MARKET SEGMENTATION USING ENHANCED RFM (RECENCY, FREQUENCY, MONETARY) MODEL

FAHED YOSEPH

UNIVERSITI SAINS MALAYSIA

2017

MARKET SEGMENTATION USING ENHANCED RFM (RECENCY, FREQUENCY, MONETARY) MODEL

by

FAHED YOSEPH

Thesis Submitted in fulfillment of the requirements for the Degree of Master of Science

February 2017

DECLARATION

Name: **FAHED YOSEPH**

Matric No:

Faculty: **SCHOOL OF COMPUTER SCIENCES**

Thesis: **HYBRID OF CLUSTERING AND CLASSIFICATION FOR MARKET SEGMENTATION**

I Fahed Yoseph, the undersigned, hereby declare that the work contained in this research is my own original work and it has never been submitted it at any university other than the University of Saint Malaysia (USM).

Students Signature: ………………………………………...   Date:

ACKNOWLEDGEMENT

I hereby wish to express my sincerest gratitude and appreciation to my supervisors: Dr. Nurul Hashimah Binti Ahamed Hassain Malim and Dr. Manmeet Mahinderjit Singh for giving me this opportunity and sharing their valuable knowledge and guidelines, especially in the area of data mining and Security.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ii

TABLE OF CONTENTS iii

LIST OF TABLES viii

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xiii

ABSTRAK xv

ABSTRACT xvii

CHAPTER 1 -INTRODUCTION

1.1 Introduction 1

1.2 Research Motivation 2

1.3 Research Problem 3

1.4 Research Questions 5

1.5 Research Objectives 6

1.6 Research Scope 6

1.7 Organization of the thesis 6

CHAPTER 2 -LITERATURE REVIEW

2.1 Introduction 8

2.2 Market Segmentation 8

2.2.1 Market Segmentation Bases 9

2.2.2 Market Segmentation Data Sources 10

2.3 Customer Value Analysis.. 11

2.3.1 Customer *Lifetime* Value (CLTV) Model 11

2.3.2 Components of CLTV.. 12

2.3.3 CLTV Retention and Churn Rates 14

2.3.4 CLTV Benefits and Limitation 14

2.3.5 Market Segmentation based on CLTV Model 15

2.4 Recency, Frequency and Monetary RFM Model 16

2.4.1 Benefits of using RFM Model... 17

2.4.2 RFM Analysis Scoring Techniques 18

2.4.3 RFM Scoring Based on Quintiles 18

2.4.4 RFM Weighted Scores 19

2.4.5 RFM with CLTV.. 20

2.4.6 Market Segmentation based on RFM Method 20

2.4.7 RFM Analysis Model Limitation.. 21

2.5 Data Mining 22

2.5.1 Data Mining Methods for Discovering Patterns 24

2.5.2 Classification using Customer Lifetime Value 25

2.5.3 Data Mining Using RFM 25

2.5.4 Data Mining and Market Segmentation 26

2.5.5 Data Mining and Clustering 27

2.5.6 Clustering with K-Means Algorithm 27

2.5.7 Clustering with EM (Expectation Maximization) Algorithm 28

2.6 Related Case Studies 28
2.7 Analysis Gap 33
2.8 Summary 34

CHAPTER 3 -RESEARCH METHODOLOGY

3.1 Introduction 35
3.2 Phase 1: Generic Data Warehouse Unification and Attribute Identification 38
3.2.1 Database Engine Schema 39
3.2.2 Feature selection based on attributes identified via interviews 40
3.2.3 Generic POS Data Warehouse Implementation 41
3.3 Phase 2: Data Pre-processing 44
3.3.1 Data Cleansing 44
3.3.2 Proposed Data transformation 45
3.4 Phase 3: Market Segmentation via the Classification of Customer Purchase Behavior 51
3.4.1 CLTV model on RFM PQ Dataset 51
3.4.2 Modified Regression on C dataset 52
3.5 Phase 4: Market Segmentation using Clustering Methods 55
3.5.1 Clustering with K-means Algorithm 56
3.5.2 Expectation- Maximization (EM) Clustering 59
3.6 Phase 5: Accuracy and Effectiveness Determination 62
3.6.1 Inter-cluster Distance 63
3.6.2 Customer Retention and Churn Rate 63

