	148
Standard		1	1	1	20	66	68
Complex		2	1	2	3	12	15
Nursing Low						1	1
Nursing Standard					4	15	20
Nursing Complex		1		5	24	30	16

Figure 4.3: Churns by Age Group and Health (aka Billing) Grade

Client churns occurred mostly in the 76-90 age group and under the *Nursing standard* grade and over. Relative to each age group's population, the highest churn rate was for 1-15 age group at 34%. Billing Grade is a category of level of home-care service associated with clients disability. *Nursing standard* and *Nursing Complex* care have significantly more churn percentages

at 79.9% and 46.3% respectively.Other socio-demographic client attributes had too many missing values (*e.g. Client Ethnicity* 77%, *Gender* 53%) to be considered in the analysis. The discharge reasons associated with churns are shown in Figure 4.4 and corresponding discharge sub-reasons shown in Figure 4.5.

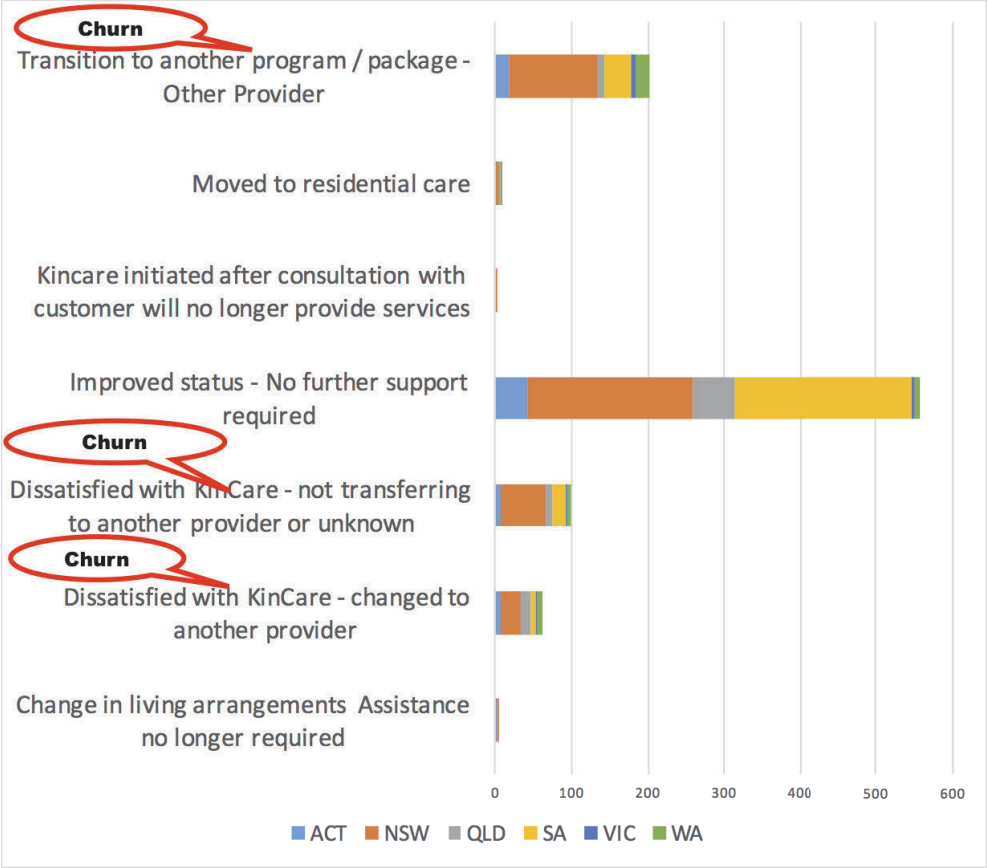


Figure 4.4: Client Discharge Reasons and Churns

Discharge reasons shown in Figure 4.4 include all client reasons and encircled reasons are associated with churns. The discharge sub-reasons shown in Figure 4.5 provide greater detail on the cause of churn. The *quality of care received*, *continued access to home care worker* and *timeliness/reliability of services* are top discharge sub-reasons provided by churned clients upon exit. Affected Programs by churned clients are shown in Figure 4.6 and associated

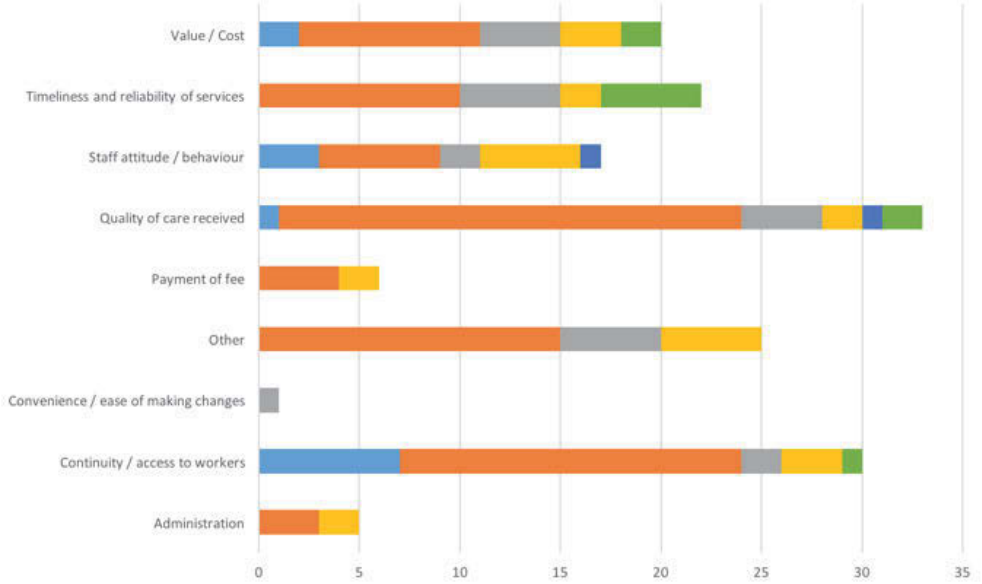
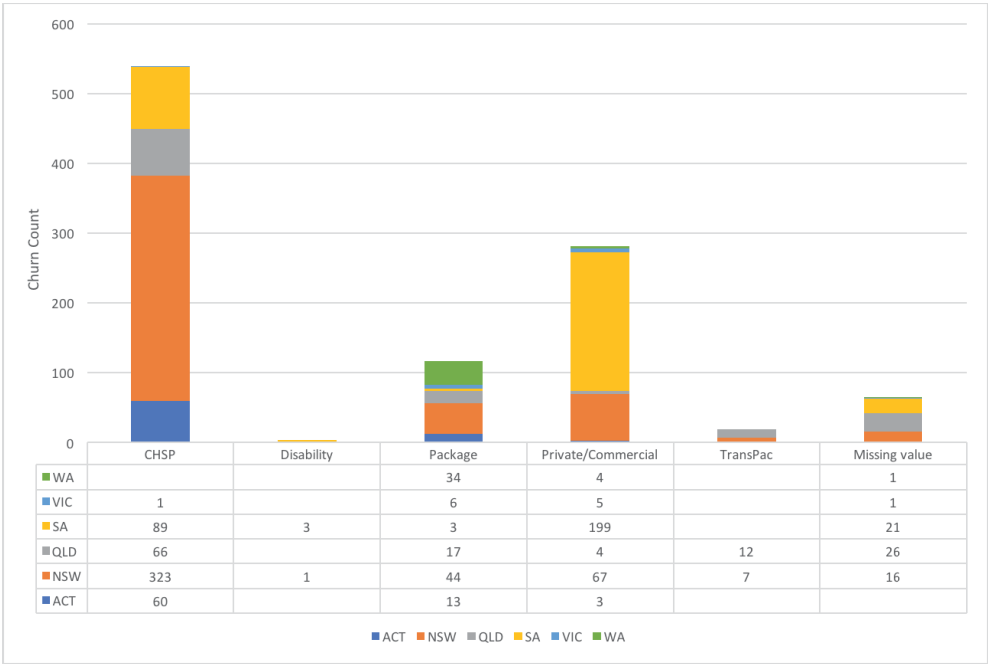


Figure 4.5: Client Discharge Subreasons and Churns

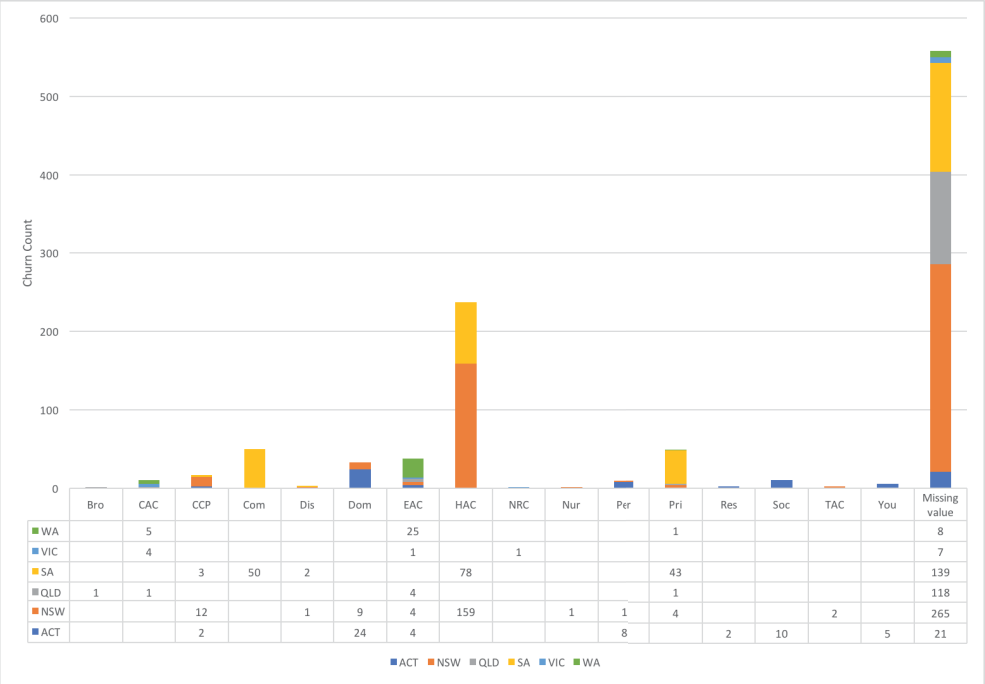
services shown in Figure 4.7.



	CHSP	Disability	Package	Private/Commercial	TransPac	Missing value
WA			34	4		1
VIC	1		6	5		1
SA	89	3	3	199		21
QLD	66		17	4	12	26
NSW	323	1	44	67	7	16
ACT	60		13	3		

Figure 4.6: Client Program enrolments and Churns

Like Discharge Reasons, some programs are excluded in churn monitoring as described in the previous section. Only programs tagged as *Core* business are monitored for churns. In Figure 4.6, churned clients belong mostly in programs under *Private/Commercial* and *CHSP* (Commonwealth Home Support Program). Significant churns occurred at *SA* for *Private/Commercial* and at *NSW* for *CHSP*. Services where churns occurred most were in *HAC* (Home and Community Care), *EAC* (Elderly Accommodation Counsel), *DOM* (Domestic Assistance) and *PRI* (Private services) as shown in Figure 4.7.



	Bro	CAC	CCP	Com	Dis	Dom	EAC	HAC	NRC	Nur	Per	Pri	Res	Soc	TAC	You	Missing value
WA		5					25					1					8
VIC		4					1		1								7
SA			3	50	2			78				43					139
QLD	1	1					4					1					118
NSW			12		1	9	4	159		1	1	4			2		265
ACT			2			24	4				8		2	10		5	21

Figure 4.7: Client Program Services and Churns

For the case company, client satisfaction is measured monthly via survey. Clients are randomly sampled to participate in a phone survey. There are ten questions in the form and a client responds to each on a Likert-type scale ranging from 1 (poor) to 5 (very good). The questions and client responses are shown in Figure 4.8 for the same observation window of November 2014 and October 2015.

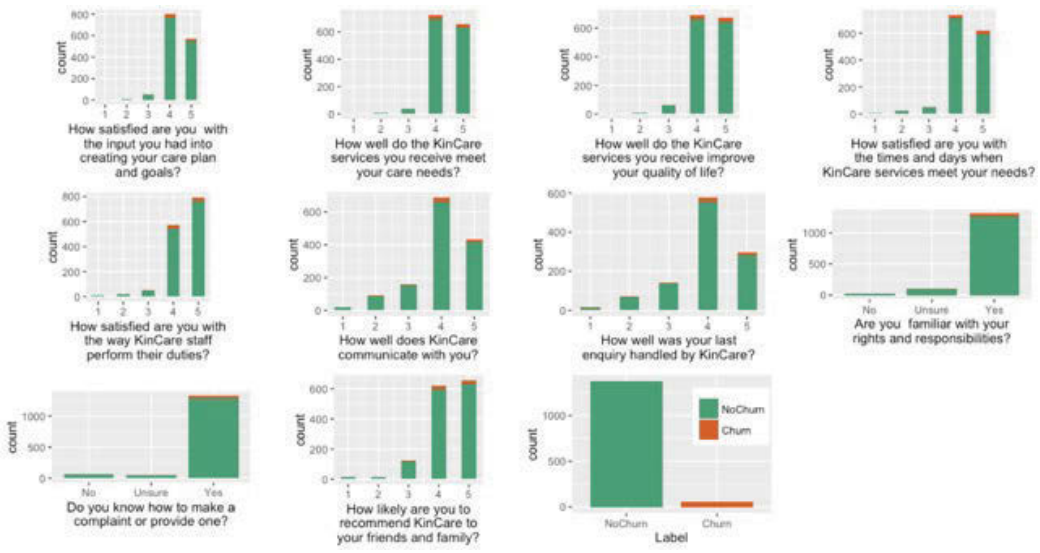


Figure 4.8: Client Satisfaction Survey Responses and Churns

When relating the responses to churns, it was observed that none of the client satisfaction survey responses is related to churns. It is important to note that the clients randomly surveyed represented less than 1.5% of the population. It is for this reason that the survey responses cannot be used as features in modelling.

