

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

Certificate	i
Acknowledgment	iii
List of Figures	vii
List of Tables	viii
List of Publications	ix
Abstract	x
Chapter 1 Introduction	1
1.1 Introduction and Context of Study	1
1.2 The Problem	2
1.3 Aim of this Study	3
1.4 Research Significance and Contribution	4
1.5 Thesis Structure	5
Chapter 2 Background	7
2.1 Introduction	7
2.2 Home care services industry	7
2.2.1 Trends for Home Care Services	8
2.2.2 Peculiarities of Home Care Services	9
2.3 Case company	10
2.4 Client Churn Prediction, Satisfaction and Retention	13
2.5 Churn Analysis and Prediction Modelling	14
2.5.1 Feature Selection Techniques	14
2.5.2 Regression and Classification	16

2.5.3	Decision Trees and Ensemble methods	17
2.5.4	Support Vector Machine	18
2.5.5	Artificial Neural Net	19
2.5.6	Ant Colony Optimisation	19
2.6	Model Bias, Variance and Imbalance Data	20
2.7	Model Performance Measures	21
2.8	General Methodology and tools used	22
2.9	Conclusion	22
Chapter 3	Literature Review	24
3.1	Introduction	24
3.2	Applied Churn Prediction Model	24
3.3	Churn associated studies on home care services	28
3.4	Client Churn Analysis	30
3.5	Conclusion	32
Chapter 4	Data Description and Churn Analysis	34
4.1	Introduction	34
4.2	Churn Definition and Measure	34
4.3	Data Collection and the Dataset	38
4.4	Data Cleansing	39
4.5	Churn Analysis in various dimensions	40
4.6	Conclusion	45
Chapter 5	Prediction Modelling	46
5.1	Introduction	46
5.2	Model Development Methodology	46
5.3	Data Preparation	48
5.4	Feature Selection	50
5.4.1	Significant variables in Logistic Regression	50
5.4.2	Important variables in Random Forest	52
5.4.3	Reduced Dimensions using Correlation Analysis	53

5.5	Candidate Prediction Models in Training	56
5.5.1	Logistic Regression	57
5.5.2	Random Forest	61
5.5.3	C5.0 model	63
5.6	Model Comparison and Evaluation	67
5.7	Selected model and tuning parameters	70
5.8	Churn Model Analysis and Insights	72
5.9	Conclusion	73
Chapter 6	Conclusion	75
6.1	Conclusion and Research Answers	75
6.2	Future Work	76
Appendix A	Attributes	78
Appendix B	Summary of Raw Categorical Data	80
Appendix C	Summary of Raw Numerical Data	82
Appendix D	Correlation Matrix	84
Appendix E	C5.0 model Decision Rules	87
Appendix F	Vocabulary of Terms	97
Appendix G	R Program and Results	98
Bibliography	99

List of Figures

2.1	Home-based care services Business Process Agents	11
4.1	Annual Client Churn Rate	37
4.2	Source data Entity Relationship Diagram	38
4.3	Churns by Age Group and Health (aka Billing) Grade	40
4.4	Client Discharge Reasons and Churns	41
4.5	Client Discharge Subreasons and Churns	42
4.6	Client Program enrolments and Churns	42
4.7	Client Program Services and Churns	43
4.8	Client Satisfaction Survey Responses and Churns	44
5.1	Model Development Observation Windows	47
5.2	Variable importance measures in RF	53
5.3	Feature-to-feature Correlation Analysis	55
5.4	RF model variable importance by decrease in accuracy	62
5.5	Comparison of Model AUC on 10-fold validation datasets	69

List of Tables

3.1	Client Churn Prediction Models reviewed	28
3.2	Churn associated studies on Home-based Care Services	30
5.1	Model Development Summary	48
5.2	Selected Features	51
5.3	Logistic Regression significant variables	52
5.4	RF variables ranked by Accuracy	54
5.5	Standardised Logistic Regression Coefficients	58
5.6	Logistic Regression model insights	59
5.7	Top C5.0 churn decision rules ranked by accuracy	66
5.8	Comparison of Prediction Model Performances	68
5.9	Pair-wise comparison of model significance (AUC)	69
5.10	C5.0 model parameter tuning	72

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.