**Predicting Customer Churn for Insurance Data**

Michael Scriney, Dongyun Nie, and Mark Roantree

Insight Centre for Data Analytics, School of Computing,

Dublin City University, Ireland

1 Introduction

One of the major goals of Customer Relationship Management is to maximize the Customer Lifetime Value (CLV) for the purpose of supporting long term business investment [8]. CLV is a measure that focuses on predicting the net profit that can accrue from the future relationship with customers [4]. This metric can be calculated by recording the behaviours of the customer over the longer term and thus, help to build a customised business strategy. It has been a popular research topic, addressed by researchers in different ways, for example, formulaic CLV [11] and Probability Model CLV [15]. One of the core elements in CLV models [5] and the calculation of CLV scores is customer retention or its opposite, customer churn. In the business sector, customer churn is commonly used not only to support CLV predictions, but to maximize customer profitability by establishing resource allocation decisions for marketing, sales, and customer interaction. As a result of the benefits that churn analysis provides, this topic has become popular for industrial research in recent years. Some of business based research focuses on statistical efforts, often required for CLV calculations [17]. Information technology based research generally experiments with data mining techniques [7] to try to generate the variables needed for CLV scores. In the telecom sector, churn analysis research has been shown to require a specific set of variables [18] for effective results.

1.1 Problem Statement

One of the issues with CLV research and the generation of variables such as churn is the highly theoretical nature and focus of this area of research. There has been little research on deriving necessary attributes from real world datasets which require activities such as the construction, transformation and enrichment of datasets to make them suitable for existing CLV calculations. Researchers in [2] highlighted the variables that are required as input to CLV calculations. Here, a is the acquisition rate; A is the acquisition cost per customer; d is the yearly discount rate; m is the net income of a transaction; R is the retention cost per customer per year; and r is the yearly retention rate.

On the surface, it appears as if the generation of CLV scores for all customers is a straightforward process but in reality, many of these variables are not easily extracted from enterprise databases. From [2], m and d can be deduced using feature extraction and a detailed clustering process but all others will generally require some form of imputation after a segmentation process. Customer segmentation is regarded as a natural process to help companies to classify customers and plan market investment strategies such as direct sales. As such, it has been widely adopted by industry planners or in data warehousing similar efforts include fragmentation of the data and queries [13]. Moreover, it plays a critical role in the development of the company’s position by combining product differentiation and marketing segmentation to provide resources, objectives and competences to the company.