The last survey question *How likely are you to recommend the company to your family and friends* is called the Net Promoter Score (NPS) (Reichheld 2006) and is the single measure to measure over-all client satisfaction. Using pair-wise correlation (core team 2017) between the NPS score and the Churn/No-churn target variable, the two are not significantly correlated. In (Geron et al. 2000) however, it is claimed that a single item global score to measure client satisfaction is not adequate in capturing the complexity of services in health, mental health or long-term care and instead proposed an alternative survey instrument in measuring client satisfaction.

4.6 Conclusion

From the descriptive churn analysis and data visualisations presented, churns were occurring mostly for clients aged 76 years and older. These clients required a higher grade of home service and were enrolled into the governments Commonwealth Home Support Program (CHSP), previously known as Home Assistance Community Care (HACC). They were churning because of the quality of home care services and continued access / promptness of homecare workers, most prevalent at NSW and SA. Highest churn rate was for under-15 age group who are presumably disabled or under private care.

To label a client as churn or not, discharge reason from enrolled program was used as indicator of a unfavourable outcome. This approach is the same as adopted in (Madigan & Curet 2006). The literature also mentioned limitations of using the reason for labelling such as the inherent bias in disclosing the true reason for churning and the ambiguity of the mutually exclusive discharge reasons in use. Note that in this thesis, *Others* was significantly used as discharge reason/sub-reason suggesting a need to refine and distinguish clearly.

As in literature surveyed, the analysis also showed the level of health or well-being of the patient is critical. This single attribute *Billing Grade* used in this study, however, is too loaded. It incorporated a clients welfare, physical disability, level of customer care required and paid for. This attribute should be decoupled and captured into many variables.

Chapter 5

Prediction Modelling

5.1 Introduction

In this chapter, development of the churn prediction model and test results are described in detail. It starts with an overview of the methodology and the selection of features as common set to be use by all candidate models. The candidate prediction models were the logistic regression(GLM), Random Forest (RF) and C5.0. Comparison of results using several criteria ensured selection of most appropriate model, C5.0 The chapter ends with tuning the model and a churn analysis from model results.

5.2 Model Development Methodology

It was claimed that clients past behaviour determines future outcomes (Coussement & den Poel 2009). For the purpose of capturing clients historical behaviour, available data was divided into two windows as shown in Figure 5.1.

The cut-off date chosen was November 17, 2014, as this was the date when the case company started to capture *discharge reasons*. This attribute is a key variable in labelling instances as churn or non-churns according to the definition in Equation 4.1.

The aim of dividing the data into two observation windows was to include

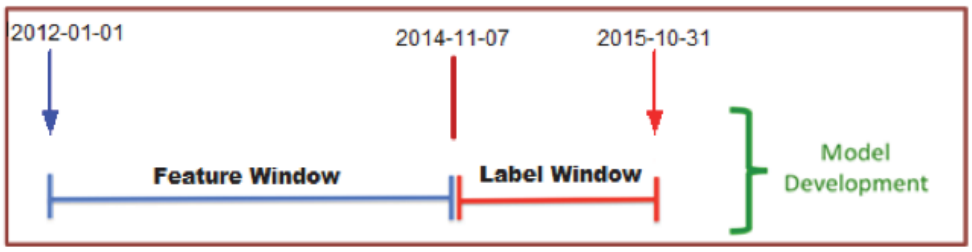


Figure 5.1: Model Development Observation Windows

behavioural variables in the immediately preceding two years to predict churn or no-churn for the following year. In the feature window, client historical attributes such as a count of *Issues_raised* and *HCW_Ratio* (distinct attending HCW person) were aggregated and included in the dataset.

Logistic regression and RF embedded feature selection algorithm were used to select candidate features. Highly correlated features were then filtered-out to come up with a set of features commonly used by all candidate prediction models.

All models generated churn prediction probabilities. The probability of each instance is rounded to nearest class membership of either churn (1) or non-churn (0) and threshold used was 50%. To visualise prediction accuracy with changing threshold values, Receiver Operating Characteristic (ROC) curve is shown and the area under the ROC computed.

In comparing prediction accuracies, the overall prediction accuracy, which measures all correct classifications over the population, was the main criteria used.

A summary of the methodology is shown on Table 5.1.

All steps in the development summary were included in a single program written in R Language and published `https://github.com/raulmanongdo/R-PredictionModel/blob/master/kc_client_churn_FINAL_G.md`. Model results presented in this paper are the same as published at GitHub.

Table 5.1: Model Development Summary

Dataset at Label Window	Train (4,130 obs.) & Test (1,000 obs.) with 74 variables
	Binary outcome variable (Churn or No Churn)
Feature Selection	Significant Logit vars. + RF important vars. - Pearson Correlated vars.
Candidate prediction models	Logistic Regression (Logit), Random Forest, C5.0
Prediction probability threshold	50% for all models and then
	to mean probability of (TP+TN) to adjusts for population bias
Model Comparison	Over-all prediction accuracy and others
	Area under ROC
Validation	10-fold cross validation (90%-10% training/test split)
Attempted Misclassification penalty	2 to 1 in favour of Type II(misclassified churns)

5.3 Data Preparation

The raw data used in this study is shown in Appendix B and Appendix C for categorical and numerical variables respectively. The dataset, which was previously cleansed for descriptive analysis as described in previous chapter, needed to be further prepared for prediction modelling. In data preparation, the following were undertaken:

- Variables with more than 50% missing values were ignored
- Missing numeric values were set to mean, negative values set to 0 or Null
- Missing categorical values were set to most commonly used value. Where applicable, sensible defaults were assigned (*e.g. SpokenLanguageRequired* set to *English*)
- Outliers were replaced with quartile-based boundaries (*i.e.* 3 times interquartile range)
- Normalisation was applied to numeric variables (regression only)

For the logistic regression model only, Z-score normalisation was performed. This normalisation technique set the range of numeric values to standard deviation with a variables mean value as zero (Han & Kamber 2006).

For tree-based models RF and C5.0 which are robust to outliers, normalisation was not performed. This choice was for the purpose of maintaining the original values of attributes used as split points of decision trees that led to better churn analysis and interpretation.

The results of the data preparation step performed on the population with 5,130 instances and 74 variables are summarised as follows:

- 31 variables had too many missing values and were dropped. Some of these were *ClientAge*, *Gender*, *EthnicGroupRequired* and *SpokenLanguageRequired* that had 50% missing values.

- Many missing values for remaining categorical attributes were set to most-used value.

  ClientType was set to *'HAC'* for 2,459 instances,

  default_contract_group was set to *'CHSP'* for 182 instances

- Significant mean imputation for numeric attributes were performed

- Many outliers existed and were replaced with quartile-based boundaries. An outlier is 1.5 times the interquartile range, defined accordingly as the difference between the 1st and 3rd quantile values. An outlier value is replaced with the limit defined as 1st or 3rd quantile value +/- inter-quantile range.

  AverageCoreProgramHours was reset 392 times to upper limit

  Issues_Raised was reset 177 times to upper limit

  Client_Initiated_Cancellations was reset 157 times to upper limit

  HCW_Ratio was reset 141 times to upper limit

  AgeAtCreation was reset 31 times to lower limit

- Many numeric attributes were extracted in error.

  FirstCoreServiceHours,*LastCoreServiceHours* had all zero values.

  CoreProgramsNums,*CoreRecordsRate*,*Issues_Requiring_Action*

Escalated_Issues had same single numeric value

It is worth noting that many attributes with irregularities mentioned above were critical attributes identified in literature review described in Chapter 3.

5.4 Feature Selection

For feature selection, the embedded technique was used wherein important and significant variables are identified by a selection algorithm innate to a pre-defined model. That is, Logistic regression and Random Forest model algorithms filter important attributes and the succeeding correlation analysis identfied irrelevant attributes for exclusion. Table 5.2 shows the final set of features using this approach.

Succeeding subsections describe how the feature under each model or technique were determined.

5.4.1 Significant variables in Logistic Regression

Significant variables using logistic regression to select features are shown in Table 5.3.

p-value (3rd column) is the measure of variable significance and is derived from standard error. The lower the value, the more significant and more asterisks (*) are shown for the variable. The relative contribution (i.e. log odds) of the independent variable to the dependent variable(*i.e.* churn/non-churn) is indicated by its coefficient (*i.e.* Estimate). A coefficient can be a positive or negative value indicating the direction of the association (*e.g.* positive coefficient leads to a churn). The higher the absolute value of the coefficient, the more discriminating is the variable. Common p-value threshold used is .05 which is equivalent to 95% confidence interval level of the coefficient value.

Table 5.2: Selected Features

HOME_STR_state	Client residential home state
ClientType	Type of client service usually performed
AgeAtCreation	Age at first program enrollment
MostUsedBillingGrade	Level of health care mostly used for billing
Client_Programs_count _at_observation_cutoff	Count of subscribed programs at start of label window
default_contract_group	Contract group of client's enrolled program
Client_Initiated_Cancellations	Count of client initiated cancellations of scheduled services
AverageCoreProgramHours	Average service hours of subscribed program monitored for Churn
HCW_Ratio	Number of distinct attending home care workers at start of label window
Issues_Raised	Count of issues raised at start of label window
PCNeedsFlag	Personal Needs Care (Yes/No)

Table 5.3: Logistic Regression significant variables

	Estimate	$Pr(>\|z\|)$	Signif.
ClientTypeYou	4.4516080	3.084973e-03	**
MostUsedBillingGrade.L	0.9074320	4.072909e-04	***
FrequentschedStatusGroupCancelled	0.6448600	3.036731e-02	*
Issues_Raised	0.4031573	1.482545e-16	***
default_contract_groupPrivate/Commercial	0.3977963	7.502689e-03	**
PCNeedsFlagY	0.2815010	3.345663e-02	*
Client_Programs_count_at_observation_cutoff	0.1484453	7.854017e-03	**
MinCoreProgramHours	0.1203954	2.865223e-02	*
MaxCoreProgramHours	-0.2197452	2.671882e-03	**
Client_Initiated_Cancellations	-0.2415312	5.610230e-03	*
AgeAtCreation	-0.3542742	7.664414e-17	***
RespiteNeedsFlagY	-0.4273024	3.742558e-02	*
HOME_STR_stateNSW	-0.5914624	1.272792e-02	*
MostUsedBillingGrade.C	-0.6602870	4.304246e-02	*
HOME_STR_stateVIC	-1.1414711	1.011449e-02	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note that some variables in Table 5.3 included a value (e.g. *MostUsedBillingGrade='Low'*, *ClientType='You'*). This inclusion is shown for all categorical variables (*i.e.* ordered and unordered factors in R).

5.4.2 Important variables in Random Forest

In RF, there are two basic types of measure for variable importance; importance by decreasing mean accuracy and Gini index. Figure 5.2 shows available variable importance measures as applied to the training set.

The aim is to select the least number of variables that mostly discriminates the dependent variable. From Figure 5.2, it was observed that the most discriminating measure is accuracy for the negative class (non-churn) or for both classes. Since the non-churn class is of less interest, the former measure is eliminated.

In RF, variable importance measure is computed from observations not selected from RFs bootstrap sampling (out-of-bag) algorithm. The RF looks at how much prediction error increases when data for that variable is permuted

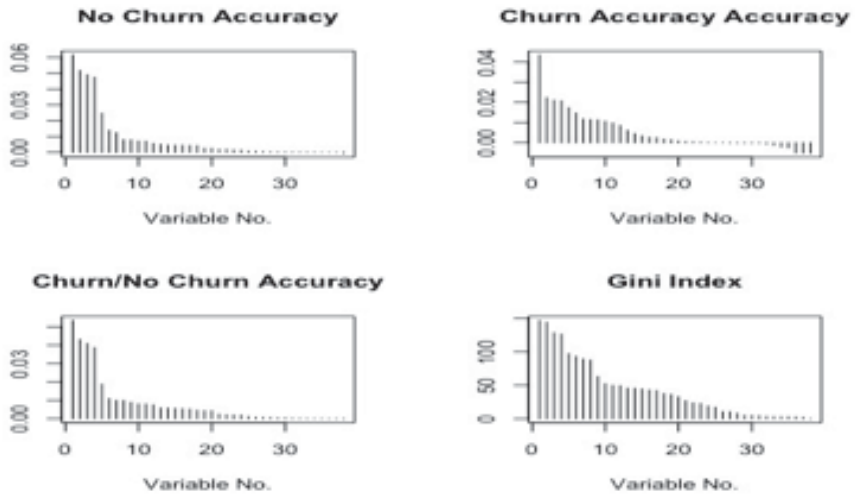


Figure 5.2: Variable importance measures in RF

while all others are left unchanged (Liaw & Wiener 2002). A permutation based measure is claimed to be superior for determining variable importance than Gini index (Strobl & Zeileis 2008).

Using this measure, the important variables using RF are shown in Table 5.4.

The *mTry* model parameter used is the default for classification (*i.e.* $\sqrt{N}$ where N is the number of variables to randomly select used for splitting tree nodes. The number of trees in the forest (*nTree*) to be generated is typically set to a high number as recommended (Liaw & Wiener 2002) and for this case, was set to 1,000.

The top 15 variables shown at Table 5.4 were selected using RF. The other variables not shown in this table but included in the set of predictors in Table 5.2 came from the GLM model.

