International Journal of Information Management 48 (2019) 238–253

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

A big data analytics model for customer churn prediction in the retiree segment

Farid Shirazi[a,*], Mahbobeh Mohammadi[b]

[a] Ted Rogers School of Information Technology Management, Ryerson University, Toronto, Canada
[b] Ted Rogers School of Management, Ryerson University, Toronto, Canada

ARTICLE INFO

Keywords:
Big data
Business intelligence
Churn prediction model
Hadoop
Customer lifetime value
Classification
Regression tree

ABSTRACT

Undoubtedly, the change in consumers' choices and expectations, stemming from the emerging technology and also significant availability of different products and services, created a highly competitive landscape in various customer service sectors, including the financial industry. Accordingly, the Canadian banking industry has also become highly competitive due to the threats and disruptions caused by not only direct competitors, but also new entrants to the market.

The primary objective of this paper is to construct a predictive churn model by utilizing big data, including the structured archival data, integrated with unstructured data from sources such as online web pages, the number of website visits and phone conversation logs, for the first time in the financial industry. It also examines the effect of different aspects of customers' behavior on churning decisions. The Datameer big data analytics tool on the Hadoop platform and predictive techniques using the SAS business intelligence system were applied to study the client retirement journey path and to create a churn prediction model. By deploying the above systems, we were able to uncover a wealth of data and information associated with over 3 million customers' records within the retiree segment of the target bank, from 2011 to 2015.

1. Introduction

Unarguably, the financial industry is evolving with an increasing magnitude and fast pace, according to the discernable changes in consumers' choices and expectations, which are stemming from the emerging technology and significant availability of different products and services. As a result, the banking industry has also become highly competitive due to all the threats and disruptions caused by not only direct competitors, but also all the new and innovative entrants, such as Apple, Google, and new start-ups. Hence, sustaining a competitive advantage and maintaining the Point of Differentiation (POD), in order to remain in clients' financial paths, is considered one of the highest priorities in strategic planning related to client attraction and more importantly, retention within the retail banking sector.

Despite the fact that different types of Customer Relationship Management (CRM) strategies have been in existence for many decades, it became the center of focus for many researchers and practitioners, after the business world shifted its marketing focus from product-centric strategies towards customer-centric strategies. As a result, the relationships between customers and companies have evolved in a way that many new marketing opportunities have been created (Ngai, 2005). Moreover, customer retention has become one of the main areas on which most of the CRM strategies focus on. As mentioned by Coussement, Benoit, and Poel (2015), one of the cornerstones of CRM is customer churn prediction, where one attempts to predict whether or not a customer will leave the company. Moreover, the churn rate is defined as the annual percentage rate at which customers cease to subscribe to a service or terminate a business relationship.

Reducing the rate of churn and retaining current customers are the most cost-effective marketing approaches that will maximize shareholder's value (Ekinci et al., 2012; Ngai, Xiu, & Chau, 2009; Ryals & Knox, 2005; Van den Poel & Lariviere, 2004) Asmentioned by Lin, Tzeng, and Chin (2010), with so much competition, companies need to focus on retaining existing customers by effectively satisfying their needs; otherwise, they risk losing their customers; and losing customers offers competitors the opportunities to attract them.

Accordingly, the client retiree segment has been identified as one of the highest client strategic priorities within the Personal and Commercial Banking (P&CB) of "target bank"[1]. In order to own the retirement segment and become the retirement advisor and provider of

* Corresponding author at: 350 Victoria St, Toronto, ON M5B 2K3, Canada.
E-mail addresses: f2shiraz@ryerson.ca (F. Shirazi), mmohamma@ryerson.ca (M. Mohammadi).
[1] Name restricted due to confidentiality

https://doi.org/10.1016/j.ijinfomgt.2018.10.005
Received 12 February 2018; Received in revised form 6 October 2018; Accepted 8 October 2018
Available online 17 October 2018

choice in Canada, a new research project is sponsored by the CRM department, which focuses on “Mass Affluent” pre-retiree, and retiree client segments, which aims to evaluate and examine various marketing approaches towards maximizing retention rates.

Generally, each industry has its definition of the mass affluent segment, but according to the Canadian Banking industry, mass affluent segments include clients who have investable assets between $100,000 CAD and $1,000,000 CAD. According to the financial industry, clients who are categorized solely based on their assets, belong to one of the following categories:

- Non-Mass Affluent or Mass Retail: Clients with less than $100,000 investable assets
- Mass Affluent: Clients with investable assets between $100,000 and $1,000,000
- High Net-Worth: Clients with higher than $1,000,000 investable assets

As indicated on the Bank Administration Institute (BAI) report of 2014, mass affluent clients are difficult to be penetrated due to their specific knowledge base, their expectations of the market, and the return on their investments. Therefore, gaining a thorough understanding of their expectations and their investment criteria is necessary to not only successfully attract them but also to retain them (Bank Administration Institute, 2014). For a reason mentioned above, analyzing customers' shopping behavior through big data techniques will provide an opportunity to indicate the products that a specific client has expressed an interest in and has reviewed through external websites. This type of scrutiny provides an opportunity for the marketing team to design and offer customized products and services that a specific client desires. According to Lewis and Soureli, this marketing approach will boost customers' satisfaction and will likely increase the probability of retention in return (Lewis & Soureli, 2006). It is noteworthy that none of the previous churn prediction studies have focused specifically on this multi-billion dollar market.

1.1. Background

Currently, as illustrated in Fig. 1, there is a well-defined predictive model in place that estimates the likelihood of retirement in the next 10–22 months, by using demographic data. However, there is a need to study the client's retirement journey path in order to identify the moment when a specific client is naturally more receptive of retirement advice, the so-called "receptive point", in order to win the "moment of truth" by gaining customer's loyalty and minimizing the possibility of churn during their retirement journey path. This enables the business to not only provide the best-in-class services to clients but also provides proactive guidance and the most relevant advice at the right time, which will ultimately increase the likelihood of retention.

Although emerging technologies and the improvement of data mining techniques have made it possible to perform comprehensive studies about customer churn predictions within many industries, including the financial industry, there is an apparent absence of big data analytics in the customer churn prediction. Furthermore, as Chen et al. claimed, big data analytics is being used in many science and technology-driven projects but has received less attention in service industries, such as the financial services (Chen, Chiang, & Storey, 2012).

Since this research study deals with the prediction of customer churn in the financial industry using big data analytics techniques, it will help to fill the analytical gap. The study will contribute to the existing literature through a) developing a customer churn prediction model for the financial sector by using big data analytic techniques, b) building a model based on a dataset which consists of only mass affluent clients from the retiree segment, c) producing multiple outcomes in this model. Compared to most of the existing churn literature, which focus on the binary status of churned and not-churned clients; our proposed model will return four different types of customers: churned, not yet churned (potential clients), partially churned (the second potential client segment), and retained clients. According to the current literature, it is the first study that uses Classification and Regression Tree (CRT) in conjunction with regression analysis for implementing the churn prediction model on big data, incorporating different sources of data.

This paper serves as the preliminary assessment and measurement for the construction of the Churn Prediction Model (CPM).

The remainder of this paper is structured in the following order. Section 2 reviews the existing literature, while Section 3 highlights the rationale behind the need for the current research study, and is followed by the main question and a list of hypotheses. Section 4 will explain the research approach, data population, and all the cues that are utilized to predict the churn event for each client category within the population of this study. In Section 5 we will explain our big data model followed by Section 6, which outlines a comprehensive discussion regarding the above findings, along with the limitations and strengths of this study. Finally, in Section 7, a detailed conclusion along with all the recommendations is stated, which the target bank’s Marketing Strategy team will use.

2. Review of literature

For the scope of this paper, the literature studies are categorized into three main groups: 1) Customer Lifetime Value (CLV) measurement: to indicate the best available model for creating the population dataset for this study, 2) churn predictive indicators and models: in order to locate the most relevant indication factors in a churn event, 3) attrition analysis methods: which will be used as part of the final recommendation made to the marketing executives, in order to design a sound retention campaign for customers who are identified as potential clients.

2.1. Customer lifetime value

Customer Lifetime Value (CLV) is a total of net value that a single client possesses for the company over the lifetime of their relationship (Hoekstra & Huizingh, 1999). For example, in this study, CLV is measured based on the net value each client possesses for the bank, from the

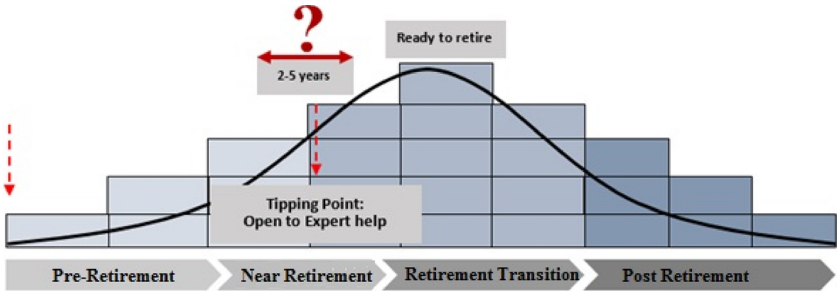


Fig. 1. Client Retirement Journey Path.

beginning of their banking relationship.