3.7 Implementation 64

3.8 Conclusion 65

CHAPTER 4 -RESULTS AND DISCUSSION

4.1 Introduction.. 66

4.2 Classification on RFM and RFMPQ Dataset 67

4.3 Control Experiment: market segmentation on RFM Dataset 70

4.3.1 Demographic Data Clustering.. 70

4.3.2 Cluster Quality Assessment for RFM Dataset 74

4.4 Market Segmentation on RFMPQ Dataset.. 76

4.4.1 RFMPQ Results Based on Best Segment 76

4.4.2 RFMPQ Results Based on the Spender Segment.................................... 79

4.4.3 RFMPQ Results on Frequent Segment 82

4.4.4 RFMPQ Results on Uncertain Segment.. 85

4.5 RFMPQ Dataset Summary.. 88

4.6 Comparison Between RFMPQ and Traditional RFM... 90

4.7 Market Segmentation On RFM C Dataset 91

4.8 RFMC Dataset Summary 95

4.9 Customer Lifetime Value based on Retention and Churn Rate 96

4.9.1 Historical CLTV 97

4.9.2 Predictive CLTV Matrix 97

4.9.3 CLTV Ranking Matrix.. 99

4.10 Suggested Marketing Strategies... 101

4.10.1 Best Segment.. 102
4.10.2 Spender Segment.. 104
4.10.3 Frequent Segment 104
4.10.4 Uncertain Segment... 105
4.11 Summary 108

CHAPTER 5 -CONCLUSION

5.1 Conclusion 109
5.2 Contribution 112
5.3 Recommendation and Future Research.. 114

REFERENCES 115
APPENDICES 123
APPENDDIX A 124
APPENDDIX B 131
APPENDDIX C 145
APPENDDIX D 149

LIST OF TABLES

		Page
Table 2.1	Illustrates the difference between Quantitative and Qualitative methods	10
Table 2.2	Overview of the literature studied in this research	31
Table 4.1	Average of four clusters generated by k-means algorithm (RFM dataset)	70
Table 4.2	Average of four clusters generated by EM algorithm (RFM dataset)	71
Table 4.3	Cluster Quality Assessment uses K-means algorithm	72
Table 4.4	Cluster Quality Assessment using EM Expectation-Maximization algorithm	73
Table 4.5	Average of four clusters generated by the K-means algorithm (RFMPQ dataset Best segment)	74
Table 4.6	Average of four clusters generated by the EM algorithm (RFMPQ dataset Best Segment)	75
Table 4.7	Cluster Quality Assessment for the Best Segment using K-means algorithm	76
Table 4.8	Cluster Quality Assessment for the Best Segment using EM algorithm	76
Table 4.9	Average of four clusters generated by the K-means algorithm (RFMPQ dataset Spender Segment)	77
Table 4.10	Average of four clusters generated by EM algorithm (RFMPQ dataset Spender Segment)	78

Table 4.11	Cluster Quality Assessment for the Spender Segment using K-means Algorithm	79
Table 4.12	Cluster Quality Assessment for the Spender Segment using K-means Algorithm	79
Table 4.13	Average of four clusters generated by the K-means algorithm (RFMPQ dataset Frequent Segment)	80
Table 4.14	Average of four clusters generated by EM algorithm (RFMPQ dataset Frequent Segment)	81
Table 4.15	Cluster Quality Assessment for the Frequent Segment using K-means Algorithm	82
Table 4.16	Cluster Quality Assessment for the Frequent Segment using EM Algorithm	82
Table 4.17	Average of four clusters generated by the K-means algorithm (RFMPQ dataset Uncertain Segment)	83
Table 4.18	Average of four clusters generated by the EM algorithm (RFMPQ dataset Uncertain Segment)	84
Table 4.19	Cluster Quality Assessment for the Uncertain Segment using K-means Algorithm	85
Table 4.20	Cluster Quality Assessment for the Uncertain Segment using EM Algorithm	85
Table 4.21	Comparison Between RFM (PQ) and Traditional RFM	86
Table 4.22	Average of four clusters generated by the EM algorithm (RFMC dataset Best Segment)	90
Table 4.23	Average of four clusters generated by the EM algorithm (RFMC dataset Spender Segment)	91
Table 4.24	Average of four clusters generated by the EM algorithm (RFMPQ dataset Frequent Segment)	91