5.4.3 Reduced Dimensions using Correlation Analysis

With the combined variables from RF and GLM, correlation analysis was then performed to come up with the final set of predictor variables shown in

Table 5.4: RF variables ranked by Accuracy

Variable	Mean Decrease in Accuracy
TotalCoreProgramHours	4.76E-02
AllRecordsNums	4.18E-02
CoreRecordNums	3.73E-02
AverageCoreProgramHours	3.66E-02
CoreTotalKM	2.02E-02
HCW_Ratio	1.24E-02
HOME_STR_state	9.58E-03
AgeAtCreation	9.53E-03
MaxCoreProgramHours	8.76E-03
Client_Programs_count_at_observation_cutoff	8.76E-03
AverageCoreServiceHours	8.25E-03
MostUsedBillingGrade	7.85E-03
Canned_Appointments	6.69E-03
default_contract_group	6.53E-03
Closed_Issues	6.30E-03
Issues_Raised	6.12E-03
ClientType	6.05E-03
Client_Initiated_Cancellations	5.68E-03
MinCoreProgramHours	4.92E-03
MostUsedPayGrade	4.29E-03
FirstCoreServiceDelayDays	2.69E-03
SmokerAccepted	2.46E-03
PCNeedsFlag	2.35E-03
Grade	1.57E-03
Company_Initiated_Cancellations	1.42E-03
DANeedsFlag	1.18E-03
TransportNeedsFlag	5.89E-04
SocialNeedsFlag	5.23E-04
NCNeedsFlag	4.59E-04
GenderRequired	3.76E-04
FrequentschedStatusGroup	3.67E-04
RespiteNeedsFlag	2.99E-04
RequiredWorkersFlag	9.39E-05
PreferredWorkersFlag	6.54E-05

Table 5.2. This step is to further reduce the dimension of variables by removing highly correlated variables which are redundant and irrelevant. Removing highly correlated variables aid towards a more concise interpretation of model results. Use of this technique also fulfils the requirement in regression for independence and non-collinearity amongst independent variables.

Pearson correlation (National Institute of Health Sep 2012) was used which is the default method in R. The cut-off value of +/ .80 was chosen for this application domain. The cut-off value is arbitrary and other applications, such as in health clinical data, requires a higher threshold value. Figure 5.3 shows the correlation map.

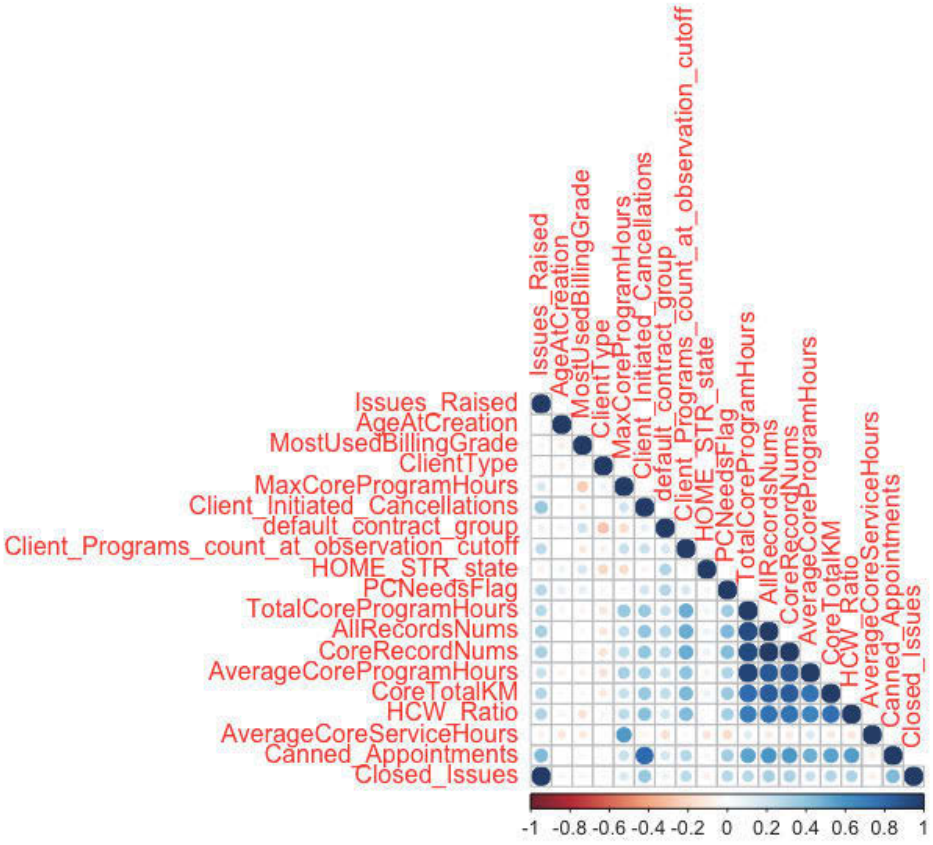


Figure 5.3: Feature-to-feature Correlation Analysis

The darker the colour of a cell, the higher the correlation between the pair

of variables intersecting at that cell. Hence, for a pair of highly correlated variables, one attribute in the pair must be excluded from the set of candidate predictors. This step identified and eliminated around 10 variables. Appendix D shows the Pearson correlation matrix in texts format used for Figure 5.3.

5.5 Candidate Prediction Models in Training

The three candidate prediction models were chosen for the following reasons:

- Logistic regression is known as robust technique and had been in existence for a long time. It is widely used in marketing studies usually as a baseline for comparison with other models
- C5.0 and Random Forests are ensemble classifiers which use new data mining techniques bagging and boosting, proven to improve prediction accuracy. C5.0 is the latest of successively improved versions of tree-based algorithms
- C5.0 and Random Forests are tree-based models that generate decision trees and business rules that are easy to understand and interpret
- Literature surveyed in Chapter 3 utilised C5.0 for studies on home-care services and RF for churn prediction

The candidate models have different ways to measure variable importance and model accuracy. GLM computes this by way of coefficients and error rates; RF by way of error-rates from its out-of-bag instances and C5.0 by way of frequency of correct prediction over misclassifications. The succeeding sections describe the candidate prediction models and the results from training.

As mentioned, the intent of this thesis is more comprehensibility and interpretability and hence, SVM and ANN were not attempted. Although

SVM and ANN are claimed to be superior, it is regarded as an incomprehensible black-box models (W. Verbeke 2011) and may impede churn analysis and model interpretation.

5.5.1 Logistic Regression

The GLM package in Cran R (core team 2017) was used for logistic regression on the training dataset and the result is shown in Table 5.5. Regression model algorithm looks for linearity between the independent and dependent (response) variables.

The variable coefficients, standard errors and p-Value are calculated with Wald z-statistic technique that measures the significance of independent variables to the dependent variable (core team 2017). Coefficients can be expressed as log odds and computed using GLM logit function. To convert to a relative odd of a churn, the exponent function is applied to the coefficient. The computed odds were included as the last column in Table 5.5. The significance of a predictor variable is visually denoted by one or more asterisks. A single asterisk denotes 95% significance level, two asterisks is 99% significance level and three asterisks, 99.9% significance level.

Table 5.6 shows insights deduced from the GLM model.

The logistic regression model is shown below.

```
Call:
 glm(formula = Label ~ ., family = binomial(link = "logit"),
      data = train.scaled)

 Deviance Residuals:
 Min       1Q   Median       3Q      Max
 -2.2030  -0.6639  -0.4741  -0.1776   3.3285

 (Dispersion parameter for binomial family taken to be 1)
 ..
```

Table 5.5: Standardised Logistic Regression Coefficients

	Estimate	Std. Error	$Pr(>\|z\|)$	Significance	Odds
(Intercept)	0.4337	145.78538	0.997626		
Issues_Raised	0.34927	0.05183	1.60E-11	***	1.42
AgeAtCreation	-0.36461	0.04645	4.17E15	***	0.69
MostUsedBillingGrade.L	0.93084	0.19525	1.87E-06	***	2.54
MostUsedBillingGrade.Q	-7.24166	445.36776	0.987027		
MostUsedBillingGrade.C	-0.26828	0.20914	0.199576		
MostUsedBillingGrade.4	7.59609	493.3888	0.987716		
MostUsedBillingGrade.5	0.58078	0.1972	0.003228	**	1.79
MostUsedBillingGrade.6	-10.04761	671.41712	0.98806		
ClientTypeCAC	1.68353	1.21489	0.165823		
ClientTypeCCP	1.10212	1.16913	0.345839		
ClientTypeCom	-0.16902	1.13103	0.881206		
ClientTypeDem	-13.3878	350.41535	0.969524		
ClientTypeDis	-0.59841	1.44835	0.679484		
ClientTypeDom	1.21376	1.15942	0.295158		
ClientTypeDVA	-3.40964	1725.34418	0.998423		
ClientTypeEAC	1.53377	1.15688	0.184913		
ClientTypeHAC	0.94026	1.12027	0.40129		
ClientTypeNRC	-11.87389	390.89921	0.975767		
ClientTypeNur	-1.02369	1.62273	0.528142		
ClientTypePer	0.61148	1.22969	0.619005		
ClientTypePri	0.92657	1.15184	0.421151		
ClientTypeRes	-0.06105	1.61001	0.96975		
ClientTypeSoc	1.95938	1.30036	0.131862		
ClientTypeTAC	0.85389	1.57873	0.588598		
ClientTypeTCP	-11.19081	905.04904	0.990135		
ClientTypeVHC	-11.81507	926.6227	0.989827		
ClientTypeYou	4.16097	1.49018	0.005234	**	64.13
Client_Initiated_Cancellations	-0.36419	0.06843	1.03E-07	***	0.69
default_contract_groupDisability	0.90829	0.88461	0.30453		
default_contract_groupDVA/VHC	-11.41443	926.62244	0.990172		
default_contract_groupPackage	-0.04747	0.16953	0.779483		
default_contract_groupPrivate/Commercial	0.40155	0.14958	0.007265	**	1.49
default_contract_groupTransPac	-0.70167	0.37191	0.059204	.	
Client_Programs_count_at_observation_cutoff	0.09586	0.05459	0.079069	.	
HOME_STR_stateNSW	-0.65864	0.25812	0.01072	*	0.52
HOME_STR_stateQLD	-0.14533	0.2854	0.610607		
HOME_STR_stateSA	-0.07075	0.28029	0.800722		
HOME_STR_stateVIC	-1.25828	0.47128	0.007587	**	0.28
HOME_STR_stateWA	-0.39709	0.36099	0.271331		
PCNeedsFlagY	0.18885	0.14038	0.178552		
AverageCoreProgramHours	-0.72001	0.09461	2.74E-14	***	0.49
HCW_Ratio	-0.3221	0.09466	0.000667	***	0.72

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5.6: Logistic Regression model insights

For *ClientType*, the odds to churn of a client under *'You'*(young, below 25 years) is 64.13 more than a *Client Type ='Bro'*
For every increase in the level of *MostUsedBillingGrade* the odds to churn is 2.54 more than *'BGrade5'* (Nursing Standard) and 1.79 times more than a client under *'BGrade1'* (Low care)
For *default_contract_group*, the odds a client enrolled in *'Private and Commercial'* is 1.49 times more than a client under *'CHSP'*
For every unit increase in *Issues_Raised*, the odds of churning increases by 1.418
For clients in *NSW* and *'VIC'* , the odds to churn were .52 and .28 when compared to clients at *'ACT'*
For every distinct additional *HCW_ratio* (*i.e.* new attending HCW), the odds of churning increases by .28
For every unit increase in *AverageCoreProgramHours*, the odds of churning increases by .49

```
Null deviance: 4133.3  on 4129  degrees of freedom
Residual deviance: 3477.8  on 4087  degrees of freedom
AIC: 3563.8

Number of Fisher Scoring iterations: 14
```

The null deviance measures how well the model predicts the value of the dependent variable using only the intercept while the residual deviance measures the variance from the null model. A higher deviance denotes badness-of-fit and the result showed that the predictor variables were critical and that a null model is to be rejected. Residual deviance reduced by 655.5 with a loss of 42 degrees of freedom.

The AIC which is calculated based on deviance and is used to assess the quality of the model in comparison with each iteration of the GLM model. Model iterations stop when no improvement in AIC can be attained. The AIC score for the above logistic regression model was 3563.8. The Fisher score stated the model needs 14 iterations to fit data into a model.

The GLM model in training yielded an over-all prediction accuracy of 81.62%.

```
    fitted.results.glm
       0      1
 0 3199  105
 1  654  172
```

To ascertain, the p-value of less than 0.001 below tells us that our model was statistically and significantly better than the null model.

```
> with(kc.glm, pchisq(null.deviance - deviance,
            df.null - df.residual, lower.tail = FALSE))
 [1] 3.995819e-111
```

5.5.2 Random Forest

For Random Forest, we followed the suggestion to use a large number of trees (*i.e.* nTree= 1,000) as model parameter. The same value was used as model parameter in feature selection. This parameter sets the number of trees in the forest constructed by the bootstrap sampling algorithm. The sample is equal in size to the original dataset but selected with replacement (Breiman 2001).

In constructing tree nodes, two variables were used to split each node (*mTry* model parameter). This value was determined by the algorithm itself (*i.e.* RFTune) as opposed to using the default in feature selection. (*i.e.* square root of the total number of variables).

The model result is shown below:

```
Call:
randomForest(formula = Label ~ ., data = train, mtry = RFmtry_param,
     ntree = RFnTree_param, keep.forest = TRUE, importance = TRUE)
Type of random forest: classification
Number of trees: 1000
No. of variables tried at each split: 2

OOB estimate of  error rate: 17.7%
Confusion matrix:
0   1 class.error
0 3164 140  0.04237288
1  591 235  0.71549637
```

In RF, error rates are computed from the observations not sampled, for all trees in the forest and is termed as Out-of-Bag. It is an average of prediction accuracies for all the 1,000 trees generated. The models prediction accuracy was 82.3% (*i.e.* 1 - out-of-bag error rate of 17.7%) which was close tp the accuracy of 83.1%. when later tested.

The important variables are as shown in Figure 5.4.