More than 62% of customer relationship studies are related to customer retention (Ngai et al., 2009). According to current studies, CRM is segmented into two strategies: "Operational" and "Analytical." While operational CRM strategies focus on process automation and optimization, the analytical CRM strategies primarily focus on customer behavioral studies by classifying and analyzing customers' data to provide appropriate guidelines for CRM marketing (Rosset, Neumann, Eick, Vatnik, & Idan, 2002). In general, CRM can be classified into four major dimensions: customer identification, attraction, retention, and development (Ngai et al., 2009; Au, Chan, & Yao, 2003; Kracklauer, Mills, & Seifert, 2004; Ling & Yen, 2001). As mentioned by De Caigny, Coussement, and De Bock (2018) customer churn prediction is an important research discipline within a CRM context. It introduces not only loss in profit but also imposes other adverse effects on the operation of a business; such as the increased cost of attracting new customers and loss of up sale opportunities (De Caigny et al., 2018). As such, it is not surprising that the customer relationship management and more specifically customer retention, has received a growing amount of attention (Verbeke, Martens, & Baesens, 2014) in recent years.

Current studies also confirm that six different data mining algorithms are widely used by CRM researchers (Ngai et al., 2009). This paper is mainly focused on the assessment of the association rule, regression analysis, and the decision tree. Stated applied algorithms are also suggested for big data analytics (Chen et al., 2012). By analyzing customers' behavior, firms will be able to predict a churn event and identify potential churners to design a customized incentive program as part of the retention strategies (Ngai et al., 2009). Likewise, as indicated by Ekinci, Uray, and Ülengin (2014), in order to address the drawback of previous literature in measuring a customer's lifetime value, the measurement must be modeled based on the product and industry-specific indicators.

2.2. Churn prediction, indicators and models

Although different methods return different accuracy rates, selecting the right method for our churn prediction model significantly impacts how accurate the outcome would be. Therefore, detecting the accuracy of the defection rate is a noticeable objective for many studies to define the best churn prediction model (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). As argued by Verbeke, Martens, Mues, and Baesens (2011), the accuracy of the churn prediction model allows the company to target future churners in a retention marketing campaign correctly.

On the one hand, current studies confirm that models have staying power for at least three months because it has been proven that the result of the current churn models remains valid for approximately three months after each run (Neslin et al., 2006). Alternately, the cost of maintaining and running the churn prediction models on a regular basis is significantly high for companies, including banks. Hence, a more in-depth analysis was conducted as part of this study to validate the above findings. After performing the constructed churn model in different periods and comparing the results, we were able to confirm that the three months staying power is also applicable to bank customers. This finding will be reflected in the final recommendations of this study.

Out of the five distinct estimation and variable selection techniques which are used in the churn prediction models, the logistic and decision-tree approaches, perform relatively well compared with other techniques (Moro, Cortez, & Rita, 2014; Neslin et al., 2006). For this study, a decision tree is used to confirm the results of the churn model and the probability of each churn event. The decision tree is a familiar technique and has had many successful applications to real-world problems, due to its symbolic learning technique that organizes information extracted from a training dataset (Nie, Rowe, Zhang, Tian, & Shi, 2011). It can build models using datasets, including numerical and categorical data (Nie et al., 2011).

With this background and confirmation that the selected method will significantly impact the accuracy of the results, more papers regarding different churn prediction methods were evaluated, in order to find the most suitable technique for the current research. For example, by reviewing the most distinguished work of Verbeke et al. (2011), it is confirmed that correct and accurate results of churn models increase the effectiveness of a retention campaign (Neslin et al., 2006; Prasad & Madhavi, 2012; Verbeke et al., 2011).

Verbeke et al. (2011) compared two commonly used techniques in predicting customer churns Active Learning Based Approach (ALBA) and AntMiner+. They also included traditional rule-based classification techniques, such as C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and confirmed the findings of previous literature that AntMiner + is the most comprehensible classification technique since it allows for adding of domain knowledge as customized rule-set variables (Verbeke et al., 2011). Their work also supported the finding of Prasad and Madhavi's research (2012) that C4.5 and C5 (higher version of 4.5) are not the best performing classification decision trees because C4.5 and its next-generation C5, require more information and use a large number of nodes, which makes the model very complex. Therefore, AntMiner + techniques are used to classify the customer segmentation in the current study. However, since the AntMiner + algorithm is based on the binary method, and this research is building a multi-result model, we combined AntMiner + techniques with Neural Network analysis in order to create multi-level customer segmentation.

Evaluating various churn prediction models using data mining techniques such as C5 and in comparison with the Classification and Regression Tree Techniques revealed that CRT returns a higher rate of accuracy (Prasad & Madhavi, 2012; Verbeke et al., 2011).

Therefore, the most suitable approach for building an accurate churn model is to use a dynamic time-period for each customer instead of a static one (Prasad & Madhavi, 2012). Although studies that were conducted suggested that the best indicator of churn possibility depends on the past three months' activity. While the three months approach may address churn requirements in different client segments other than retiree segments, for this research and according to the financial industry standards, a an extended period is considered since customers usually begin planning their retirement decisions 10–22 months prior to their actual date of retirement.

2.3. Attrition analysis methods

In terms of defining a churn prediction model with high rates of accuracy, the proportional hazard technique, which is a class of survival analysis, is being used by many researchers, primarily because this model not only enables the prediction of the churn event, but it also ascertains the time of the churn (Van den Poel & Lariviere, 2004). Current literature indicates that demographic and environmental changes have significant impacts on retention (Van den Poel & Lariviere, 2004), whereas customer behavior predictors reflect only a limited impact. Although our study partially rejects this theory as the behavioral factors considered in previous works were only internal data, such as paying off all the credit debts and mortgages or enquiring with financial planners in banks about available options. On the contrary, this study outlines the online behavior of customers on external websites and products are analyzed to identify the correlation between customers' behavior and the churn decision. The findings of this research confirm that slowing down the banking relationship is an indication of churn; however, compared to online shopping behavior, this factor is more significant within the retiree segment, as compared to the non-retiree segment.

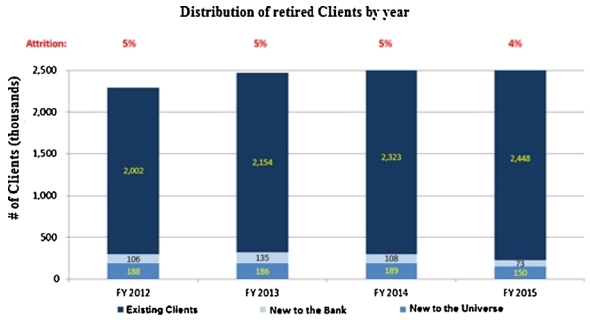


Fig. 2. Distribution of retired clients 2012–2015.

As shown in Fig. 3, an analysis of the retired clients confirmed that mass affluent clients are approximately two times more likely to retire with the target bank, compared to non-mass affluent clients; meaning that mass-retail clients attrite twice more, in comparison to mass-affluent clients.

3. Rational, hypothesis and objective

3.1. Rational

As illustrated in Fig. 2, preliminary analysis of our population, which includes mass and non-mass affluent clients, revealed that the main source of the retiree population is existing clients who have had a long history with the company. This study also shows that the attrition rate has remained at 5% flat, year by year.

As shown in Fig. 3, an analysis of the retired clients confirmed that mass affluent clients are approximately two times more likely to retire with the target bank, compared to non-mass affluent clients; meaning that mass-retail clients attrite twice more, in comparison to mass-affluent clients.

The above numbers confirm that mass-affluent clients who have a long history with the bank are more likely to remain loyal for at least a few years after retirement. This result can also be confirmed by looking at Fig. 4, which compares the attrition rate between mass affluent and non-affluent clients, year over year. As demonstrated, the attrition rate of affluent clients' accounts for 1%, versus that of 7% for non-mass affluent clients. The analysis also exposed that the attrition rate is stable, year over year, which is an essential factor for defining the retention strategy.

Based on the stated facts, designing and implementing a retention strategy for mass affluent clients seems necessary as it will return a higher rate of success by improving and strengthening the relationship with this population. To do so, using the current prediction system of the target bank is not sufficient, and there is a need for understanding the clients' needs and their shopping behavior in order to offer the right product and services. The current research study is being designed to address this objective by answering the main question and stated hypotheses that are outlined in the next section.

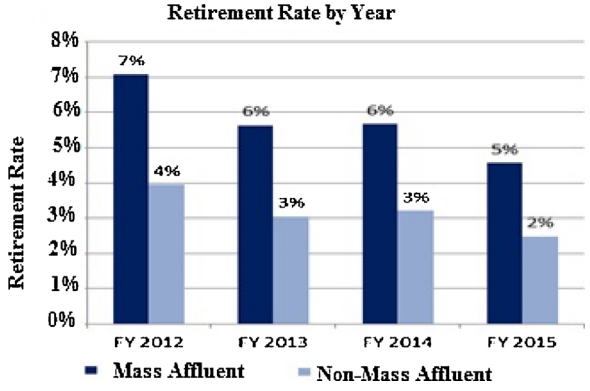


Fig. 3. Retirement rate 2012–2015.