Table 4.25	Average of four clusters generated by the K-means algorithm (RFMC dataset Uncertain Segment)	92
Table 4.26	Comparison Between RFM (PQ) and Traditional RFM	93
Table 4.27	Experimental Results on Customer Lifetime value (CLTV) using Ranking Matrix	98
Table 4.28	Suggested Marketing Strategies (Best segment)	101
Table 4.29	Suggested Marketing Strategies (Spender segment)	102
Table 4.30	Suggested Marketing Strategies (Frequent segment)	103
Table 4.31	Suggested Marketing Strategies (Uncertain segment)	104
Table 4.32	Summarized recommended strategies for the four customer segments	105

LIST OF FIGURES

		Page
Figure 2.1	CLTV formula	12
Figure 2.2	Retention rate formula	13
Figure 2.3	Customer Value Matrix	16
Figure 2.4	RFM distribution:125 RFM values and the number of customers	19
Figure 2.5	Knowledge Discovery in Database Process	
Figure 2.6	IPO(Input, Process, Output) diagram of the proposed method	23
Figure 3.1	Methodologies for the Hybrid of Classification and Clustering for market segmentation	26
Figure 3.2	Original POS Database	26
Figure 3.3	Data transfer from the production POS database to the production Generic POS database	39
Figure 3.4	Shows the final design of BUSUNI Schema and the needed objects to develop a Generic POS Data warehouse	41
Figure 3.5	Shows the Star-Schema with the relevant facts and dimension objects	42
Figure 3.6	Illustrates sample of the architecture design of the Star-Schema POS_STAR	42
Figure 3.7	Computing the Purchase slope for a sample customer purchase	43
Figure 3.8	Cluster Quality Assessment	52
Figure 4.1	Effective customer pyramid classification	61
Figure 4.2	Illustrate Historical CLTV	67

Figure 4.3	Illustrates Predictive CLTV of customers over three years' period	95

LIST OF ABBREVIATIONS

BI	Business Intelligence
MS	Market Segmentation
RDBMS	Relational Database Management System
ETL	Extract, Transform and Load
DM	Data Mining
DW	Data warehouse
SP	Stored Procedures
DSS	Decision support system
OLAP	Online analytical processing engine
MOLAP	Data Mining and OLAP
OLTP	Online transaction processing systems
ANN	Artificial Neural Networks
MBA	Market Basket Analysis
SMR	Small- and Medium-Sized Retailer

RFM	Recency, Frequency and Monetary
PQ	Purchase Power of Monetary and Product
C	Purchase Change Rate
CLTV	Customer Lifetime Value
MPS	Most Profitable Segment

SEGMENTASI PASARAN MENGGUNAKAN MODEL RFM (KEBARUAN, KEKERAPAN, KEWANGAN)