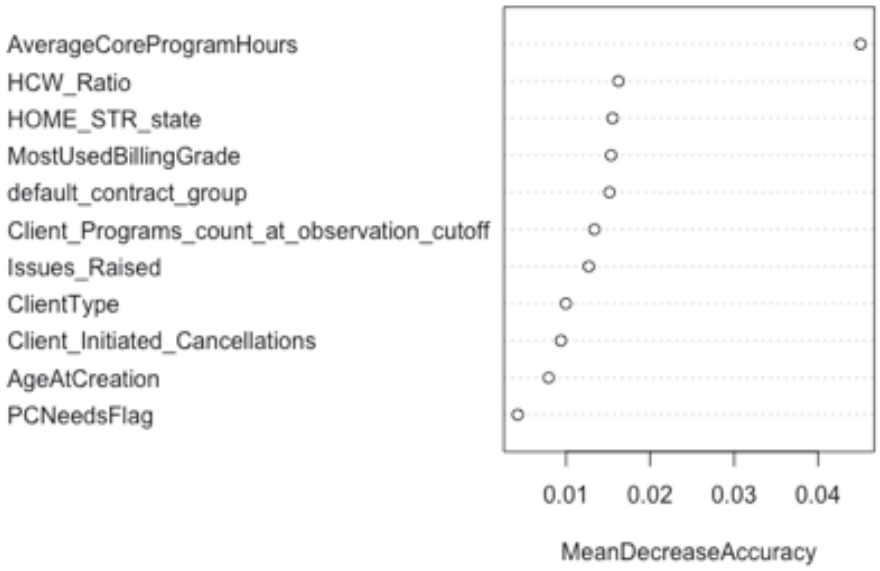


Figure 5.4: RF model variable importance by decrease in accuracy

As previously discussed, the variable importance measure chosen is mean decrease in accuracy for both churn and non-churn classes. Variables with a large mean decrease in accuracy are more important. The most important variable in the RF model was *AverageCoreProgramHours*. Its mean decrease in accuracy of 4.5% was significantly larger than other variables as shown in Figure 5.4. For a population size of 4,130 clients, dropping only this variable equates to an additional 186 misclassified instances on top of 591 already misclassified. Other variables were around the same degree of importance and were less critical.

Shown below are two non-churn rules extracted from the first generated RF tree.

```
Client_Initiated_Cancellations > 2.62096444626073   &
    MostUsedBillingGrade  > 3.5  &
    Client_Programs_count_at_observation_cutoff < 0.840036378334681
```

```
    => Nonchurn

ClientType > 1038323  &
    Client_Programs_count_at_observation_cutoff > 0.340036378334681
     & Client_Initiated_Cancellations < 2.62096444626073
     &  MostUsedBillingGrade > 3.5 &
    Client_Programs_count_at_observation_cutoff < 0.840036378334681
     => Nonchurn
```

Note that *ClientType* shown above is numeric and needs to be translated to its corresponding categorical value. The translation is done through a formula and is not shown in this paper.

More importantly, due to the algorithm in RF, the decision rules generated in the tree cannot be merged. This is primarily due to random selection of variables in building tree trunks and branches which made the induced rule an expert only of its limited domain. Consequently, the generated rules cannot cannot be merged with other rules in order to represent the population.

5.5.3 C5.0 model

The model parameter *Trials* was initially set to 10 which instructs the model to generate 10 boosted iterations in succession in building the tree. The model results are shown below

```
Call:
 C5.0.default(x = train[-ndxLabel], y = train$Label, trials
  = C5.0Trials_param, rules = FALSE,
  control = C5.0Control(earlyStopping  = TRUE))

 Classification Tree
 Number of samples: 4130
 Number of predictors: 11
```

```
 Number of boosting iterations: 10
 Average tree size: 17.6
...
 boost             641(15.5%)   <<

     (a)   (b)    <-classified as
    ----  ----
    3191   113    (a): class 0
     528   298    (b): class 1
     Trial        Decision Tree
      -----      ----------------
         Size      Errors

         0     25  671(16.2%)
         1     11  788(19.1%)
         2     15  946(22.9%)
         3     22  938(22.7%)
         4     15 1026(24.8%)
         5     15  881(21.3%)
         6     18  920(22.3%)
         7     20  823(19.9%)
         8     21  734(17.8%)
         9     14  745(18.0%)
      boost            641(15.5%)   <<

          (a)   (b)    <-classified as
         ----  ----
         3191   113    (a): class 0
          528   298    (b): class 1
```

```
Attribute usage:

100.00% AgeAtCreation
100.00% MostUsedBillingGrade
100.00% ClientType
100.00% Client_Initiated_Cancellations
100.00% AverageCoreProgramHours
 99.98% Issues_Raised
 99.98% HCW_Ratio
 67.92% HOME_STR_state
 67.80% Client_Programs_count_at_observation_cutoff
 66.80% default_contract_group
 46.97% PCNeedsFlag
```

The number of actual iterations can be less than the model parameter (*i.e.Trials*) if no further increase in accuracy can be attained. For the C5.0 model, it proceeded to the full 10 iterations to achieved the minimum boosting error rate of 15.5%. After 10 iterations, 641 out of population size of 4,130 were misclassified (528 misclassified churns and 113 misclassified non-churns).

The average tree size on all 10 boosted trees was 17.6, Tree size in C5.0 is defined as the number of leaf (terminal) nodes. (Research March 2017)

The model summary identified 7 attributes with the highest usage. Usage is based on how often the attribute was used as split points in tree constructions. These were *AgeAtCreation, MostUsedBillingGrade, ClientType, Client_Initiated_Cancellations, AverageCoreProgramHours, Issues_Raised* and *HCW_Ratio*.

C5.0 model can extracts business rules from its decision trees. All these rules are shown in Appendix E.

A rules performance is given a notation of the form n/m where n is the number of training cases correctly covered by the rule and m are cases incorrectly classified by the rule. The rule's accuracy is estimated by Laplace

ratio defined as $(n - m + 1)/(n + 2)$. The liftis the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set (Research March 2017).

A rule is of the form *if-then-else* and concludes to a class label; churn or no-churn. (*i.e.* TP or TN) Ranked by lift scores, the top 8 churn rules is shown in Table 5.7.

Table 5.7: Top C5.0 churn decision rules ranked by accuracy

1) DefaultContractGroup='Disability' & Clienttype='HAC' & AvgCoreProgamHrs<=46
2) 9<AvgCoreProgamHrs<= 24 & HCW<10 & AgeCreation>75 & Clienttype[Bro, CCP]
3) AgeAtCreation<=88 & AvgCoreProgramHrs<=18.5 & DefaultContractGroup='Disability' & BillingGrade[BG1, BG2, BG3] & Clientinitcancel <= 3
4) AvgCoreProgamHrs > 9 & HCW<10 & AgeCreation<= 75 & DefaultContractGroup= 'CHSP'
5) AgeAtCreation<=84 & Client type='Dom' & AvgCoreProgramHrs <= 17.9 & BillingGrade[BG1-BG3]
6) AvgCoreProgramHrs <= 37.5 & Clientinitcancel <= 3 & BillingGrade [BG4-BG9] & State [Act/QLD/SA] & AgeAtCreation<=71
7) AvgCoreProgramHrs <= 18.5 & Clientinitcancel<= 3 & BillingGrade='BG3' & State [Act/QLD/SA/NSW/VIC] & AgeAtCreation<= 88 & DefaultContractGroup [CHSP,Priv/Commercial] & Issues_raised >0.77
8) 6.8< AvgCoreProgamHrs <= 43 & Issues_raise > 0.77 & AgeCreation <= 61

Least common denominators of split points used in rule conditions can be deduced from Table 5.7. These split points can serve as alarms for home-care agents to take extra care and attention to clients reaching the threshold

values.

- *ClientAgeAtCreation* <= 65 and *DefaultContractGroup* = CHSP
- *ClientAgeAtCreation* between 75 and 83 and *HCW_ratio*(count of distinct home care worker < 2
- *AvgCoreProgramHours* < 19 and *DefaultContractGroup* = Disability

Business rules can alternately be obtained by passing instead a different model parameter (rules = TRUE). The generated rules with this parameter enabled is different in that it not extracted from a hierarchical tree structure. The rules are expressed in the same *if-then-esle* format but the rule set consists of unordered collections of rules that simply associates to churn or no-churn outcome. The rule sets obtained with the rule parameter enabled is shown in R Markdown document published at GitHub and not shown in this paper.

5.6 Model Comparison and Evaluation

Poor model performance is caused by over-fitting or under-fitting the data and becomes more evident when a model is tested using unseen data. Comparing training set, where the models are learnt, against the testing dataset, there was an average change of 1.483% in prediction accuracies for all models.

Various prediction accuracy metrics were used in this study such as precision, recall, F score and AUC. In general, resampling-based measures such as cross-validation should be preferred over theoretical measures such as Akaike's Information Criteria.The main evaluation criterion used was overall prediction accuracy as yielded during cross-validation. This measure is the percentage of all correctly predictions from total population (*i.e.* $(TP + TN)/(TP + TN + FP + FN)$).

In testing, the dataset was divided into 80% training and 20% testing. The candidate models were further tested and compared using combined datasets used in 10-fold cross-validation.

The modelling does not consider imbalance in the minority class (churns), penalties for misclassified predictions and trade-off between precision and recall.This will be discussed more in the next sections.

The model comparison of various prediction accuracies as applied to three datasets is shown in Table 5.8.

Table 5.8: Comparison of Prediction Model Performances

	Train	Test					10-fold	
Population	4,130	1,000					Cross Validation	
	Accuracy	Accuracy	Precision	Recall	F-score	AUC	Accuracy	AUC
Logit	.8162	0.832	0.7353	0.250	0.3731	0.7993	0.8162	0.7594
RF	.8447	0.831	0.8039	0.205	0.3267	0.7817	0.8259	0.7822
C50	.8230	0.838	0.9524	0.200	0.3306	0.7793	0.8283	0.7772

C5.0 model yielded the highest accuracy on testing at 83.8% and 82.83% on cross-validation. This value was close to its training accuracy of 82.3% suggesting a possible modest under-fitting. RF, however, outperformed others in training at 84.47%.

The ROC curves of candidate models under 10-fold cross validation are shown at Figure 5.5. When comparing AUC in 10-fold validation, RF outperformed others at 78.22% and followed by C5.0. An ROC curve can be drawn for the testing dataset but with significantly less number of observations, the ROC curve was less smooth and is not shown in this paper.

A ROC curve measures the performance of a binary classifier as its discrimination threshold. With the ROC curves shown, the use of 50% threshold value to round prediction probabilities to nearest class membership was well justified.

It is important to note that AUC, the area under a ROC curve, shows trade-off of correct and wrong predictions for the positive class (Churn) only. Predictions for labelled negative class (non-churn) was not considered.

The prediction accuracies of candidate models were close and the variances between candidate models needed to be investigated. A test for model

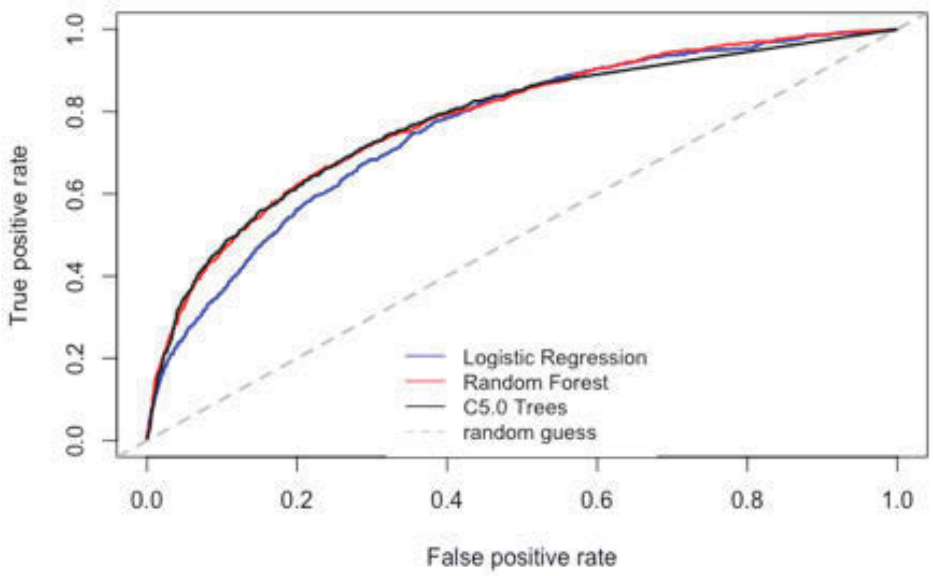


Figure 5.5: Comparison of Model AUC on 10-fold validation datasets

significance is shown in Table 5.9.

Table 5.9: Pair-wise comparison of model significance (AUC)

	p-Value Test	p-Value 10-X Test	95% C.I. on Test?	95% C.I. on 10-fold X Val
Logit to RF	0.2912	3e-04	No	Yes
Logit to C50	0.2074	0.0046	No	Yes
RF to C50	0.8668	0.304	No	No

The default method (Delong) in R was used for significance testing (core team 2017). At 95% confidence level, Random Forest and C5.0 were both better than Logistic Regression. Between Random Forest and C5.0, the difference was not statistically significant. With AUC as performance measure,

RF was a better model.

5.7 Selected model and tuning parameters

C5.0 model was selected as it yielded the highest overall prediction accuracy in the test environments. It was simpler and generates business rules that are easy to understand and interpret by any user. As C5.0 rules are *If-then-else* statements, it does not require any computing aid to interpret and apply.

In this sense, C5.0 was suitable for companies in this industry that are characterised as having fewer resources to develop automated business processes. For the case company, the *If-then-else* rules readily integrate to the business rule engine currently in use.

RF model also yielded high prediction accuracy but unlike C5.0, RF trees cannot represent the population as previously discussed. The complete list of rules extracted from the C5.0 model is shown at Appendix E. The top 8 churn rules ranked by accuracy is shown in Table in 5.7.

To improve the prediction accuracy of the selected mode, two approaches are adjusting for bias in the population and introducing misclassification penalties. Both approaches entail new model parameters to be introduced and the model retrained. Another way is to adopt a different accuracy measure (*i.e.* F-score, recall) which in effect views the problem from a different perspective. These approaches were attempted and are discussed below.

Although 1 out 5 active clients had churned, the entire population was used in this study, not a subset and thus the dataset provided a complete representation. For purposes of computing prediction accuracy, adjusting for bias may yield better accuracy rating but is not needed. Also, C5.0 model inherently adjusts for misclassification in its algorithm for boosting (model iterations). In addition, the case company prefered churn analysis and insights and hence no further work was expended in improving prediction accuracy. For the same reason, the trade-of between precision versus recall and determining the optimum balance from a business standpoint was

not covered. These two technqiues however were actually implemented post model selection and explained below.