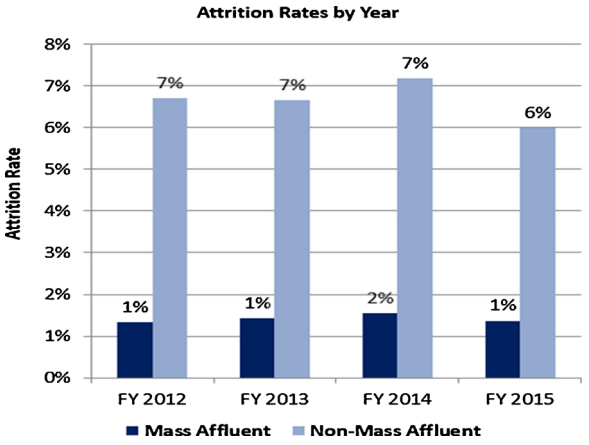


Fig. 4. The attrition rate of Mass Affluent clients vs. Non-Mass Affluent 2012–2015.

3.2. Hypothesis

The following main question is going to be addressed in this research: Is there any association between a customer's behavior and the attrition rate?

Our model of Churn prediction for big data underwent two types of prediction modeling; a) using a big data mining approach called Decision Tree with the growing method known as *Classification and Regression Tree* (CRT), and b) deploying the classical General Linear Model (GLM) analysis using SAS system on a selected group based on CRT. Gandomi and Haider (2015) argue that prediction analytics compromise a variety of techniques that predict future outcomes based on historical/archival and current data. These techniques may include the classical linear regression analysis, in conjunction with machine learning tools such as neural network and/or data mining techniques such as Decision Tree. In particular, CRT helps big data mining for classification and "segmentation of population into meaningful subsets" (Lemon, Friedmann, & Rakowski, 2003:172). As mentioned by Wei and Chiu (2002), data mining refers to a process of extracting previously unknown, valid and actionable patterns or knowledge from large databases, involving methods at the intersection of machine learning, statistics, and database systems for critical business decision support.

Based on the literature and in order to develop a model for churn prediction using big data techniques, the following hypothetical model, as depicted in Fig. 5, was set up. We investigate the following hypothesis to answer the main question:

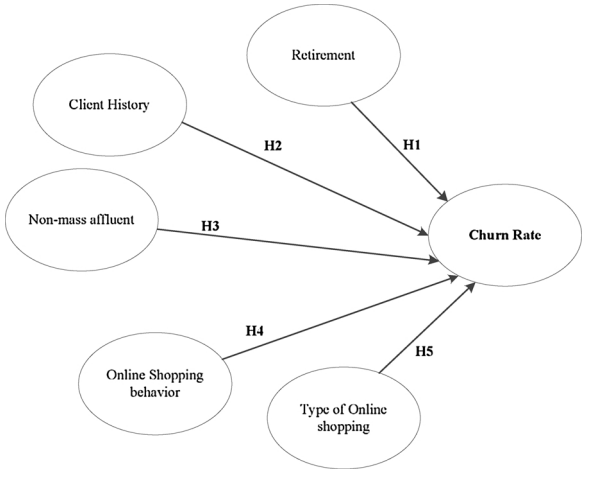


Fig. 5. The hypothetical research model.

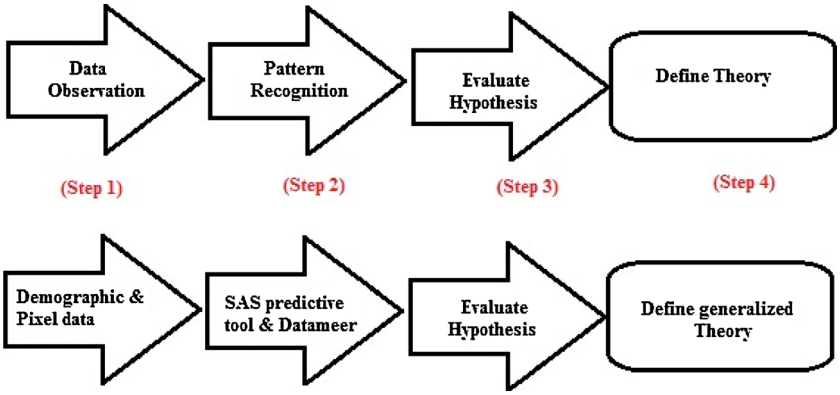


Fig. 6. Inductive Research Approach.

H1. Churn rate has a significant correlation with the retirement event

As argued by McNeal (1999), a characteristic of the financial services industry is that banking and especially insurance products are perceived as "adults only" products (cf. Larivie`re & Van den Poel, 2007). This study investigates the retention proneness of customers during important lifetime events, namely the retirement stage, which is characterized by the need to generate income from previous investments and to maintain the standard of living during the senior years (Larivie`re & Van den Poel, 2007). Based on the historical data, it is known to the target bank that the customer churn rate is correlated with the events of retirement. When it comes to retirement planning, there are steps one may want to take, including but not limited to planning an extended trip abroad, investment plans, a major renovation of the existing property, financing a new vehicle or supporting their children (e.g., education, business start-ups, marriage).

H2. The length of the client history has a significant correlation with churn rate.

A long-term customer has been able to establish mutual trust with their financial provider, despite the fact they are not necessarily a loyal customer (Lejeune, 2001). Bhattacharya (1998) argues that the length of time for which one has been a member (a customer) indicates the base level of satisfaction with the organization. Therefore it is expected that the length of customer history with their financial institute is influenced, among others, by the availability of good service, relevant content in an enjoyable and satisfactory context (Khailfa & Liu, 2007; Koufaris, 2002)) and the trust established between the customer and the financial planner. As argued by Kandampully (1998), in order to protect their long-term interest, service organizations such as the target bank are seeking ways to forge and to maintain an ongoing relationship with their customers by offering service loyalty, thereby effecting responsive and sustained patronage.

H3. The churn pattern of mass affluent clients is not the same as non-mass affluent.

According to the Bank Administration Institute (BAI) report of 2014, mass affluent clients are difficult to be penetrated due to their specific knowledge base and their expectations of the market and the return on their investments. Therefore, gaining a thorough understanding of their expectations and their investment criteria is necessary to not only successfully attract them but also to retain them (Bank Administration Institute, 2014). As such, it is expected that the mass affluent clients who usually have a stronger relationship with their financial planners remain longer with their financial institutes, compared to the non-mass affluent segment.

H4. Customers' online shopping behavior has a significant correlation with the churn rate.

Koufaris (2002) argues that a fundamental difference between online and offline consumer behavior is that the online consumer is generally more powerful, demanding, and utilitarian in their shopping expeditions. As such, it is expected that the intra-relationship among online customers impact their shopping habits via the process of information sharing. Khailfa and Liu (2007) argue that the effects of online shopping habits on online repurchase intentions are two-fold: (1) mediated through satisfaction and (2) moderating the relationship between satisfactions and repurchase intentions (p.783). As such, it is expected that customers with a high rate of online shopping habits are more likely to search for competitors' products and churn with a higher rate, compared at the other retiree groups.

H5. The type of shopping behavior has a significant correlation with the churn pattern.

A related hypothesis to H4 is the type of online activities namely: online banking (e-banking) and the traditional non-online (i.e., brick-and-mortar) banking. As such, it is expected the attrition rate amongst the non-online banking, retiree segment to be higher than those who used online banking and the information seeking pages. Lassar, Manolis, and Lassar (2005) suggest that the type of consumer innovation matters in understanding the adoption of e-banking processes. This supports the notion that online shoppers are distinct from traditional non-online shoppers since they possess more experience in searching for products and service information online. The empirical study of Dwivedi, Rana, Jeyaraj, Clement, and Williams (2017) found that attitude towards IS/IT innovations plays a central role in the acceptance and use of technology. Also, this implies that individuals may use innovation based on the strength of their attitudes, even though they may not consciously intend to use the innovation (Dwivedi, Rana et al., 2017).

4. Method

4.1. Approach to the research question

As shown in Fig. 6, we follow an inductive approach as our study begins with data collection and observation in order to support or reject the hypotheses and finally define the theory. Our panel data is a combination of archival demographic data constructing the population, as well as pixel data regarding different websites and the external products that the specific client has visited during the period of this study.

The first two steps are essential in supporting the hypothesis and defining a defendable general theory; therefore a comprehensive cue analysis was performed to define and assess all the possible cues. This result was concluded after evaluating different behavioral cues, which are considered churn signals. The cues, were generated using various documents, and information through the target bank's website, including the Frequently Asked Questions (FAQ), emails, search engines,

weblogs, as well as user demographics, contractual data, call center and the financial planner notes of interacting with customers, the exact type, time, and duration of each interaction (Verbeke et al., 2014) through the big data setup as depicted in Fig. 8. As suggested by Ngai (2005), as well as assessing the likelihood of predicted events for every client within the population, clients will be assigned to one of the following sub-groups:

1) Retires @ Target Bank (Retained) – Clients who retire & consolidate all their investments and savings with the target bank
2) Retires @ Target Bank(Partially Churned) – Clients who retire & consolidate most of their investments and savings with OFI, but maintain a banking relationship with the target bank
3) Not Retired – Clients that do not display a retirement event
4) Attrited – Clients who have no active banking relationship with the target bank

4.2. Conceptualization of the model

While there are many well established steps defined by pioneers regarding a predictive model conceptualization, according to Inadomi (2004), Roberts et al. (2012) the process of conceptualization can be divided into two major components: first conceptualizing the problem by understanding the actual research question, which will guide the observation and assessment of events; the second step in this process is the actual conceptualization of the stated model. Drawing, based on the proposed model by Roberts et al. (2012), Fig. 7, illustrates the primary steps in developing a churn predictive model. It is noteworthy to mention that 2 and 3 are representing the two main modeling methods which are used in this study: decision tree and regression analysis.