ABSTRAK

Kepentingan strategi pemasaran tepat sasaran merupakan prinsip yang mengubah fokus peruncit dari berorientasikan produk kepada pengutamaan pelanggan. Ia telah menarik minat industri dan akademik. Pelanggan adalah berbeza dalam pelbagai cara termasuk dalam pilihan membeli. Pendekatan yang digunakan secara meluas untuk mendapat gambaran tentang kepelbagaian tabiat membeli pelanggan dan keuntungan dinamakan segmentasi pasaran. Ia merujuk kepada pembahagian pasaran massa ke pasaran homogen yang lebih kecil berdasarkan persamaan pembelian dan kepelbagaian pelanggan. Model segmentasi pasaran konvensional sering kekurangan bukti empirikal terhadap pengiraan mereka yang diambil berdasarkan tempoh masa tertentu dengan mengabaikan hakikat bahawa tingkah laku pelanggan mungkin berubah dari masa ke masa. Ini menyebabkan peruncit terpaksa menggunakan sumber yang terhad dan berharga untuk menawarkan perkhidmatan kepada pelanggan walaupun ia tidak menguntungkan. Dalam usaha untuk memberikan pandangan menyeluruh ciri-ciri khusus pelanggan dan tingkah laku pembelian, kajian ini melihat kepada integrasi dua model dinamik, iaitu model Nilai Pelanggan Sepanjang Hayat (CLTV) dan model Kebaharuan, Kekerapan, Kewangan (RFM) untuk menyelidiki segmentasi pasaran untuk saiz peruncit sederhana di Negara Kuwait. Tiga variasi analisa RFM (iaitu pembolehubah P, Q dan C) yang baru turut dicadangkan dimana-mana ketiga-tiganya mempunyai kelebihan berbanding model RFM tradisional. Penyelidikan ini menggunakan ciri yang penting bagi integrasi CTLV dan RFM melalui kaedah transformasi data untuk mengubah dan memproses sumber data mentah tempat jualan (POS) dan disatukan kepada gudang data berasaskan skema bintang generik. Penyelidikan ini menggunakan algoritma K-cara, algoritma pengelompokan Penjelasan Maksimum (EM) dan algoritma regresi yang diubahsuai untuk analisis tabiat membeli pelanggan. Ketiga-tiga algoritma mempunyai konsep yang berbeza untuk pengelasan data dan segmentasi. Menggunakan penilaian kualiti kelompok, kajian ini menyimpulkan bahawa algoritma EM mengatasi algoritma K-cara. Oleh itu, strategi pemasaran dicadangkan selaras dengan keputusan yang dihasilkan oleh algoritma pengelompokan EM. Strategi pemasaran telah diimplementasikan kepada pelanggan peruncit di mana peruncit menyaksikan peningkatan pada kadar pertumbuhan jualan sehingga 6%.

MARKET SEGMENTATION USING ENHANCED RFM (RECENCY, FREQUENCY, MONETARY) MODEL

ABSTRACT

The importance of targeted marketing strategy, a principle method for transforming retailers from being product-oriented to customer centric, has attracted interest from both industry and academia. It is well known fact that consumers differ in various ways, and have contrasting buying preferences. A widely used approach for gaining insight into the heterogeneity of customer buying behavior and profitability is market segmentation. It refers to the division of a mass market into smaller homogeneous markets based on purchase similarity and the diversity of customers. Conventional market segmentation models are often lack the empirical evidence to their calculations and derived based on a specific time frame, which thereby often ignore the fact that customers' behavior may evolve over time, therefore retailers often consume limited and valuable resources attempting to service unprofitable customers. In order to provide a holistic view of customers' specific characteristics and purchasing behavior, this research looks into the integration of two dynamics models, which are the Customer Lifetime Value (CLTV) model and Recency, Frequency, Monetary (RFM) model that are being investigated for market segmentation for a medium size retailer in the State of Kuwait. Also, three new RFM variation analysis methods (i.e. P, Q, C) are proposed which have superior advantages with respect to the traditional RFM model. This research applies a critical feature for the CTLV and RFM integration, using data transformation method to transform and processes the raw Point-of-Sales (POS) data into consolidated generic Star-Schema data warehouse. This research applies K-means, Expectation Maximization (EM) clustering algorithms and modified regression algorithm to the analysis of customer buying behavior. These three algorithms bear different philosophy for data classification and segmentation. Using cluster quality assessment, this research concludes that the EM algorithm outperformed k-means algorithm. Therefore, marketing strategies are suggested in accordance with the results generated by EM clustering algorithm. The marketing strategies were implemented in to the retailer customers where they witnessed in their sales growth rate up to 6%.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The US Small Business Administration has traditionally defined the small to medium size retail industry (SMR) businesses as fewer than 500 employees (SBA, 2001). According to the European Commission's document, the SMR industry forms the backbone of the economy employing 50% of the private work force (Gal, 2010). Based on the historical facts, the SMR industry have had the privilege of developing close and mutually beneficial relationships with their customers, using mass marketing strategies which are mainly based on marketing experts and sales manager's opinions of the market (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Goyat, 2011).