When determining class membership from an instances prediction probability, the usual and default threshold value used is 50%. This value may not be appropriate when the training data used is significantly imbalanced. A way to adjust is to lower the threshold that reclassifies instances close to the threshold value and retrain the model. This was attempted in this study and the threshold was replaced with a mean probability of all correctly classified churns(TP): 25%. With this adjustment alone, recall performance measure doubled at the expense of reducing precision and over-all accuracy.

Another way to improve the accuracy is to introduce misclassification costs. When an instance is wrongly predicted, a user can factor in a misclassification penalty that gets used for the next model iteration and accuracy computation. Type 1 error is for a non-churned client classified as churned (FP) and Type 2 error is for churned client predicted as non-churned (FN). This was attempted in this study using a 2 to 1 weight for Type 2 misclassification cost. The misclassification cost in effect favoured recall over precision as a prediction accuracy measure.

Lastly, a different accuracy measure can be used by giving a different weight to prediction precision and recall. Precision ($TP/(TP + FP)$) is based on the predictions while recall ($TP/(TP + FN)$) measures correct predictions over observed churns, not predicted churns. The choice to give more weigh in one measure over the other depends on the business objective of the prediction modelling. If for example, the aim is to get a measure of how much operating budget is to be allocated for client retention campaigns, the more appropriate accuracy measure is Recall. If the aim, however, is identifying high-risk clients as the target of limited special offers, precision is more appropriate. A weighted measure of the precision and recall is F-score and commonly used are F_2 and $F_{0.5}$; the former weighs recall twice more than precision accuracy (by placing more emphasis on false positives) and the latter weighs precision over recall (by attenuating the influence of false

negatives). (Van Rijsbergen 1979)

The result of the three approaches to handle imbalance data and misclassification penalties is described in Table 5.10.

Table 5.10: C5.0 model parameter tuning

	Accuracy	Precision	Recall	F-score
No tuning	0.838	0.9524	0.200	0.3306
Probability threshold set at 25%	0.812	0.5405	0.400	0.4598
Type 2 misclassification penalty of 2 to 1	0.829	0.8372	0.180	0.2963

5.8 Churn Model Analysis and Insights

As the prediction accuracies were not far apart, this section describes model insights from all prediction models

The three models have top five common predictors; namely *MostUsedBillingGrade*, *AgeAtCreation*, *ClientType*, *AverageCoreProgramHours* and *Issues_Raised*. As to how each predictor affects churn outcome, the author refers to GLM's Table 5.5 and RF's Figure 5.4. For the GLM model, separate insights are shown n Table 5.6.

The predictors can be classified into two for practical use; attributes known upon client enrolment to a home-care program and attributes historically accumulated about services and client interactions.

On client socio-demographics, most churners were enrolled in CHSP (Commonwealth Home Support Package) previously called Home Assistance Community Care(HACC) . These clients were mostly frail or aged and under a higher level of home care service (*i.e.* Nursing care and above). Churned clients were mostly between 75 to 85 years old. The age of cut-off varies

according to other variables such as disability and services required (*e.g.* personal care).For every incremental increase in age, a client is less likely (-.069) to churn and remain as a client. Private and Commercial clients were also at-risk as presumably, they do not avail of government support and are more critical of quality rendered such as in domestic and private services. Clients in the under-25 age group were most at-risk. Clients from the states of QLD and SA were found to more likely to churn.

Through time, a customer accumulates service and interaction history and key predictors were issues raised and client-initiated service cancellations. For every issue raised, the odds of churning increases by around 1.42 and for every service cancellations, by .69. For each additional hour of home service, a client is less likely to churn by a factor of .49. Contrary to popular belief, for every additional HCW attending to one client, the client will not churn (.72).

It is however observed that there were significant data quality issues as described in Section 5.3 in this chapter that affected this analysis.

5.9 Conclusion

C5.0 model was selected over logistic regression and Random Forest models as it yielded a marginally higher overall prediction accuracy and produced simple business rules that are easy to interpret and apply.

In feature selection, embedded models by way of using GLM and RF were used as this approach maintained meaning of attributes as opposed to feature extraction techniques wherein a variables original meaning will likely be lost.

The candidate models selected were popular and successfully used in many industries. As the aim of the study is to build an initial model, more sophisticated and recent prediction models were not attempted. More value can be attained by incorporating new features and addressing data quality issues.

Tuning on the model was attempted by introducing misclassification penal-

ties and adjusting for imbalance in the minority class (churn). In this thesis, the entire population, not a sample, was used and hence there was no bias. Also, prediction probability, which uses misclassification penalties in computing percentages, was of less concern to the case company.

In a major IT company-sponsored tournament in developing client churn systems participated by practitioners and academics alike, it was concluded that Logistic Regression and tree approaches perform well and were good techniques to begin with by companies starting up a predictive modelling function. Exploring several estimation techniques to develop one model may not pay off. (Neslin et al. 2006).

Chapter 6

Conclusion

6.1 Conclusion and Research Answers

This thesis developed a binary client churn prediction model as applied to home-based care services using real data from an anonymous company in this industry. Logistic regression, Random Forest and C5.0 were the candidate models attempted and all yielded over-all accuracy of over 83%. C5.0 prediction model attained a marginally higher prediction accuracy and was chosen because the induced model decision trees are easier to understand, interpret and more suitable for use by the case company.

The main predictors of client churning were found to be health grade (*i.e.* Billing Grade, Disability), enrolled programs/services, number of issues and complaints raised and client age. Most churners were enrolled in CHSP (Commonwealth Home Support Package) previously called Home Assistance Community Care(HACC) and who were at least at the standard Nursing level of care.. Private and Commercial and clients in the under-25 age group were also most at-risk. These findings are consistent with literature surveyed on home-care services overseas where age and health grade were identified emphasised.

The major challenge the author experienced was in data capture and data quality pertaining to client churns. Churn classification and reporting were

not existence in the case company and had to be formulated for this study only. It was also observed that associated client satisfaction measures were inadequate and retention strategies not existing.

Are the results of this study applicable to other companies in the home-care industry? It was claimed that there were 2,000 companies operating in this local industry with a diversity of service offerings (*i.e.* services only, retirement village, etc.) and business models (*e.g.* private, not-for-profit). Although the company is significantly larger than the multitude of small-to-medium size companies, further study is required if the company is typical and the results of this study are deemed applicable to companies in this industry in general. At best, the churn analysis will apply to similar programs and services offered operating under the same business model. At worst, data needs to be captured in the context of other companies and remodeled again.

The results of this study confirmed the claim made in (Neslin et al. 2006) that logistic regression and tree-based models are good prediction models, to begin with by a company.

6.2 Future Work

As stated, the aim is for an initial model only that can be subsequently enhanced. This can be achieved by expanding further the set of variables especially ones that are particular to home-based care services.

It is recommended that other than expanding the set of candidate variables as model features, refining churn definition and improving data quality be the next focus in enhancing the prediction model for the case company.

In the literature survey for home-based care services, many important variables were identified that were not currently being captured. The identified attributes can be included and have potential to become predictor variables.

- Has own personal care worker and his/her relation to client
- if client is living alone

- Self-rating of the clients health
- If client is a new user of home care services and if not, prior length of usage
- Source of payment for enrolled programs and services specially for non-government programs (CHSP).

Other successfully proven prediction models not attempted in this thesis can be tried out in the subsequent enhancement of this prediction model. Support Vector Machine (SVM) and Artificial Neural Network are more recent and popular models claimed to be successfully used as classifiers especially in the telecommunication industry. Business rules can be extracted from these two black-box models using a technique described in (W. Verbeke 2011) (Ant Miner+, ALBA).

For data preparation, home care programs and services can be further classified, The study in (Mylod & Kaldenberg 2000) and (Geron et al. 2000) described clustering of services into the home-maker, home health aide, care management, home-delivered meal and grocery. Dimensions of client satisfaction empirically identified were staff competency, system adequacy and dependency, positive and negative interpersonal client interaction and service convenience.

In (Coussement & den Poel 2009), service recovery activities arising from client complaints were considered and text mining was used to differentiate emotions in client interactions. With the rise in digital communications, other sources such as websites, blogs and twitter feeds can be used as additional sources of client feedback.

The case company can also include more variables under Recency, Frequencyand Monetary Value (RFM) scheme. For this thesis, there were no 'Monetary Value' features included. The author suggests that a segmentation analysis of customers be performed and segments profiled according to profitability, sustainability and potential life-time value.

Appendix A

Attributes

Attribute Name	Derived?	Description
myUniqueClientID	Y	unique client based on name and date of birth
Label	Y	churn/nochurn label
Category: Demographics		
HOME_STR_state		Ciient's home state
Sex		Gender
ClientAge	Y	Age of client upon program enrollment
AgeAtCreation	Y	Age of client upon first program enrollment
Category: Client Needs & Preferences		
SmokerAccepted		HCW smoker allowed
GenderRequired		HCW required gender
EthnicGroupRequired		HCW EthnicGroupRequired
SpokenLanguageRequired		HCW required ethnic group
RespiteNeedsFlag		Respite service needed for client's own carer
DANeedsFlag		Domestic Assistance
NCNeedsFlag		Nursing care
PCNeedsFlag		Personal care
SocialNeedsFlag		Social needs
TransportNeedsFlag		Transport needs
RequiredWorkersFlag		Specific HCW to serve
PreferredWorkersFlag		Preferred HCW to attend
Category: Client Service Profile		
ClientType		Client Type
Grade		Billing Grade, used to classify client's required level of care
MostUsedBillingGrade	Y	Most used Grade
MostUsedPayGrade	Y	Most used Pay Grade of attending HCW
Client_Programs_count_at _observation_cutoff	Y	count of subscribed programs prior to label window
default_contract_group	Y	contract group of the client's subscribed program
TotalDaysWithCompany_all_Programs	Y	number of days as client since initial program enrollment
HCW_Ratio	Y	number of distinct attending HCW prior to label window
Category: Client Interaction		
complainttier	Y	highest complaint escalation tier reached
Issues_Raised	Y	count of issues raised
Issues_Requiring_Action	Y	count of issues requiring action
Escalated_Issues	Y	count of issues escalated
Closed_Issues	Y	count of closed issues
Client_Initiated_Cancellations	Y	client initiated service cancellations

Company_Initiated_Cancellations	Y	Company initiated service cancellations
Canned_Appointments	Y	Cancelled service appointments
Category: Service Delivery		
AssDaysAfterCreation	Y	days between program subscription and assessment
AllRecordsNums	Y	A client's total service record times
CoreProgramsNums	Y	A client's total core programs
CoreRecordNums	Y	A client's total core services
TotalCoreProgramHours	Y	A client's total core programs service hours
MaxCoreProgramHours	Y	A client's max core programs service hours
MinCoreProgramHours	Y	A client's min core programs service hours
AverageCoreProgramHours	Y	A client's average core programs service hours
AverageCoreServiceHours	Y	(TotalCoreProgramHours/AllRecordsNums)
CoreRecordsRate	Y	CoreRecordNums/AllRecordsNums
CoreTotalKM	Y	A client's total KM run for programs tagged as core.
FirstCoreServiceDelayDays	Y	Start Date - Admitted Date
FirstCoreServiceHours	Y	The service hour for the first core service
LastCoreServiceHours	Y	The service hour for the last core service
NoneCoreProgramsNums	Y	A client's total none-core programs
NoneCoreRecordNums	Y	A client's total none-core services
TotalNoneCoreProgramHours	Y	A client's total none-core programs service hours
MaxNoneCoreProgramHours	Y	A client's max none-core programs service hours
MinNoneCoreProgramHours	Y	A client's min none-core programs service hours
AverageNoneCoreProgramHours	Y	A client's average none-core programs service hours
AverageNoneCoreServiceHours	Y	(Totalnone-coreProgramHours/AllRecordsNums)
NoneCoreRecordsRate	Y	none-coreRecordNums / AllRecordsNums
NoneCoreTotalKM	Y	A client's total none-core KM
FirstNoneCoreServiceDelayDays	Y	Start Date - Admitted Date
FirstNoneCoreServiceHours	Y	The service hour for the first none-core service
LastNoneCoreServiceHours	Y	The service hour for the last none-core service
FrequentschedStatusGroup	Y	The Most Frequent billingGrade
late_first_service_count_all_programs	Y	count of service late by more than 5 days from assessment
avg_first_service_days_all_programs	Y	average days before waiting for first service
Category: Client Satisfaction		
Responses_1		How satisfied are you with the input you had into creating your care plan and goals?
Responses_2		How well do the services you receive meet your care needs?
Responses_3		How well do the services you receive improve your quality of life?
Responses_4		How satisfied are you with the times and days when services met your needs?
Responses_5		How satisfied are you with the way agency staff perform their duties?
Responses_6		How well does communicate with you?
Responses_7		How well was your last enquiry handled?
Responses_8		Are you familiar with your rights and responsibilities?
Responses_9		Do you know how to make a complaint or provide one?
Responses_10		How likely are you to recommend to your friends and family? (NPS)

Appendix B

Summary of Raw Categorical Data

HOME_STR_state	Sex	ClientType	Grade	SmokerAccepted
ACT: 288	Female:1586	HAC :1379	Grade1:2829	No :2739
NSW:2950	Male : 784	EAC : 310	Grade2:100 5	Yes :2390
QLD: 547	NA's :2760	Com : 304	Grade3: 276	NA's: 1
SA : 982		CCP : 163	Grade4: 16	
VIC: 95		Dom : 136	Grade5: 55	
WA : 268		(Other): 379	Grade6: 164	
NA's :2459		NA's : 785		
FrequentschedStatusGroup	MostUsedBillingGrade	MostUsedPayGrade	RespiteNeedsFlag	DANeedsFlag
Active :4641	BGrade1:3030	PGrade2:2927	N :4500	N :1484
Cancelled : 58	BGrade2:1170	PGrade3:1130	Y : 448	Y :3464
Company_Initiated: 21	BGrade3: 5 06	PGrade1: 487	NA's: 182	NA's: 182
NA's : 410	BGrade4: 2	PGrade6: 208		
	BGrade5: 91	PGrade5: 108		
	BGrade6: 260	(Other): 6		
	BGrade9: 71	NA's : 264		
TransportNeedsFlag	RequiredWorkersFlag	PreferredWorkersFlag	default_contract_group	complaint_tier
N :4473	N :4733	N :4470	CHSP :2844	Tier1: 47
Y : 475	Y : 215	Y : 478	Disability : 34	Tier2: 18
NA's: 182	NA's: 182	NA's: 182	DVA/VHC : 3	Tier3: 5
			Package : 993	Tier4: 1
			Private/Commercial: 914	NA's :5059
			TransPac : 160	
			NA's :182	
SpokenLanguageRequired	SocialNeedsFlag	EthnicGroupRequired	PCNeedsFlag	Label
English:1288	N :3943	Australia: 954	N :3790	Churn :1026
Others : 43	Y :1005	Others : 112	Y :1158	NoChurn:4104
Arabic : 34	NA's: 182	Italy : 19	NA's: 182	
Italian: 33		Greece : 16		
Greek : 31		England : 12		
(Other): 50		(Other) : 21		
NA's :3651		NA's :3996		
GenderRequired	NCNeedsFlag			
Either:2405	N :4711			
Female:2619	Y : 237			
Male : 106	NA's: 182			