As argued by Wei and Chiu (2002), Churn prediction is a concern for many industries, including the financial institutes.

Sharma and Panigrahi (2011) argue that churn prediction from the business intelligence perspective is a process under the customer relationship management (CRM) framework with two major analytical modeling tasks. Firstly, the prediction of those customers who are about to churn and the second task is assessing the most effective way that a service provider (e.g., a financial institute) can offer special promotion programs or do nothing (Sharma & Panigrahi, 2011). This study intends to assist the target bank in its quest for offering promotional programs to retain its customers in the retiree segment by using a data mining technique which is called a decision tree. This allows the target bank to improve the efficiency of retention campaigns which aim to prevent customers from churning, by directing personalized retention efforts to the customers that are effectively about to churn (Verbeke et al., 2014). Churn prediction and management is a concern for many industries, but it is particularly acute in a strongly competitive environment (Wei & Chiu, 2002) such as the banking industry.

Wei and Chiu (2002) argue that past research on churn prediction in industries such as banking and telecommunication employed classification analysis techniques for the construction of churn prediction models, using predictors such as user demographics, contractual data, call center data such as communication logs describing the identity (i.e. the phone number) and operator of interacting subscribers, and the precise type, time, and duration of each interaction (Verbeke et al., 2014).

4.3. Population

The target population for this study consists of all mass and non-mass affluent clients of the target bank between the ages of 50 to 71 years old, with one or more open and active products and who are not partially or fully retired at the beginning of the analytic window, as we want to observe the retirement event. This analysis is based on an Analytic Window from November 1, 2011, to September 30, 2015, to provide a large enough population in order to conduct behavior analysis.

Although the main purpose of this paper is to examine the churn behavior of mass-affluent clients, to address all the hypotheses; high-net-worth and non-mass affluent clients who share homogeneous characteristics, such as age or shopping behavior are also included in this study. The target population and universe are being used interchangeably throughout this paper.

4.3.1. Age criteria

The first criteria for the target population selection are based on the critical ages for retirement, for instance, age 50 is identified as the retirement age for most of the personal clients who own a business and when their business transactions begin to decrease. On the other hand, 55, 60 and 65 are the critical ages for personal clients who are identified, based on the first deposit of pension payouts into their Direct Deposit Accounts (DDA) or transferring Registered Retirement Savings Plan (RRSP) funds into a Registered Retirement Income Fund (RRIF) or any other type of retirement income fund. In order to validate the retirement age, currently retired clients of target banks are examined and segmented, based on their age and also transaction accounts.

4.3.2. Exclusions criteria

Based on the age criteria, the initial population included more than 3 million clients. Though, further analysis revealed that more consideration is required to minimize the bias in the dataset as much as possible. Therefore, the following exclusions are applied to the initial

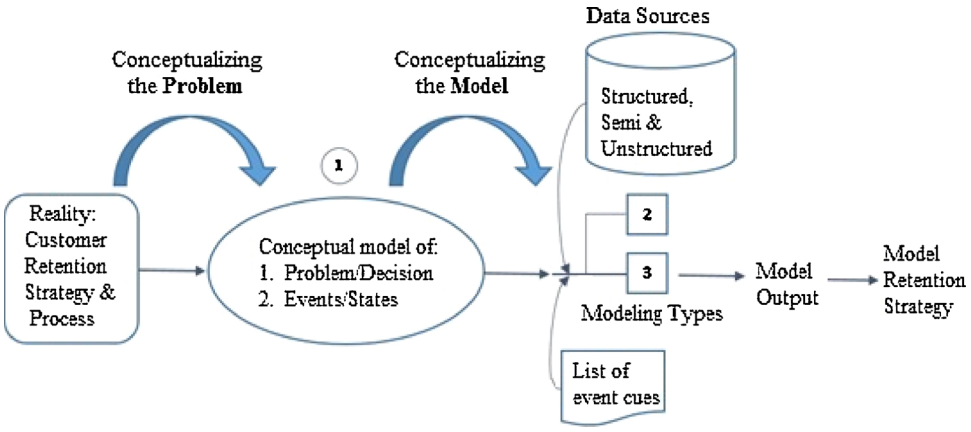


Fig. 7. Schematic presentation of the development of the Churn Prediction Model.

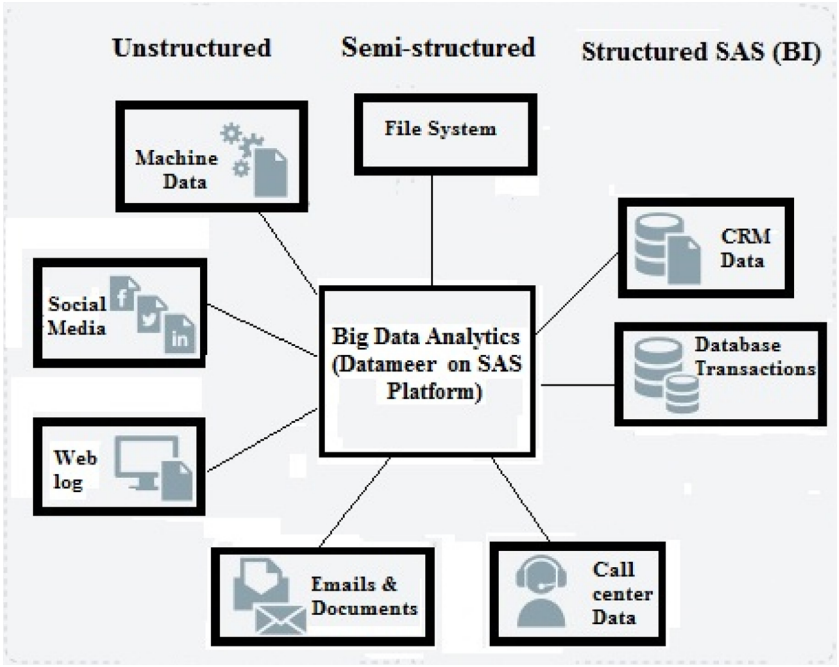


Fig. 8. A schematic view of multi-structured data.

population of 3,144,132, to ensure there are sufficient observation months for the journey development during the analytic window (November 1, 2011, to September 30, 2015) to create the final universe for this study. A quick analysis of the target population revealed that the three months' staying power of behavior models, which was discussed by Neslin et al. (2006) is also valid for our population. Therefore, the following criteria were applied to the initial Retirement Universe to exclude all the clients who had less than three months' history.

A Clients with retirement dates prior to February 1, 2012, OR retirement dates after June 30, 2015
B Clients who attrited prior to February 1, 2012
C Clients with fewer than 3 months of observations in the analytic window
D Clients with < 3 months of observations before retiring
E Clients with < 3 months of observations before attrition
F Clients with < 3 months of observations before the end of the analytic window

4.3.3. Population segmentation

In order to assess the possibility of a correlation between attrition and customer loyalty, the retirement and attrition behaviors were captured through different ways in which each client joins the Universe. Therefore, the Analytic Universe is segmented, based on following three groups of clients: Existing Clients, New to the bank and also those clients who are new to the Universe.

A **Existing Clients** who represent ~ 92% of the Universe consist of all clients between the ages of 50–71, who have an existing banking relationship with the target bank as of Oct. 31, 2011, and joined the Universe as of Nov 1, 2011. This population includes active, personal clients who may or may not have a Direct Investing profile, and with at least one or more active/open products. This group does not include those who are already retired prior to Nov 1, 2011, or deceased clients.
B **New to the Bank clients** who represent ~ 3% of the Universe are those clients who joined the Universe during the Analytic Window and were not bank clients prior to Oct 31, 2011.
C **New to the Universe** This category covers existing bank clients who did not reach the critical ages of retirement at the beginning of the analytic window, but later on became eligible to retire within the period of this study. This group represents ~ 5% of the Universe.

5. The big data model

Decision support systems (DSSs) are an integral part of today's business intelligence (BI), in which the structured, semi-structured and unstructured decision problems integrated with big data analytics tools, allows for organizations to better understand, process, analyze and predict less well structured and underspecified customer experiences. Big Data customer analytics, powered by modern BI platforms allow organizations, and in particular financial institutions, to effectively mine into a large set of data in order to better understand, analyze and predict less well structured and underspecified customer experiences. It offers new insights into decision making and innovation (Negash, 2004; Phillips-Wren, Iyer, Kulakarni, & Ariyachandra, 2015). By deploying the Datameer software (Datameer, 2016), integrated with the target bank's SAS business intelligence platform (see Fig. 8), our study was able to uncover a wealth of data and information, which are associated with nearly 3 million customers' records from 2011 to 2015.

5.1. Structured and unstructured data

Big data has become a popular term to describe the exponential growth, availability and use of information, both structured and unstructured (SQL and NoSQL) (Alshammari, Bajwa, & Lee, 2015).