Appendix C

Summary of Raw Numerical Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
myUniqueClientID	5,130	17,068.530	9,788.399	105	10,039.2	22,334.8	48,723
AssDaysAfterCreation	350	16.503	29.644	0	0	20.8	89
AgeAtCreation	5,130	78.826	14.318	0	73	87	115
Responses_1	554	4.330	0.587	1	4	5	5
Responses_2	554	4.406	0.582	2	4	5	5
Responses_3	554	4.388	0.582	2	4	5	5
Responses_4	554	4.359	0.601	1	4	5	5
Responses_5	554	4.471	0.659	1	4	5	5
Responses_6	532	4.071	0.829	1	4	5	5
Responses_7	419	4.005	0.818	1	4	5	5
Responses_10	554	4.330	0.747	1	4	5	5
AllRecordsNums	5,130	138.365	410.444	1	9	100	11,064
CoreProgramsNums	5,130	1.309	0.637	1	1	1	7
CoreRecordNums	5,130	136.476	409.664	1	9	99	11,064
TotalCoreProgramHours	5,126	214.874	1,515.200	0.250	12.500	143.000	73,077.330
MaxCoreProgramHours	5,126	3.607	10.703	0.250	2.000	3.000	728.000
MinCoreProgramHours	5,126	0.787	0.965	−22.000	0.500	1.000	24.000
AverageCoreProgramHours	5,126	175.905	1,490.563	0.250	11.000	107.229	73,077.330
AverageCoreServiceHours	5,126	1.655	1.629	0.250	1.016	1.875	24.000
CoreRecordsRate	5,130	0.957	0.203	0	1	1	1
CoreTotalKM	5,130	634.027	1,925.491	0.000	22.990	442.857	44,144.310
FirstCoreServiceDelayDays	5,130	156.425	746.627	−995	0	8	32,153
FirstCoreServiceHours	5,130	0.000	0.000	0	0	0	0
LastCoreServiceHours	5,130	0.000	0.000	0	0	0	0
NoneCoreProgramsNums	222	1.027	0.163	1	1	1	2
NoneCoreRecordNums	222	43.658	44.798	1	6	65	195
TotalNoneCoreProgramHours	222	60.305	121.967	0.333	8.562	87.688	1,710.983
MaxNoneCoreProgramHours	222	2.988	3.510	0.333	1.500	3.000	24.000
MinNoneCoreProgramHours	222	1.102	2.787	0.000	0.500	1.000	24.000
AverageNoneCoreProgramHours	222	55.187	72.274	0.333	8.375	86.812	855.492
AverageNoneCoreServiceHours	222	1.701	2.811	0.333	1.000	1.500	24.000
NoneCoreRecordsRate	222	0.000	0.000	0	0	0	0
NoneCoreTotalKM	222	158.189	216.638	0.000	12.090	224.368	1,262.180
FirstNoneCoreServiceDelayDays	222	20.221	198.013	−20	0	3	2,905
FirstNoneCoreServiceHours	222	0.000	0.000	0	0	0	0
LastNoneCoreServiceHours	222	0.000	0.000	0	0	0	0
Client_Programs_count_at_observation_cutoff	4,948	0.680	0.805	0	0	1	8
TotalDaysWithProvider_all_Programs	2,273	459.538	673.760	0	83	518	5,939
Issues_Raised	4,894	0.767	1.416	0	0	1	33
Issues_Requiring_Action	4,894	0.044	0.258	0	0	0	4
Escalated_Issues	4,894	0.067	0.329	0	0	0	6
Closed_Issues	4,894	0.767	1.416	0	0	1	33
Client_Initiated_Cancellations	4,894	2.242	3.883	0	0	3	46
Company_Initiated_Cancellations	4,894	0.565	1.370	0	0	1	25
Canned_Appointments	4,894	3.237	6.812	0	0	3	129
late_first_service_count_all_programs	1,751	1.225	0.516	1	1	1	7
avg_first_service_days_all_programs	1,751	486.215	1,120.369	1	4	628	32,153
HCW_Ratio	5,130	9.532	12.376	0	2	12	133

Note : Missing values is number of observations N less population size of 5,130.

Appendix D

Correlation Matrix

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1 Issues_Raised	1.00000	0.04139	-0.05359	0.02143	0.16178	0.38683	0.07596	0.26856	-0.07407	0.28127
A2 AgeAtCreation	0.04139	1.00000	-0.08934	-0.07299	-0.02256	-0.02323	0.05736	0.01062	0.07115	0.10612
A3 MostUsedBillingGrade	-0.05359	-0.08934	1.00000	0.05542	-0.25038	-0.09569	0.20373	-0.11917	0.18038	0.11013
A4 ClientType	0.02143	-0.07299	0.05542	1.00000	0.07004	0.01334	-0.27324	-0.06885	-0.20027	-0.03428
A5 MaxCoreProgramHours	0.16178	-0.02256	-0.25038	0.07004	1.00000	0.20369	-0.17738	0.22795	-0.18072	0.03855
A6 Client_Initiated_Cancellations	0.38683	-0.02323	-0.09569	0.01334	0.20369	1.00000	0.10154	0.22004	-0.05790	0.18117
A7 default_contract_group	0.07596	0.05736	0.20373	-0.27324	-0.17738	0.10154	1.00000	0.16588	0.30838	0.29671
A8 Client_Programs_count _at_observation_cutoff	0.26856	0.01062	-0.11917	-0.06885	0.22795	0.22004	0.16588	1.00000	-0.02333	0.17444
A9 HOME_STR_state	-0.07407	0.07115	0.18038	-0.20027	-0.18072	-0.05790	0.30838	-0.02333	1.00000	0.05738
A10 PCNeedsFlag	0.28127	0.10612	0.11013	-0.03428	0.03855	0.18117	0.29671	0.17444	0.05738	1.00000
A11 TotalCoreProgramHours	0.28152	-0.04114	-0.05988	-0.12508	0.38149	0.37317	0.23964	0.50403	0.06034	0.35815
A12 AllRecordsNums	0.33770	0.02237	-0.03550	-0.12637	0.25277	0.40694	0.30261	0.50988	0.08249	0.45516
A13 CoreRecordNums	0.32643	0.01614	-0.03561	-0.13541	0.25386	0.39695	0.27845	0.50397	0.08889	0.43993
A14 AverageCoreProgramHours	0.22359	-0.05947	-0.06933	-0.14483	0.33241	0.33924	0.22765	0.40379	0.06690	0.30879
A15 CoreTotalKM	0.29003	0.02712	-0.03885	-0.11188	0.23314	0.38905	0.25267	0.45845	0.09696	0.37220
A16 HCW_Ratio	0.30826	0.04071	-0.14422	-0.04890	0.26002	0.42339	0.19526	0.44850	-0.00221	0.33590
A17 AverageCoreServiceHours	-0.07498	-0.11547	-0.11036	0.02909	0.57467	-0.05542	-0.16395	-0.02813	-0.13254	-0.18843
A18 Canned_Appointments	0.44010	0.00492	-0.02039	0.00864	0.22625	0.79422	0.17819	0.29448	-0.04683	0.34912
A19 Closed_Issues	1.00000	0.04139	-0.05359	0.02143	0.16178	0.38683	0.07596	0.26856	-0.07407	0.28127

	A11	A12	A13	A14	A15	A16	A17	A18	A19
A1 Issues_Raised	0.28152	0.33770	0.32643	0.22359	0.29003	0.30826	-0.07498	0.44010	1.00000
A2 AgeAtCreation	-0.04114	0.02237	0.01614	-0.05947	0.02712	0.04071	-0.11547	0.00492	0.04139
A3 MostUsedBillingGrade	-0.05988	-0.03550	-0.03561	-0.06933	-0.03885	-0.14422	-0.11036	-0.02039	-0.05359
A4 ClientType	-0.12508	-0.12637	-0.13541	-0.14483	-0.11188	-0.04890	0.02909	0.00864	0.02143
A5 MaxCoreProgramHours	0.38149	0.25277	0.25386	0.33241	0.23314	0.26002	0.57467	0.22625	0.16178
A6 Client_Initiated_Cancellations	0.37317	0.40694	0.39695	0.33924	0.38905	0.42339	-0.05542	0.79422	0.38683
A7 default_contract_group	0.23964	0.30261	0.27845	0.22765	0.25267	0.19526	-0.16395	0.17819	0.07596
A8 Client_Programs_count _at_observation_cutoff	0.50403	0.50988	0.50397	0.40379	0.45845	0.44850	-0.02813	0.29448	0.26856
A-9 HOME_STR_state	0.06034	0.08249	0.08889	0.06690	0.09696	-0.00221	-0.13254	-0.04683	-0.07407
A10 PCNeedsFlag	0.35815	0.45516	0.43993	0.30879	0.37220	0.33590	-0.18843	0.34912	0.28127
A11 TotalCoreProgramHours	1.00000	0.91641	0.92151	0.95139	0.79391	0.71689	0.13529	0.54748	0.28152
A12 AllRecordsNums	0.91641	1.00000	0.99531	0.86572	0.84832	0.75713	-0.11161	0.59721	0.33770
A13 CoreRecordNums	0.92151	0.99531	1.00000	0.87174	0.85340	0.75887	-0.11013	0.58674	0.32643
A14 AverageCoreProgramHours	0.95139	0.86572	0.87174	1.00000	0.75670	0.67966	0.15359	0.49545	0.22359
A15 CoreTotalKM	0.79391	0.84832	0.85340	0.75670	1.00000	0.77104	-0.11579	0.54169	0.29003
A16 HCW_Ratio	0.71689	0.75713	0.75887	0.67966	0.77104	1.00000	-0.12599	0.56423	0.30826
A17 AverageCoreServiceHours	0.13529	-0.11161	-0.11013	0.15359	-0.11579	-0.12599	1.00000	-0.08013	-0.07498
A18 Canned_Appointments	0.54748	0.59721	0.58674	0.49545	0.54169	0.56423	-0.08013	1.00000	0.44010
A19 Closed_Issues	0.28152	0.33770	0.32643	0.22359	0.29003	0.30826	-0.07498	0.44010	1.00000