Santos et al. (2017) argue that big data is often seen as a catchword for smarter and more insightful data analysis. However, it is more than that, it is about new challenging, large and complex data (Dwivedi, Janssen et al., 2017) sources helping to understand business at a more granular level, creating new product or services and responding to business changes as they occur. This includes various sources of data

from large databases to call centers, machine and metadata, configuration file systems, log files, documents, emails, web and social media data, Internet search index, XML, among others, that are arriving at high volumes with characteristics such as velocity (e.g. the speed of time-sensitive data, collected, processed and analyzed), variety (e.g. data in multiple formats) variability (event-triggered loads such as social media), value proposition and complexity (Dwivedi, Janssen et al., 2017). In this context, by analyzing large and feature-rich data, organizations, including the financial institutes, seek to store and analyze greater levels of transaction details, as well as web and machine-generated data, to gain a better understanding of customer behaviour and business (Sharda, Delen, & Turban, 2018), and meet the competitive challenges in developing new product and services (Santos et al., 2017) that they may not have otherwise. Associated with big data is its technical infrastructure (hardware and software) that allows parallel processing and the deployment of non-relational storage capabilities in order to process structured, unstructured and semi-structured data, as depicted in Fig. 8 above. For the sake of this study, the target bank deployed a third party version of the Hadoop ecosystem, integrated with the SAS system. SAS 9.4 offers an Intelligence Platform (SAS, 2016) called SAS Data Loader for Hadoop web application, in which it allows SAS analytical programs to run in a Hadoop environment. For real-time, data stream is stored into a NoSQL database (Santos et al., 2017) namely Hadoop's HBase. Even though Hadoop is an open source framework for processing massive amounts of distributed, unstructured data (Sharda et al., 2018), the target bank deployed a commercial version of this ecosystem called Datameer that supports the Hadoop Distributed File System, as depicted in Fig. 8. The system, in particular, integrates NoSQL (Not only SQL) to process large volumes of structured, semi-structured and unstructured data on the SAS platform.

The structured and unstructured data for the target bank include the archival data on the client demographics and banking products to identify the research population; online data and Interactive Voice Response (IVR) data for the identified population. The online data is mostly the pixel data which tracks clients' behavior on various websites. The provided pixel data demonstrates, among others, how many times a client visited an external service or product, which is either offered or endorsed, by the target bank. This data also provides visibility on each client's online usage.

5.2. Data limitation

There are two crucial and problematic challenges associated with churn data, the imbalance in the data distribution (Xie, Li, Ngai, & Ying, 2009) and the structure of data. The most important limitation in this study is the data related to the retirement and attrition status, especially for those who fall under the category of business clients or that retired at other financial institutes, as the current process is based on self-reported status, but apparently, many clients do not call the bank to change their status from employed to retired. Personal clients who have a Direct Deposit Account (DDA) can be identified as soon as the first CPP payment is made to their personal DDA accounts. However, business owners and other personal clients who do not hold a DDA account with the target bank are not easily detectable, mainly because most of the churners retain one or two active accounts with the target bank, while the majority of their retirement investment is already transferred to another financial institute. To address this limitation, the CLV measurement technique, as well as the calculation of Money-In (MI) balance, which is currently being used as an indicator to distinguish between active and passive clients, is used in this study to predict the retirement status. These techniques are discussed in the Retirement Consolidation Model.

5.3. Pattern recognition cues

In order to identify the retirement event and also the prediction of possible churn events, a wide range of cues was analyzed to identify the most relevant ones. Cues are categorized into four different groups: Life Style, Client Feedback, Information Seeking, and Behavioral cues.

After applying the above cues on the eligible population, it is first divided into two groups of retired and non-retired clients, and then each group would be further divided into two subgroups of retained and attrited; the outcome of this analysis is discussed in the results section. As stated in the data limitation section, the major obstacle in any retention planning strategy is to target the right clients. Since many churners retain at least one active account with the target bank while they have already transferred all of their investments out, it is very likely that some of the clients who were identified as retired and retained are already churned. As a result, implementing a consolidation model is necessary to differentiate loyal clients (those who retired and consolidated all their investment with the target bank) from the actual churners. This model is the subject of the next section of discussion.

5.4. Retirement consolidation model

In order to increase the accuracy rate of our findings and prevent any biases in data, the CLV and MI techniques are applied to the Retired and Retained (R&R) segment to separate loyal clients from churners and also, partially churned clients. Henceforth, the partially churned clients who can be considered as potential clients are labeled Committed at Risk, as their continuance relationship with the bank depends on their level of satisfaction, as well as the competence level of products offered by the target bank. The outcome of this investigation will help to reduce the churn rate after retirement and will increase the client's share of wallet to the optimal level.

Bauer, Hammerschmidt, and Braehler (2003) discussed a comprehensive formula for calculating the CLV for each client. Although they have used many different variables to calculate the revenue and also the cost of each client in a specific period, their recommended formula stems from three basic elements:

I Client NPV (Net Present Value) over time = (all Revenues – all Costs)
II Length of Service (LoS)
III Discount Factors which is used in the NPV calculation

Although the calculation of the CLV is the preferred method for identifying the most valuable clients, the data for the population was collected only for the past three years; therefore the calculated value of the existing clients who form the majority of the population does not represent the actual values of these clients. To address the issue and eliminate any bias that may occur due to the flaw, the current Money-In balance is compared to the future Money-In balance, based on the forecasted cash flow technique over a period of 12 months. Hence, the change in MI balance 6 months prior to retirement (-6 months) and 6 months post-retirement (+6 months) is used to determine if the consolidation of assets occurred with the target bank or other financial institutes.

Money In (MI) balances include all the money a client holds in various accounts, such as Checkings/Savings/eSavings, Guaranteed Investment Certificates (GIC), Mutual Funds, and Direct Investing (stocks, ETFs, Mutual Funds, and so forth).

In order to categorize clients based on their MI, the first step is to define the criteria, such as the minimum/maximum balance and also the growth rate. Therefore, the result of this analysis revealed that those clients who have less than $4000 in their account for a period of 6

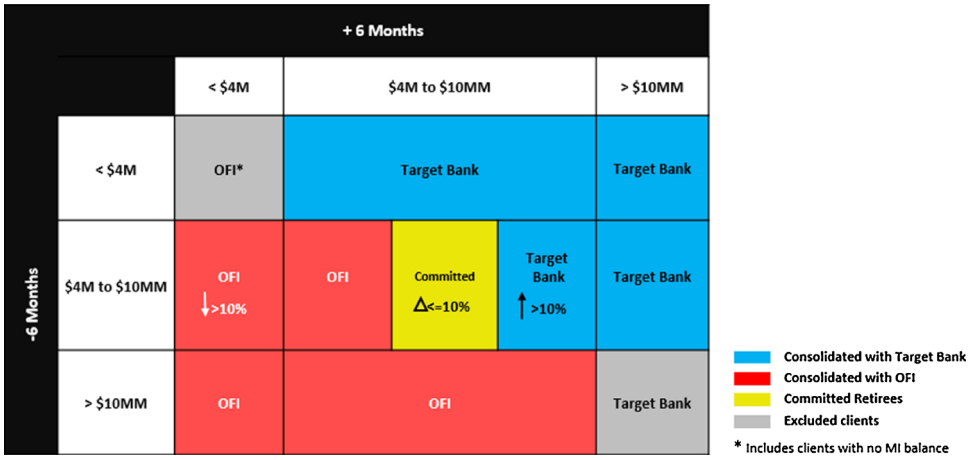


Fig. 9. Schematic view of Retirement Consolidation Model.

months, either before or after their retirement, are most likely inactive clients with no severe banking relationship with the target banks, regardless of retiring with it, meaning that these are already attrited. This analysis also revealed that clients who hold more than $10,000,000 in their balance for the above period are high net-worth clients who most likely will remain loyal. Thus, all the clients who fall under these two extreme categories were excluded from the study.

Further analysis shows that clients who increased their MI balance by more than 10% within 6 months post retirement are most likely to remain loyal and retained, whereas those who decreased their MI by more than 10% are in the process of attrition and migration to other financial institutes. Therefore, as illustrated in Fig. 9, clients with starting and ending MI balances between $4 M and $10 M are categorized based on the following:

- If the MI balance increases by more than 10% then Consolidated with the target bank
- If the MI balance decreases by more than 10% then Consolidated with OFI
- If the MI balance increases/decreases by less than 10% then Committed Retirees

6. Analysis and results

6.1. Big data analysis of client segmentation

After building a predictive model, it is essential to evaluate the performance of the classifiers. In order to assess the churn behavior, the following steps were applied to develop segmentation for the retirement journey analysis. The result of this segmentation is summarized in Fig. 10.

As noted by Lemon et al. (2003), CRT, as depicted in Fig. 10, begins with one node (root) containing the entire sample and it then examines all possible splitting variables to the root by selecting the most different according to a binary tree pattern. By applying CRT to various groups within our sample, we were able to identify cues. As shown in Fig. 10, out of 2,813,276 clients, approximately 10.45% (294,036) fall under the retired group, while the remaining 89.55% of clients are part of the non-retirees sub-group.