Appendix E

C5.0 model Decision Rules

```
## C5.0 [Release 2.07 GPL Edition]      Sat Jun  3 22:32:26 2017
## -------------------------------
##
## Class specified by attribute ‘outcome’
##
## Read 4130 cases (12 attributes) from undefined.data
##
## -----  Trial 0:  -----
##
## Decision tree:
##
## MostUsedBillingGrade in [BGrade4-BGrade9]:
## :...HOME_STR_state in {ACT,NSW,VIC,WA}: 0 (129/28)
## :   HOME_STR_state in {QLD,SA}:
## :   :...MostUsedBillingGrade in [BGrade4-BGrade6]: 1 (184/50)
## :       MostUsedBillingGrade = BGrade9:
## :       :...Issues_Raised <= 0: 0 (6)
## :           Issues_Raised > 0:
## :           :...AverageCoreProgramHours <= 2.5: 0 (5)
## :               AverageCoreProgramHours > 2.5: 1 (33/8)
## MostUsedBillingGrade in [BGrade1-BGrade3]:
## :...AverageCoreProgramHours > 17.875: 0 (2689/297)
##     AverageCoreProgramHours <= 17.875:
##     :...AgeAtCreation > 84: 0 (342/62)
##         AgeAtCreation <= 84:
##         :...ClientType in {Bro,CAC,CCP,Com,Dem,Dis,DVA,NRC,Nur,Per,Pri,Res,Soc,
##             :              TCP,VHC,You}: 0 (50/18)
##             ClientType in {EAC,TAC}: 1 (14/3)
##             ClientType = Dom:
##             :...Client_Programs_count_at_observation_cutoff > 0: 1 (8/1)
##             :   Client_Programs_count_at_observation_cutoff <= 0:
```

```
##             :   :...AverageCoreProgramHours <= 12.875: 1 (15/5)
##             :       AverageCoreProgramHours > 12.875: 0 (10/1)
##             ClientType = HAC:
##             :...default_contract_group = Disability: 1 (1)
##                 default_contract_group in {DVA/VHC,TransPac}: 0 (10)
##                 default_contract_group in {CHSP,Package,Private/Commercial}:
##                 :...Issues_Raised > 0.7666531: [S1]
##                     Issues_Raised <= 0.7666531:
##                     :...AverageCoreProgramHours > 11.25: 0 (98/11)
##                         AverageCoreProgramHours <= 11.25:
##                         :...HOME_STR_state in {ACT,VIC,WA}: 0 (21/6)
##                             HOME_STR_state = QLD: 1 (53/23)
##                             HOME_STR_state = SA: [S2]
##                             HOME_STR_state = NSW: [S3]
##
## SubTree [S1]
##
## default_contract_group = Private/Commercial: 0 (12/2)
## default_contract_group in {CHSP,Package}:
## :...HOME_STR_state in {ACT,QLD,SA,WA}: 0 (38/14)
##     HOME_STR_state in {NSW,VIC}: 1 (129/58)
##
## SubTree [S2]
##
## Client_Programs_count_at_observation_cutoff <= 0: 0 (86/32)
## Client_Programs_count_at_observation_cutoff > 0: 1 (5/1)
##
## SubTree [S3]
##
## default_contract_group in {CHSP,Package}: 0 (157/41)
## default_contract_group = Private/Commercial:
## :...AgeAtCreation <= 36: 0 (6/1)
##     AgeAtCreation > 36: 1 (29/9)
##
## -----  Trial 1:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 37.5:
## :...ClientType in {Bro,CAC,CCP,Com,Dem,Dis,Dom,DVA,EAC,HAC,NRC,Nur,Per,Pri,Res,
## :   :              TAC,TCP,VHC}: 0 (1808.6/345.3)
## :   ClientType in {Soc,You}: 1 (22.3/4)
## AverageCoreProgramHours <= 37.5:
## :...Client_Initiated_Cancellations > 3: 0 (182.2/44.8)
##     Client_Initiated_Cancellations <= 3:
##     :...MostUsedBillingGrade in [BGrade2-BGrade9]:
##         :...AgeAtCreation > 73: 0 (625.4/277.9)
```

```
##         :   AgeAtCreation <= 73:
##         :   :...AgeAtCreation <= 58: 0 (78.1/31.6)
##         :       AgeAtCreation > 58: 1 (248.9/91.6)
##        MostUsedBillingGrade = BGrade1:
##        :...Issues_Raised > 0.7666531: 0 (449.1/209.4)
##            Issues_Raised <= 0.7666531:
##            :...PCNeedsFlag = Y: 1 (15.8/4.8)
##                PCNeedsFlag = N:
##                :...HCW_Ratio > 4: 0 (174.1/21.9)
##                    HCW_Ratio <= 4: [S1]
##
## SubTree [S1]
##
## Client_Programs_count_at_observation_cutoff <= 0.6800728: 0 (448.2/131)
## Client_Programs_count_at_observation_cutoff > 0.6800728: 1 (77.2/30)
##
## -----  Trial 2:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 43:
## :...Issues_Raised <= 0: 0 (699/107.2)
## :   Issues_Raised > 0:
## :   :...AgeAtCreation > 78: 0 (436.1/101.6)
## :       AgeAtCreation <= 78:
## :       :...Client_Initiated_Cancellations > 4: 0 (117.8/29.4)
## :           Client_Initiated_Cancellations <= 4:
## :           :...Client_Initiated_Cancellations <= 1: 1 (159.9/63.9)
## :               Client_Initiated_Cancellations > 1: 0 (169.5/74.3)
## AverageCoreProgramHours <= 43:
## :...AverageCoreProgramHours <= 6.833333:
##     :...default_contract_group in {Disability,DVA/VHC,TransPac}: 0 (7.8)
##     :   default_contract_group in {CHSP,Package,Private/Commercial}:
##     :   :...Issues_Raised > 0.7666531: 1 (153.1/47.1)
##     :       Issues_Raised <= 0.7666531:
##     :       :...HOME_STR_state in {ACT,QLD,SA}: 1 (506.6/199.8)
##     :           HOME_STR_state in {NSW,VIC,WA}: 0 (314/147.8)
##     AverageCoreProgramHours > 6.833333:
##     :...Issues_Raised <= 0.7666531:
##         :...PCNeedsFlag = N: 0 (730.6/215.3)
##         :   PCNeedsFlag = Y: 1 (92.7/41.6)
##         Issues_Raised > 0.7666531:
##         :...AgeAtCreation <= 61: 1 (30.1/4.8)
##             AgeAtCreation > 61:
##             :...Client_Initiated_Cancellations > 3: 0 (118.3/39.1)
##                 Client_Initiated_Cancellations <= 3:
##                 :...default_contract_group in {CHSP,Package}: 1 (540.5/229.9)
```

```
##                     default_contract_group in {Disability,DVA/VHC,
##                                                Private/Commercial,
##                                                TransPac}: 0 (54/17.3)
##
## -----  Trial 3:  -----
##
## Decision tree:
##
## AverageCoreProgramHours <= 9.666667:
## :...Client_Initiated_Cancellations > 2.241929: 0 (43.9/14)
## :   Client_Initiated_Cancellations <= 2.241929:
## :   :...default_contract_group in {Disability,
## :       :                          Private/Commercial}: 1 (388/170.8)
## :       default_contract_group in {DVA/VHC,Package,TransPac}: 0 (83.5/37.7)
## :       default_contract_group = CHSP:
## :       :...AgeAtCreation <= 84: 1 (531.9/237.8)
## :           AgeAtCreation > 84: 0 (213.9/89.8)
## AverageCoreProgramHours > 9.666667:
## :...HCW_Ratio > 41: 0 (64.8/3)
##     HCW_Ratio <= 41:
##     :...Issues_Raised > 0.7666531:
##         :...AgeAtCreation > 90: 0 (86.5/16.4)
##         :   AgeAtCreation <= 90:
##         :   :...HOME_STR_state in {ACT,WA}: 1 (213.6/89.7)
##         :       HOME_STR_state in {NSW,QLD,SA,VIC}: 0 (1191.2/482)
##         Issues_Raised <= 0.7666531:
##         :...Client_Initiated_Cancellations > 9: 0 (27.5)
##             Client_Initiated_Cancellations <= 9:
##             :...HCW_Ratio > 10:
##                 :...AgeAtCreation <= 42: 1 (17/5.9)
##                 :   AgeAtCreation > 42: 0 (258/29.1)
##                 HCW_Ratio <= 10:
##                 :...AgeAtCreation > 75:
##                     :...AverageCoreProgramHours > 24.125: 0 (374.1/69.1)
##                     :   AverageCoreProgramHours <= 24.125:
##                     :   :...ClientType in {Bro,CCP}: 1 (14.5/1.1)
##                     :       ClientType in {CAC,Com,Dem,Dis,Dom,DVA,EAC,HAC,NRC,
##                     :                      Nur,Per,Pri,Res,Soc,TAC,TCP,VHC,
##                     :                      You}: 0 (224.5/75.4)
##                     AgeAtCreation <= 75:
##                     :...default_contract_group in {Disability,
##                         :                          DVA/VHC}: 0 (6.5)
##                         default_contract_group = TransPac: 1 (5.7)
##                         default_contract_group in {CHSP,Package,
##                         :                          Private/Commercial}:
##                         :...MostUsedBillingGrade in [BGrade3-BGrade9]: 0 (44.7/9.6)
##                             MostUsedBillingGrade in [BGrade1-BGrade2]:
```

```
##                              :...PCNeedsFlag = Y: 1 (33.8/14.8)
##                                  PCNeedsFlag = N:
##                                  :...HCW_Ratio > 6: 0 (65.6/14.9)
##                                      HCW_Ratio <= 6: [S1]
##
## SubTree [S1]
##
## Client_Programs_count_at_observation_cutoff <= 0: 0 (122.1/43.6)
## Client_Programs_count_at_observation_cutoff > 0: 1 (118.7/51.3)
##
## -----  Trial 4:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 46:
## :...HCW_Ratio > 41: 0 (58.1/3.3)
## :   HCW_Ratio <= 41:
## :   :...Issues_Raised <= 0: 0 (556.5/113)
## :       Issues_Raised > 0:
## :       :...AgeAtCreation > 78: 0 (347.8/107.2)
## :           AgeAtCreation <= 78:
## :           :...Client_Programs_count_at_observation_cutoff <= 0: 0 (46.3/13.1)
## :               Client_Programs_count_at_observation_cutoff > 0:
## :               :...default_contract_group in {CHSP,Package,
## :                   :                          Private/Commercial}: 0 (363.9/177.5)
## :                   default_contract_group in {Disability,DVA/VHC,
## :                                              TransPac}: 1 (19.7/3.6)
## AverageCoreProgramHours <= 46:
## :...AgeAtCreation > 92: 0 (61.4/14.4)
##     AgeAtCreation <= 92:
##     :...Client_Programs_count_at_observation_cutoff > 0:
##         :...Issues_Raised <= 0.7666531: 0 (442.5/207.3)
##         :   Issues_Raised > 0.7666531: 1 (388.3/170.7)
##         Client_Programs_count_at_observation_cutoff <= 0:
##         :...HCW_Ratio > 5: 0 (257.2/78.3)
##             HCW_Ratio <= 5:
##             :...MostUsedBillingGrade in [BGrade6-BGrade9]: 0 (187.6/66.8)
##                 MostUsedBillingGrade in [BGrade1-BGrade5]:
##                 :...MostUsedBillingGrade = BGrade1:
##                     :...AgeAtCreation <= 57: 1 (44/14.7)
##                     :   AgeAtCreation > 57: 0 (764.9/316.9)
##                     MostUsedBillingGrade in [BGrade2-BGrade5]:
##                     :...AgeAtCreation <= 70: 0 (143.1/58.3)
##                         AgeAtCreation > 70: 1 (448.5/194.8)
##
## -----  Trial 5:  -----
##
```

```
## Decision tree:
##
## Client_Initiated_Cancellations > 5: 0 (314.7/91.5)
## Client_Initiated_Cancellations <= 5:
## :...HCW_Ratio > 40: 0 (30.8)
##     HCW_Ratio <= 40:
##     :...Issues_Raised > 0:
##         :...MostUsedBillingGrade in [BGrade4-BGrade9]: 1 (157.9/63.5)
##         :   MostUsedBillingGrade in [BGrade1-BGrade3]:
##         :   :...Issues_Raised <= 0.7666531: 0 (116.7/41.4)
##         :       Issues_Raised > 0.7666531:
##         :       :...AgeAtCreation > 78: 0 (741.8/314.5)
##         :           AgeAtCreation <= 78:
##         :           :...Issues_Raised <= 1: 1 (540.6/225.9)
##         :               Issues_Raised > 1: 0 (218.8/96.1)
##         Issues_Raised <= 0:
##         :...AverageCoreProgramHours > 35.5: 0 (563.7/145.8)
##             AverageCoreProgramHours <= 35.5:
##             :...MostUsedBillingGrade = BGrade9: 0 (8.1)
##                 MostUsedBillingGrade in [BGrade1-BGrade6]:
##                 :...HCW_Ratio > 4: 0 (200.2/70.5)
##                     HCW_Ratio <= 4: [S1]
##
## SubTree [S1]
##
## Client_Programs_count_at_observation_cutoff > 0: 1 (162.3/60.2)
## Client_Programs_count_at_observation_cutoff <= 0:
## :...HOME_STR_state in {ACT,NSW,VIC,WA}: 0 (532.5/220.8)
##     HOME_STR_state = QLD: 1 (120.6/51.8)
##     HOME_STR_state = SA:
##     :...MostUsedBillingGrade = BGrade1: 0 (120.4/53.3)
##         MostUsedBillingGrade in [BGrade2-BGrade6]: 1 (301/143)
##
## -----  Trial 6:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 46:
## :...AgeAtCreation > 78: 0 (456.8/43.1)
## :   AgeAtCreation <= 78:
## :   :...ClientType in {Bro,CCP,Com,Dem,Dis,Dom,DVA,EAC,HAC,NRC,Nur,Per,Pri,Res,
## :       :              TAC,TCP,VHC}: 0 (671.8/231)
## :       ClientType in {CAC,Soc,You}: 1 (32/5.6)
## AverageCoreProgramHours <= 46:
## :...ClientType in {Bro,CAC,Com,Dem,Dis,Dom,DVA,NRC,Nur,Per,Pri,Soc,TAC,TCP,VHC,
##     :              You}: 0 (431.4/187.7)
##     ClientType in {CCP,EAC,Res}: 1 (98.5/38.1)
```

```
##     ClientType = HAC:
##     :...default_contract_group = Disability: 1 (1.7)
##         default_contract_group in {DVA/VHC,Package,TransPac}: 0 (194.1/70.8)
##         default_contract_group = CHSP:
##         :...Issues_Raised > 0:
##         :   :...AgeAtCreation <= 82: 0 (576.2/268.4)
##         :   :   AgeAtCreation > 82: 1 (359.3/150.7)
##         :   Issues_Raised <= 0:
##         :   :...AverageCoreProgramHours > 11.25: 0 (308/65.6)
##         :       AverageCoreProgramHours <= 11.25:
##         :       :...AgeAtCreation <= 66: 1 (55/15.2)
##         :           AgeAtCreation > 66: 0 (482/222.4)
##         default_contract_group = Private/Commercial:
##         :...HOME_STR_state = WA: 0 (8)
##             HOME_STR_state in {ACT,NSW,QLD,SA,VIC}:
##             :...HCW_Ratio > 5: 1 (29.8/6.6)
##                 HCW_Ratio <= 5:
##                 :...Issues_Raised > 0: 0 (38.8/11)
##                     Issues_Raised <= 0:
##                     :...MostUsedBillingGrade = BGrade1: 1 (80/27.2)
##                         MostUsedBillingGrade in [BGrade2-BGrade9]: [S1]
##
## SubTree [S1]
##
## Client_Programs_count_at_observation_cutoff <= 0: 0 (212.8/94.7)
## Client_Programs_count_at_observation_cutoff > 0: 1 (19.7/6.5)
##
## -----  Trial 7:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 40: 0 (1039.9/149.5)
## AverageCoreProgramHours <= 40:
## :...Client_Programs_count_at_observation_cutoff > 2: 0 (29.7/5.8)
##     Client_Programs_count_at_observation_cutoff <= 2:
##     :...AgeAtCreation > 83:
##         :...HOME_STR_state in {NSW,QLD}: 0 (553.5/169.3)
##         :   HOME_STR_state in {ACT,SA,VIC,WA}:
##         :   :...Issues_Raised <= 0.7666531: 0 (215.8/96.9)
##         :       Issues_Raised > 0.7666531: 1 (87.4/34.2)
##         AgeAtCreation <= 83:
##         :...Client_Initiated_Cancellations > 3: 0 (106.3/31.8)
##             Client_Initiated_Cancellations <= 3:
##             :...AverageCoreProgramHours > 13.75:
##                 :...Client_Programs_count_at_observation_cutoff <= 0.6800728: 0 (285.8/97)
##                 :   Client_Programs_count_at_observation_cutoff > 0.6800728:
##                 :   :...ClientType in {Bro,CAC,CCP,Dis,EAC,Per,TAC,
```

```
##                 :       :              VHC}: 0 (12.1)
##                 :       ClientType in {Com,Dem,Dom,DVA,HAC,NRC,Nur,Pri,Res,Soc,
##                 :                      TCP,You}: 1 (258.4/112.9)
##                 AverageCoreProgramHours <= 13.75:
##                 :...Issues_Raised > 0.7666531: 1 (318.4/118.1)
##                     Issues_Raised <= 0.7666531:
##                     :...ClientType in {NRC,Soc,TAC,TCP,VHC,
##                         :              You}: 1 (0)
##                         ClientType in {CCP,Dem,Dis,DVA,Nur,
##                         :              Res}: 0 (14.3)
##                         ClientType in {Bro,CAC,Com,Dom,EAC,HAC,Per,Pri}:
##                         :...HCW_Ratio > 2: 0 (135.8/52.6)
##                             HCW_Ratio <= 2: [S1]
##
## SubTree [S1]
##
## default_contract_group in {Disability,DVA/VHC}: 1 (1.5)
## default_contract_group in {Package,TransPac}: 0 (23/10.8)
## default_contract_group = Private/Commercial:
## :...AgeAtCreation <= 40: 0 (29.2/9.1)
## :   AgeAtCreation > 40: 1 (348.1/137.1)
## default_contract_group = CHSP:
## :...ClientType in {Bro,CAC,Pri}: 1 (0)
##     ClientType = Com: 0 (5.7)
##     ClientType in {Dom,EAC,HAC,Per}:
##     :...Client_Programs_count_at_observation_cutoff > 0: 1 (134.9/44.3)
##         Client_Programs_count_at_observation_cutoff <= 0:
##         :...AverageCoreProgramHours <= 9.25: 1 (323.2/129)
##             AverageCoreProgramHours > 9.25: 0 (20.9/4.6)
##
## -----  Trial 8:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 37.5:
## :...ClientType in {Bro,CAC,CCP,Com,Dem,Dis,Dom,DVA,EAC,HAC,NRC,Nur,Per,Pri,Res,
## :   :              TAC,TCP,VHC}: 0 (892.3/51.8)
## :   ClientType in {Soc,You}: 1 (32.6/6.4)
## AverageCoreProgramHours <= 37.5:
## :...Client_Initiated_Cancellations > 3: 0 (124.1/15.8)
##     Client_Initiated_Cancellations <= 3:
##     :...MostUsedBillingGrade in [BGrade4-BGrade9]:
##         :...HOME_STR_state in {NSW,VIC,WA}: 0 (125.6/34.4)
##         :   HOME_STR_state in {ACT,QLD,SA}:
##         :   :...AgeAtCreation <= 71: 1 (97.4/12.9)
##         :       AgeAtCreation > 71:
##         :       :...MostUsedBillingGrade in [BGrade4-BGrade6]: 1 (228.2/80.5)
```

```
##         :           MostUsedBillingGrade = BGrade9: 0 (71.6/29.3)
##         MostUsedBillingGrade in [BGrade1-BGrade3]:
##         :...PCNeedsFlag = Y:
##             :...Client_Initiated_Cancellations > 2: 0 (40.4/7.3)
##             :   Client_Initiated_Cancellations <= 2:
##             :   :...Issues_Raised <= 1: 1 (294.4/124.9)
##             :       Issues_Raised > 1: 0 (58.9/16.5)
##             PCNeedsFlag = N:
##             :...AgeAtCreation > 88: 0 (117.6/10.5)
##                 AgeAtCreation <= 88:
##                 :...AverageCoreProgramHours > 18.5: 0 (465.7/125.3)
##                     AverageCoreProgramHours <= 18.5:
##                     :...default_contract_group = Disability: 1 (1.4)
##                         default_contract_group in {DVA/VHC,Package,
##                         :                          TransPac}: 0 (66.1/18.3)
##                         default_contract_group in {CHSP,Private/Commercial}:
##                         :...HOME_STR_state = WA: 0 (16.8/1.4)
##                             HOME_STR_state in {ACT,NSW,QLD,SA,VIC}:
##                             :...Issues_Raised <= 0.7666531: 0 (782.9/307.5)
##                                 Issues_Raised > 0.7666531:
##                                 :...MostUsedBillingGrade = BGrade3: 1 (20.4/3.1)
##                                     MostUsedBillingGrade in [BGrade1-BGrade2]: [S1]
##
## SubTree [S1]
##
## default_contract_group = Private/Commercial: 0 (25.8/5.3)
## default_contract_group = CHSP:
## :...AgeAtCreation <= 65: 1 (20.5/3.3)
##     AgeAtCreation > 65:
##     :...Client_Programs_count_at_observation_cutoff <= 0: 0 (261/116.6)
##         Client_Programs_count_at_observation_cutoff > 0: 1 (60.2/21.8)
##
## -----  Trial 9:  -----
##
## Decision tree:
##
## AverageCoreProgramHours > 18.75: 0 (1393.8/87.3)
## AverageCoreProgramHours <= 18.75:
## :...Client_Initiated_Cancellations > 3: 0 (34)
##     Client_Initiated_Cancellations <= 3:
##     :...ClientType in {CAC,CCP,Dem,Dis,DVA,NRC,Soc,TCP,VHC,
##         :              You}: 0 (46.8/2.8)
##         ClientType in {Bro,Com,Dom,EAC,HAC,Nur,Per,Pri,Res,TAC}:
##         :...AgeAtCreation > 83: 0 (554.2/167.4)
##             AgeAtCreation <= 83:
##             :...HCW_Ratio > 3: 0 (237.7/88.6)
##                 HCW_Ratio <= 3:
```

```
##                 :...HOME_STR_state = VIC: 0 (10.2)
##                     HOME_STR_state in {ACT,QLD,SA}:
##                     :...HOME_STR_state in {ACT,QLD}: 1 (347.8/114)
##                     :   HOME_STR_state = SA:
##                     :   :...MostUsedBillingGrade = BGrade1: 0 (104.7/25.4)
##                     :       MostUsedBillingGrade in [BGrade2-BGrade9]: 1 (274.1/63.6)
##                     HOME_STR_state in {NSW,WA}:
##                     :...AgeAtCreation <= 40: 0 (25.1)
##                         AgeAtCreation > 40:
##                         :...MostUsedBillingGrade = BGrade9: 0 (6.6)
##                             MostUsedBillingGrade in [BGrade1-BGrade6]:
##                             :...Issues_Raised > 0.7666531: 1 (222.2/64.1)
##                                 Issues_Raised <= 0.7666531:
##                                 :...AgeAtCreation <= 73: 1 (212.3/90.3)
##                                     AgeAtCreation > 73: 0 (174.5/39.2)
```