Second, by utilizing attrition cues, each group is further divided into two sub-groups of "Churners" and "Non-Churners." The result of this analysis for retired clients revealed that 8.7% of those clients who retired with the target bank have already attrited. However, more dissection was necessary for the non-retirees group as some of the clients in this group were showing signs of being passive clients. As a result,

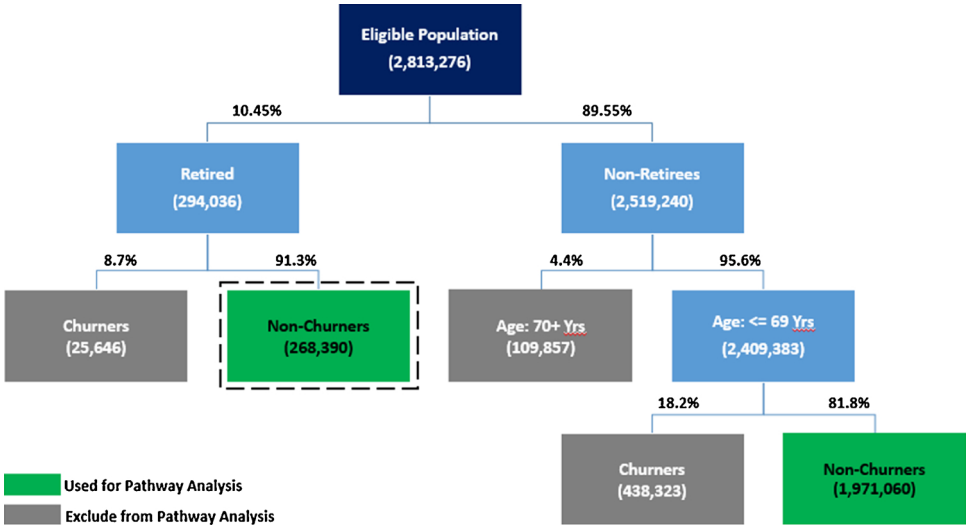


Fig. 10. CRT analysis of Target Bank population.

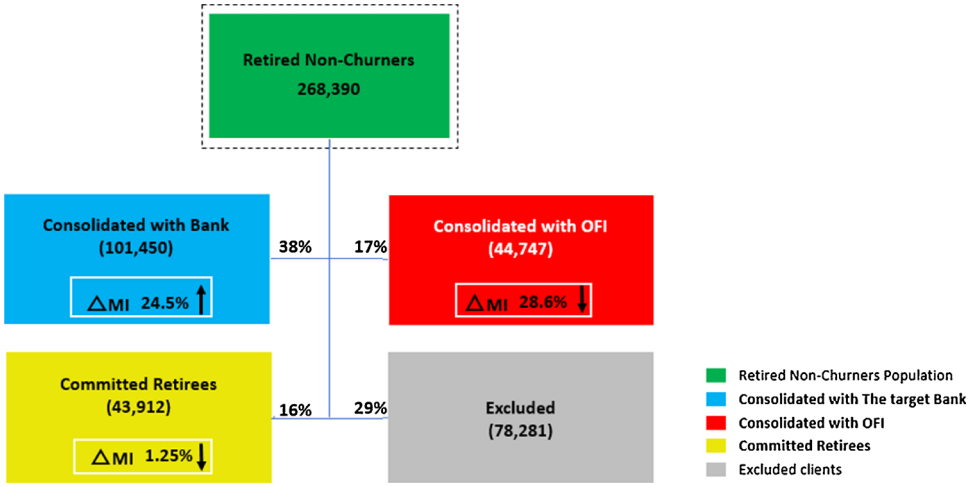


Fig. 11. Schematic view of Retirement Consolidation Results.

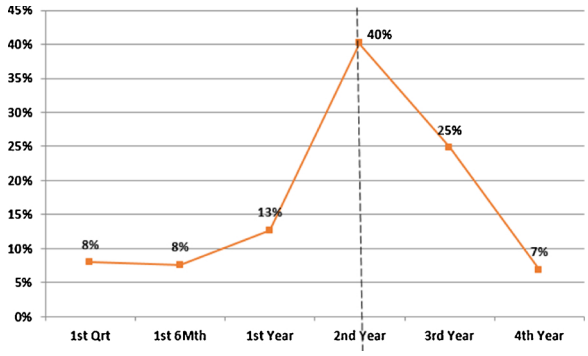


Fig. 12. Trends for Attrition Post Retirement.

about 109,857 clients, who account for 4.4% of non-retirees, are clients over the age of 70 with declining Money In balance. We assumed these clients have already retired with other financial institutes and are in the process of migrating to that financial institute. Therefore, these groups were excluded from further analysis.

Consequently, prediction analysis was applied on the remaining 95.6%. The result of this analysis confirmed that 18.2% of these clients are potential churners, while the remaining 81.8% are retained and loyal clients who will retire and remain with the target bank.

6.1.1. Retirement consolidation result

The Money-In balance of the Non-Churner clients in both retirees and non-retirees groups account for more than $140 billion, which are worthy of further scrutiny to ensure that the result of this study reflects the maximum accuracy. Accordingly, as described in the design model section, the Money-In balance technique was applied to 268,390 clients who were identified as retired and non-attrited to identify the actual loyal clients from potential churners. The result of this analysis is presented in Fig. 11.

After excluding 29% of this group with balances either less than $4000 or more than $10,000,000, the remaining clients were divided into three sub-groups: 17% of retired clients have already churned to other financial institutes and kept a low-level relationship with the target bank; 16% of clients fall under committed at risk who may or may not churn, depending on the future retention strategies. However, the remaining 38% demonstrated a high level of loyalty and consolidated all their investment with the target bank.

The result of this analysis is highly notable because as per our preliminary analysis and applying existing churn prediction techniques such as survival analysis, which is based on a binary method, only 8.7% of the retired population was identified as churned clients. However, after implementing the new model, based on big data analysis techniques and identified behavioral cues, the actual churn rate calculated as 24% of the entire retiree population.

The above finding is an indication of 175% improvement in accuracy of the result, which is highly remarkable for designing a successful retention strategy. As argued by (Verbeke et al., 2011), customer churn prediction models aim to detect customers with a high propensity to attrite. Predictive accuracy, comprehensibility, and justifiability are three key aspects of a churn prediction model. An accurate model allows for correctly targeting future churners in a retention marketing campaign, while a comprehensible and intuitive rule-set allows identifying the main drivers for customers to churn, and to develop an effective retention strategy in accordance with domain knowledge

6.2. Empirical results

The results of this study confirmed that clients' behavior has a significant correlation with their churn decision. This result was concluded after evaluating different behavioral cues, which are considered churn signals. Although mass affluent clients are twice more likely to retire after reaching the critical ages of 50, 55, 60 and 65; however their attrition rate is significantly lower than non-mass affluent clients. In the following sections, each hypothesis is discussed in more detail, referring to the statistical and correlation analysis which was conducted in SAS through a combination of T-Tests (Proc Freqs in SAS) and Anova (Proc GLM in SAS). The result of each hypothesis assessment is included in Appendix A and B.

H1. Churn rate has a significant correlation with the retirement event.

Based on the conducted analysis, this research supports the first hypothesis that the churn rate has a significant correlation with events of retirement. As discussed, the attrition rate significantly decreases after a retirement event occurs; however, as illustrated in Fig. 12, the churn rate spikes between year 2 and 3 after retirement amongst those who retired and then attired. It can be explained based on the fact that

it generally takes one year for retirees to begin managing and enjoying the new chapter of their life, and usually, after year one, they begin to assess their investment portfolio to ensure it addresses all their retirement needs.

Referring to the result of the statistical analysis in Appendix A – Fig. A1, also confirms the significant correlation between churn and retirement events due to the result of Chi-Square and also P-Value equal to Zero. As seen in the frequency behavior, the churn rate prior to retirement is 14%, while this number remarkably decreases to 4.5% after retirement occurs. Fig. A2 in Appendix A shows the result of frequency and also normal probability plot in SAS, which indicates among retired and attrited clients, year 2 has the highest rate of attrition.

H2. The length of client history has a significant correlation with the churn rate.

Same as H_1, the preliminary analysis also supported the H_2 hypothesis. As discussed earlier in this paper, the length and strength of a banking relationship have a significant correlation with churn decisions. This also can be concluded from the result of ANOVA test through the GLM procedure in SAS. As observed in Appendix A- Fig. A3, the correlation between attrition and tenure, which was considered as a dependent variable, is significantly high due to the P-Value being less than 0.0001. In this test, we have noticed some outliers in the data, which could have impacted the result; therefore, all the clients who have higher than 45 years' of tenure were excluded from this analysis.

H3. The churn pattern of mass-affluent clients is not the same as non-mass affluent.

This study also supported the H_3, as non-mass affluent clients churn at a much faster rate, compared to mass affluent clients. This can be explained based on the deep relationship that mass affluent clients have with their financial planners. The result of the frequency test and also Chi-Square in Appendix A- Fig. A4, confirms this fact. The type of client that was selected as the dependent variable and the correlation between being either mass-affluent or mass-retail and churn rate were examined. As seen, 28% of mass-retail or non-mass-affluent clients were attrited in total, while this number drops to only 7% among mass-affluent clients. The Chi-Square and P-Value result of near zero also confirms the strong correlation between the type of clients and churn decision.

H4. Customers' online shopping behavior has a significant correlation with churn rate.

The results show that the attrition rate is not changed, based on the online usage within the retiree segment. Although, this study did not support H_4; that online shopping behavior and the number of visits to retirement products have a significant relationship with the churn rate. However, as previously discussed, due to the nature of our population and the usage of online products, more online data and scrutiny is required before we can entirely reject this hypothesis.

H5. The type of shopping behavior has a significant correlation with the churn pattern.

In order to determine whether there is any difference between the churn rate of clients who use only online banking for their day-to-day activities and those who expand their usage to information seeking and research domains, some frequency tests were applied. Since the retirement universe of this study did not satisfy the required data for online usage, these T-tests were applied on not only retirement universe, but also the entire client base between November 1, 2014, and November 30, 2015.

Appendix B- Fig. B1 examines the attrition rate amongst all clients who used online banking, including transactions such as check balances and bill payments or seeking information and searching, such as looking at mortgage rates, using calculators, and so forth. In order to observe the relationship between the type of usage and attrition rates, while Fig. B2 focuses on clients who used online banking for transaction purposes only, Fig. B3 focuses on information seeking records.

As seen, the attrition rate amongst non-online clients is 7%, while this number drops to 1.6% between those who used online banking or information seeking pages. By comparing the results, it is concluded that attrition among those who used online channels for research is slightly less than those who used it only for day-to-day banking. The 0.3% difference is not significant to conclude that online activities and information seeking pages contribute to lower attrition rates.

6.3. Limitations and future directions

Similar to other studies, this study also faces some limitations. First and foremost because of the nature of our population, the available behavioral data related to clients' online research did not satisfy the result. The results of this study will be different in the next few years when the younger generation reaches their retirement stage, as the Internet and social media usage, in particular, is inevitable among this group.

The only pixel data that was procured for this study was the tracking of clients' visits through different retirement pages of target bank websites and the search engines to search for their desired products and services, therefore this study could have been more accurate if we could track our clients' online behavior through social media sites such as Facebook. There is a rich set of research articles covering the importance of social media in the context of business. For example, Pogrebnyakov and Maldonado (2018) point out that social media allow users to create content, establish connections with other users and share content with other users. They also offer several functionalities and characteristics, such as visibility of information, the establishment of connections and sharing of information and knowledge in the form of text, images, videos and web links among others. In particular, it has become an important driver for acquiring and spreading information in different business sectors today (Stieglitza, Mirbabaiea, Rossa, & Christoph Neubergerb, 2018). In this context, social media has been largely realized as an effective mechanism that contributes to the firms' marketing aims and strategy; especially in the aspects relating to customer relationship management and communication (Alalwan, Rana, Dwivedi, & Raed, 2017; Alalwan, 2018); and the big data created through social media has secured a prominent position in almost all industries (Raginia, Anandb, & Bhaskar, 2018) including the financial institutes. However, the privacy implications for both customers and the target bank were one of the primary reasons for not compiling this type of data.

6.3.1. Future directions

Few recommendations to the department of marketing and strategy of Personal and Commercial Banking (P&CB) of the target bank can be made, based on the findings of this study:

As discussed, 77% of the population demonstrated a significant opportunity for the target bank; therefore, designing and implementing a comprehensive customer-centric retention strategy, based on big data analytics, is highly recommended to ensure that this group of the population will retire and retain with the target bank. The implementation of the stated strategy will assist to decrease the attrition rate significantly and would maximize shareholder's value and the bank's profitability rate.

Designing and implementing different retention strategies for the retiree segment is another area for opportunity. If clients receive their customized and desired services and products when they most need it,

they would be less likely to attrite after retirement. This will help to decrease the post-retirement attrition. Providing the right incentives to strengthen the banking relationship, especially amongst those clients who are new to the bank, is considered a key retention factor.

The need for a separate marketing campaign is necessary to encourage current non-retiree clients to improve or strengthen their banking relationship with the target bank. This can be achieved by offering different loyalty and rewards programs.

As confirmed, the constructed model has three months' staying power. Therefore it is highly recommended to run the model every three months to; a) capture new entries to the population, b) evaluate the status of current clients and to compare the results by determining whether there have been any changes in churn decisions, especially amongst potential churners.

The Non-Retiree segment should also be analyzed using the MI data to distinguish true loyal clients from churners and potential churners. Since the current study is the first phase of an end-to-end Client Knowledge & Insights (CKI) analysis initiative, after implementing recommended retention strategies, the churn rate amongst those who were identified as potential churners should be measured to evaluate the success rate of this program.

The importance of discovering the right receptive point, which indicates when a specific client is naturally more receptive regarding retirement advice, should also be emphasized as part of future studies. This can be achieved by performing a sensitivity analysis and ascertaining out when the best time to tap into the mass market would be, in order to win the "moment of truth."

7. Conclusion

The expected results of this study support the above hypotheses based on the identified backing and warrants. Subsequently, the main purpose of the conducted research supports the claim that offering the right product at the right moment can minimize churning. This study is also attempting to verify some sub-claims such as a) the number of visits to external websites is an indication of churn decisions and/or b) the longer a client has a history with a bank, the lower the likelihood of churn.

Since the Marketing and Strategy department within the Canadian Banking division of the target bank will consume the results of this study, this study also explores the time of the churn, relevant to the retirement date, in order to provide a clear roadmap for the marketing horizon.

Furthermore, this study scrutinizes the churn pattern of mass affluent clients versus non-mass affluent clients, to identify the right strategy for each market, which will maximize the rate of retention.

This study has several strengths. According to the current literature, it is the first study that uses CRT in conjunction with linear regression analysis for implementing the churn prediction model on big data, incorporating different sources of data such as pixel data, IVR logs, phone conversations, and financial data to examine the possibility of churn for each client. Utilizing a big data analytics tool such as Datameer for developing the current churn prediction model improved the result by 175% compared to traditional churn prediction models.

According to the conducted analysis, clients' behavior has a significant correlation with churning decisions. This finding has a considerable impact on designing effective client relationship management strategies in the future because the higher the rate of accuracy in detecting and identifying potential churners, it will result in higher rates of success for more effective and efficient retention strategies.

This study confirms that mass-affluent clients are twice more likely to retire with the target bank and two times less likely to churn from the target bank. Although the attrition rate within mass affluent segments is notably lower than mass retail, as this segment owns a multi-billion dollar market, attrition or retention of clients within this segment has a significant impact in a bank's profitability. Even though the churn rate significantly drops after retirement, but among those who attrite, most of the attrition occurs between years 2 and 3 after retirement.

Clients who have longer tenure and deeper banking relationships with the target bank are less likely to churn; therefore it is concluded that deepening the banking relationship will minimize the rate of attrition. "New to the bank" clients are more likely to maintain their relationship at the borrowing level or transaction only level. While borrowing only clients demonstrated high rates of interest to deepen their relationship after retirement, transaction only clients do not show any significant changes to their banking relationship and are most likely to churn after a while.

The online usage within the retiree segment did not show any correlation with the rate of attrition, due to low usage of online banking and information seeking channels by these particular clients.

Appendix A

Table of Retire_Ind by attr_typ

Retire_Ind attr_typ

Frequency / Percent / Row Pct / Col Pct	NON-Attr	Ret	Total
NON-Attr	2176823 71.19 81.03 86.24	509617 16.67 18.97 95.49	2686440 87.86
Att	347191 11.35 93.52 13.76	24055 0.79 6.48 4.51	371246 12.14
Total	2524014 82.55	533672 17.45	3057686 100.00

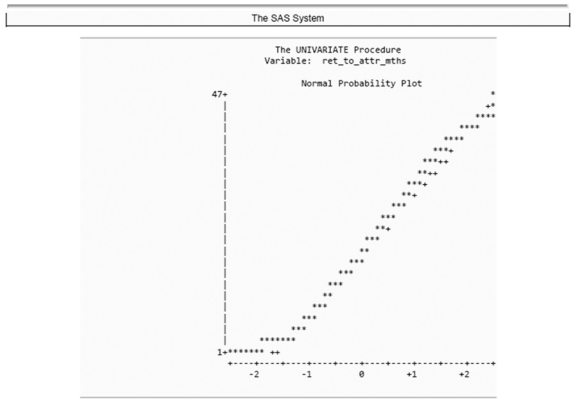
Statistics for Table of Retire_Ind by attr_typ

Statistic	DF	Value	Prob
Chi-Square	1	35320.0567	<.0001
Likelihood Ratio Chi-Square	1	43194.4035	<.0001
Continuity Adj. Chi-Square	1	35319.1898	<.0001
Mantel-Haenszel Chi-Square	1	35320.0452	<.0001
Phi Coefficient		-0.1075	
Contingency Coefficient		0.1069	
Cramer's V		-0.1075	

Fisher's Exact Test

Cell (1,1) Frequency (F)	2176823
Left-sided Pr <= F	.
Right-sided Pr >= F	.
Table Probability (P)	.
Two-sided Pr <= P	.

Fig. A1. Correlation between Churn Rate and Retirement Event.



The FREQ Procedure

RtoA_ Delta	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Fst_1Qtr	2161	8.98	2161	8.98
Fst_6m	1935	8.04	4096	17.03
Fst_Yr	3252	13.52	7348	30.55
Fourth_Yr	1524	6.34	8872	36.88
Scnd_Yr	9477	39.40	18349	76.28
Thrd_Yr	5706	23.72	24055	100.00

Fig. A2. Churn Rate after Retirement.

The SAS System

The GLM Procedure

Class Level Information

Class	Levels	Values
ind_attr	2	0 1

Number of Observations Read 3036307
Number of Observations Used 3036307

The GLM Procedure

Dependent Variable: tenure_yrs tenure_yrs

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	13946419.5	13946419.5	99002.9	<.0001
Error	3.04E6	427720794.4	140.9		
Corrected Total	3.04E6	441667213.8			

R-Square	Coeff Var	Root MSE	tenure_yrs Mean
0.031577	61.19462	11.86882	19.39520

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ind_attr	1	13946419.45	13946419.45	99002.9	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ind_attr	1	13946419.45	13946419.45	99002.9	<.0001

Fig. A3. Correlation between Attrition and Tenure.