Appendix F

Vocabulary of Terms

Term Used	Meaning
Observation	is a set of variables pertaining to one client. It is alternately termed as an instance or a row in the dataset.
Dataset	is a collection of instances or observations. Datasets are used to train a model or be used to test a model. It can be a subset or a sample of the population.
Attribute	is a column of data pertaining to a client. It is also termed as a variable. In the context of modelling, is termed as feature and in prediction models, as a predictor.
Outcome	is also called dependent variable or response Variable and is the target attribute to be predicted by prediction models.
Label	is the specific outcome value assigned to outcome/response variable. When use as a verb, it is process of attaching an outcome to an observation. It can refer to an observed or predicted outcome.
Service	a specific type of home care service performed to a client and subsequently billed as an invoice item.
Program	is a bundled set of pre-defined and customized services that is subscribe to by clients.
Core	a tag assigned to Programs which indicates inclusion for client churn monitoring.
OOB	stands for Out-of-bag and is the term use by a Random Forest model for observations excluded from its bootstrap sampling and random selection of variables used for tree nodes.
AIC	Akaike Information Criteria is a measure of model fitness used o compare successive iterations of a model using the maximum likelihood estimation (e.g. regression)
SPSS	Statistical Package for the Social Sciences is a software by IBM used for statistical analysis
Gradient Boosted Model	an ensemble of weak prediction models, typically decision trees that employs boosting methods and it generalises by optimising differentiable loss function.
CHAID	Chi-square automatic interaction detection (CHAID) is a decision tree technique, based on adjusted significance testing (Bonferroni testing)
ANOVA	Analysis of Variance, a statistical tests that measures the means of variables and determine if the variance is generalisable

Appendix G

R Program and Results

https://github.com/raulmanongdo/R-PredictionModel/blob/master/kc_client_churn_FINAL_G.md

Bibliography

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. & Scuse, D. (2016), *WEKA Manual for Version 3-8-0*, University of Waikato, Hamilton, New Zealand.

Breiman, L. (2001), Random forest, Technical report, University of California.

Brownlee, B. (2015*a*), *in* 'Connect: Insights for Business'.

Brownlee, J. (August 2015*b*).
URL: *https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000), *CRISP-DM 1.0 Step-by-step data mining guide*, CRISP-DM Consortium.

core team, R. (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Coussement, K. & den Poel, D. V. (2008), 'Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques', *Expert Systems with Applications* .

Coussement, K. & den Poel, D. V. (2009), 'Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers', *Expert Systems with Applications* .

Erhardt, E. (2017), 'Logistic regression and newtonraphson'.
URL: *https://statacumen.com/teach/SC1/SC1_11_LogisticRegression.pdf*

Fawcett, T. (January 7 2003), 'Roc graphs: Notes and practical considerations for data mining researchers'.

Fortmann-Roe, S. (June 2012), 'Understanding the bias-variance tradeoff'.
URL: *http://www.citeulike.org/group/15400/article/13849984*

Geron, S., Smith, K., Tennstedt, S., Jette, A., Chassler, D. & Kasten, L. (2000), 'The home care satisfaction measure: a client-centered approach to assessing the satisfaction of frail older adults with home care services', *Journal of Gerontology* .

Golmohammadi, D. & Radnia, N. (2016), 'Prediction modeling and pattern recognition for patient readmission', *International Journal on Production Economics* .

Group, U. S. C. (June 2017), 'Logit regression: R data analysis examples'.
URL: *https://stats.idre.ucla.edu/r/dae/logit-regression/*

Guo-en, X. & Wei-dong, J. (2008), 'Model of customer churn prediction on support vector machine', *Systems Engineering - Theory and Practice,* .

Han, J. & Kamber, M. (2006), *Data Mining Concepts and Techniques*, University of Illinois at Urbana-Champaign.

Health, A. D. (2017), 'Ageing and aged care home care packages reform'.
URL: *https://agedcare.health.gov.au/aged-care-reform/home-care/home-care-packages-reform*

Heng-liang Wu, Wei-wei Zhang, Y.-y. Z. (n.d.), 'An empirical study of customer churn in e-commerce based on data mining'.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Yanhua, L., Ni, B., Dai, W., Yang, Q. & Zeng, J. (2015), Telco churn prediction with big data, *in* 'Proceedings of ACM SIGMOD International Conference on Management of Data'.

Kim Y. S., M. S. (2012), 'Measuring the success of retention management models built on churn probability, retention probability and expected yearly revenues', *Expert Systems with Applications* .

Kuhn, M., Weston, S., Coulter, N., Culp, M. & Quinlan, R. (2015), *C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-24.
URL: *https://CRAN.R-project.org/package=C50*

Liaw, A. & Wiener, M. (2002), 'Classification and regression by randomforest', *R News* **2**(3), 18–22.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Luan, J. (2002), 'Data mining and its applications in higher education'.
URL: *https://eric.ed.gov/?id=ED474143*

Madigan, E. & Curet, O. L. (2006), 'A data mining approach in home healthcare: outcomes and service use', *BMC Health Services Research* .

Mylod, D. E. & Kaldenberg, D. O. (2000), 'Data mining techniques for patient satisfaction data in home care settings', *Home Health Care Manage Practice* .

National Institute of Health, U. N. L. o. M. (Sep 2012).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/*

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. & Mason, C. (2006), 'Defection detection: Measuring and understanding the predictive accuracy of customer churn models', *Journal of Marketing Research* .

Nguyen, E. H. X. (2011), 'Customer churn prediction for the icelandic mobile telephony market'.

P., R., Tang, L. & L., H. L. (2009), *Cross-Validation*, Springer, chapter C.

Reichheld, F. (2006), 'Net promoter'.
URL: *https://en.wikipedia.org/wiki/Net_Promoter*

Research, R. Q. (March 2017).
URL: *https://www.rulequest.com/see5-unix.html#RULES*

Ruiz-Gazen, A. & Villa, N. (2007), 'Storm prediction: Logistic regression vs random forest for unbalanced data', *Case Studies in Business, Industry and Government Statistics* .

Singh, H. & Samalia, H. V. (2014), A business intelligence perspective for churn management, *in* '2nd World Conference On Business, Economics And Management'.

Strobl, C. & Zeileis, A. (2008), 'Why and how to use random forest variable importance measures (and how you should not)'.
URL: *https://www.statistik.uni-dortmund.de/useR-2008/slides/Strobl+Zeileis.pdf*

Sunil Gupta, e. a. (2006), 'Modeling customer lifetime value', *Journal of Service Research* .

Tang, J., Alelyani, S. & Liu, H. (2015), 'Feature selection for classification: A review'.
URL: *http://eprints.kku.edu.sa/170/*

Tsai, C.-F. & Chen, M.-Y. (2010), 'Variable selection by association rules for customer churn prediction of multimedia on demand', *Expert Systems with Applications* .

Tsai, C.-F. & Lu, Y.-H. (2009), 'Customer churn prediction by hybrid neural networks', *Expert Systems with Applications* .

Van Rijsbergen, C. J. (1979), *Information Retrieval*, Butterworth.

W. Verbeke, D. M. e. a. (2011), 'Building comprehensible customer churn prediction models with advance rule induction techniques', *Expert Systems with Applications* .

Wikipedia (June 2017), 'Akaike information criterion'.
URL: *https://en.wikipedia.org/wiki/Akaike_information_criterion*

William McCausland, e. a. (1998), Churn amelioration system and method, patent number 5822410, *in* 'United States Patents'.

Y. Kim, H. L. & Johnson, J. (2013), 'Churn management optimization with controllable marketing variables and associated management costs', *Expert Systems with Applications* .