Table of ind_ma by ind_attr

Frequency / Percent / Row Pct / Col Pct	Non-Attr	Attr	Total
Non-Affl	1766359 59.20 78.22 71.93	491946 16.49 21.78 93.13	2258305 75.68
Mass-Affl	689342 23.10 95.00 28.07	36313 1.22 5.00 6.87	725655 24.32
Total	2455701 82.30	528259 17.70	2983960 100.00

Statistics for Table of ind_ma by ind_attr

Statistic	DF	Value	Prob
Chi-Square	1	106133	<.0001
Likelihood Ratio Chi-Square	1	130503	<.0001
Continuity Adj. Chi-Square	1	106132	<.0001
Mantel-Haenszel Chi-Square	1	106133	<.0001
Phi Coefficient		-0.18859	
Contingency Coefficient		0.18533	
Cramer's V		-0.18859	

Fisher's Exact Test

Cell (1,1) Frequency (F)	1766359
Left-sided Pr <= F	.
Right-sided Pr >= F	.
Table Probability (P)	.
Two-sided Pr <= P	.

Fig. A4. Churn Pattern of Mass-Affluent vs. Non-mass Affluent.

Appendix B. Churn Rate and Online Behavior

The FREQ Procedure

Table of IND_ETHR by attr_ind

IND_ETHR / attr_ind	0	1	Total
0	6604910	513238	7118148
	65.68	5.10	70.79
	92.79	7.21	
	69.55	91.75	
1	2891311	46141	2937452
	28.75	0.46	29.21
	98.43	1.57	
	30.45	8.25	
Total	9496221	559379	1.006E7
	94.44	5.56	100.00

Fig. B1. Churn Rate and Online Usage.

The FREQ Procedure

Table of IND_OLBK by attr_ind

IND_OLBK / attr_ind	0	1	Total
0	6611124	513538	7124662
	65.75	5.11	70.85
	92.79	7.21	
	69.62	91.81	
1	2885097	45841	2930938
	28.69	0.46	29.15
	98.44	1.56	
	30.38	8.19	
Total	9496221	559379	1.006E7
	94.44	5.56	100.00

Fig. B2. Churn Rate and Online Banking Transaction.

The FREQ Procedure

Table of IND_RSCH by attr_ind

IND_RSCH / attr_ind	0	1	Total
Frequency			
0	6873087	523277	7396364
Percent	68.35	5.20	73.55
Row Pct	92.93	7.07	
Col Pct	72.38	93.55	
1	2623134	36102	2659236
Percent	26.09	0.36	26.45
Row Pct	98.64	1.36	
Col Pct	27.62	6.45	
Total	9496221	559379	1.006E7
	94.44	5.56	100.00

Fig. B3. Churn Rate and Online Information Research.

References

Alalwan, A. A. (2018). Investigating the impact of social media advertising features on customer purchase intention. *International Journal of Information Management, 42*, 65–77.

Alalwan, A. A., Rana, P. N., Dwivedi, K. Y., & Raed, A. (2017). Social media in marketing: A review and analysis of the existing literature. *Telematics and Informatics, 34*, 1177–1190.

Alshammari, H., Bajwa, H., & Lee, J. (2015). Enhancing performance of Hadoop and MapReduce for scientific data using NoSQL database. *2015 IEEE Long Island Systems, Applications and Technology Conference* 2015.

Au, W. H., Chan, K. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation, 7*(6), 532–545.

Bank Administration Institute (2014). *Engaging and retaining mass affluent customers.* [. Accessed Febryary 2016] https://www.bai.org/libraries/lob-sps-downloads/massaff.sflb.ashx.

Bauer, H. H., Hammerschmidt, M., & Braehler, M. (2003). The customer lifetime value concept and its contribution to corporate valuation. *Yearbook of Marketing and Consumer Research, 1*(1), 49–67.

Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the Academy of Marketing Science, 26*(1), 31–44.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Coussement, K., Benoit, D. F., & Poel, D. V. (2015). Preventing customers from running away! Exploring generalized additive models for customer churn prediction. In M. Dato-on (Ed.). *The sustainable global marketplace. Developments in marketing science: Proceedings of the academy of marketing science.* Cham: Springer.

Datameer (2016). *Big data integration.* [. Accessed December 2016] https://www.datameer.com/product/big-data-integration/.

De Caigny, A., Coussement, K., & De Bock, W. K. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*, 760–772.

Dwivedi, K. Y., Janssen, M., Slade, L. E., Rana, P. N., Weerakkody, V., Millard, J., et al. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling. *Information Systems Frontiers, 19*(2), 197–212.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2017). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers.* https://doi.org/10.1007/s10796-017-9774-y.

Ekinci, Y., Uray, N., & Ülengin, F. (2014). A customer lifetime value model for the banking industry: A guide to marketing actions. *European Journal of Marketing, 48*(3/4), 761–784.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *Intrenational Journal of Information Management, 35*, 137–144.

Hoekstra, J. C., & Huizingh, E. K. (1999). The lifetime value concept in customer-based marketing. *Journal of Market-Focused Management, 3*(3-4), 257–274.

Inadomi, J. M. (2004). Decision analysis and economic modelling: A primer. *European Journal of Gastroenterology and Hepatology, 16*(6), 535–542.

Kandampully, J. (1998). Service quality to service loyalty: A relationship which goes beyond customer services. *Total Quality Management, 9*(6), 431–443.

Khailfa, M., & Liu, V. (2007). Online consumer retention: Contingent effects of online shopping habit and online shopping experience. *European Journal of Information Systems, 16*, 780–792.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. *Information Systems Research, 13*(2), 205–223.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). *Customer management as the origin of collaborative customer relationship management. In collaborative customer relationship management.* Berlin Heidelberg: Springer3–6.

Larivie`re, B., & Van den Poel, B. (2007). Banking behaviour after the lifecycle event of "moving in together": An exploratory study of the role of marketing investments. *European Journal of Operational Research, 183*, 345–369.

Lassar, M. W., Manolis, C., & Lassar, S. (2005). The relationship between consumer innovativeness, personal characteristics, and online banking adoption. *International Journal of Bank Marketing, 23*(2), 176–199 2005.

Lejeune, M. (2001). Measuring the impact of data mining on churn management. *Internet Research, 11*(5), 375–387.

Lemon, C. S., Friedmann, D. P., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regresion. *The society of Behavioral Medicine, 26*(1), 172–181.

Lewis, B. R., & Soureli, M. (2006). The antecedents of consumenr loyalty in retail Banking. *Journal of Consumer Behaviour, 5*(1), 15–31.

Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2010). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications, 38*, 8–15.

Ling, R., & Yen, D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. *Journal of Computer Information Systems, 41*(3), 82–97.

McNeal, J. U. (1999). *The kids market: Myths and realities.* Ithaca, NY: Paramount Market Publishing.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31.

Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems, 13*, 177–195.

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

Journal of Marketing Research, 43(2), 204–211.
Ngai, E. W. (2005). Customer relationship management research (1992-2002) an academic literature review and classification. *Marketing Intelligence & Planning, 23*(6), 582–605.
Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602.
Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert System with Applications, 38*, 15273–15285.
Phillips-Wren, G., Iyer, S. L., Kulakarni, U., & Ariyachandra, T. (2015). Business analytics in the context of big data: A roadmap for research. *Communications of the Association for Information Systems, 37*, 448–472.
Pogrebnyakov, N., & Maldonado, E. (2018). Didn't roger that: Social media message complexity and situational awareness of emergency responders. *International Journal of Information Management, 40*, 166–174.
Prasad, U. D., & Madhavi, S. (2012). Prediction of churn behavior of bank customers using data mining tools. Editorial Note Words from the Board of Editor, 5(1), 96.
Raginia, J. R., Anandb, R., & Bhaskar, B. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management, 42*, 13–24.
Roberts, M., Russell, L. B., Paltiel, A. D., Chambers, M., McEwan, P., Krahn, M., et al. (2012). Conceptualizing a model: A report of the ISPOR-SMDM modeling good research practices task force–2. *Value Health, 15*(6), 804–811.
Rosset, S., Neumann, E., Eick, U., Vatnik, N., & Idan, Y. (2002). Customer lifetime value modeling and its use for customer retention planning. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 332–340.
Ryals, L. J., & Knox, S. (2005). Measuring risk-adjusted customer lifetime value and its impact on relationship marketing strategies and shareholder value. *European Journal of Marketing, 39*(5/6), 456–472.
Santos, Y. M., Oliveira e Sá, J., Andrade, C., Vale Lima, F., Costa, E., Costa, C., et al. (2017). A big data system supporting Bosch Braga industry 4.0 strategy. *International Journal of Information Management, 27*, 750–760.
SAS (2016). *SAS Data Loader for Hadoop.* [. Accessed February 2017] https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-data-loader-hadoop-107474.pdf.
Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective* (Fourth edition). New York: Pearson.
Sharma, A., & Panigrahi, K. P. (2011). A neural network based approach for predicting customer churn in cellular network services. *International Journal of Computer Applications, 27*(11), 26–31.
Stieglitza, S., Mirbabaiea, M., Rossa, B., & Christoph Neubergerb, C. (2018). Social media analytics—Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168.
Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.
Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354–2364.
Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing, 14*, 431–446.
Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*, 103–112.
Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*, 5445–5449.