These relationships were possible because consumer's buying behavior did not change much, and the price was less of an issue due to less competition (Cabena, P et al., 1998). However the recent economic and social changes have transformed the retail industry, particularly the relationship between the retailer and customers has changed significantly. As a result of this, the SMR industry has been challenged in recent years to be more strategic in their planning and understand their customers as well as their competitors. Therefore, shifting from mass marketing strategy which is mainly focuses on selling products and services without searching for detailed knowledge concerning the customers who bought the products and services is a must. Understanding customers behavioral and establishing loyal relationships with customers has become a central concern and main strategic goal for all retailers (Dash & Mishra, 2010). Hossein (2008) noted that the retail industry is highly competitive and products are overwhelming, consumers face with a variety of products, therefore customer's demand tends to be higher and more complex. To follow this trend, modern marketing is moving from mass-marketing (products-focus) to target-marketing (customer-focus).

1.2 Research Motivation

It is undeniable that the financial crisis has been quietly taking place in the retailer industry, leading to changes in consumer's behavior. Nie, Zhao & Yu, (2010) noted that valuable consumers are dynamic and the relationship evolves over the consumer's life-cycle. Thus it is essential to understand this relationship to device an appropriate marketing strategy with promotional campaigns according to the taste of the individuals of particular market segment.

Kruger (2011) in his book top market strategy, noted that applying the famous 80/20 marketing Rule, where 20% of customers generates 80% of the profit. The next question to ask is: how well does the SMR industry know their customers to predict this profitable 20% segment?

The retail industry is being the subject of many researches and studies about their market segmentation, and modeling of their customers' purchase behavior. However, most of the studies classified the current segmentations as one of the two study. Brand purchase behavior investigates the customer's purchase of one specific brand or product, where, store purchase behavior, investigate customer’s behavior based on all ranges of the products purchase in one store (Wedel & Kamakura, 2012). Nonetheless, there have been no studies considering the customer’s purchase behavior in interaction among different products based on a variety of market segmentation demographic variables and characteristics in general and targeting the SMR industry in particular.

The research motivation is to design and develop market segmentation data mining model to extract hidden knowledge from large POS data warehouse and address customer segments with specific needs and propose a customer retention strategy to help the SMR industry grow through an effective target marketing strategy, which will eventually result in producing commercial advantage and tangible on ROI.

1.3 Research Problem

Customer purchase behavior and customer churn are the focal concerns of all retailers. Among all retailers which suffer from this issue, the small to medium size retail industry (SMR) in the State of Kuwait. The (SMR) is faced with an increasing serious competition from large retailers and mass merchandisers, mainly because of their variety of different products and low prices that appeal more to consumers as a direct result from the adaptation of market segmentation strategies giving them the ability to target profitable customers.

Market segmentation has been the subject of many studies. Unfortunately, most of these studies do not address individual consumers based on their purchase behaviors, but only look in the rear-view of consumer's historical data on the assumptions of what makes consumers similar to one another. This technique hides critical facts about individual consumers.

Sohrabi & Khanlari (2007) proposed customer lifetime value (CLV) measurement based on RFM model. The authors used K-Means clustering approach to determine the customer lifetime value and segment them based on RFM measures and then proposed customer retention strategy. However, their techniques suffer from number of weaknesses, firstly, the research only utilizing RFM model to measure customer lifetime value and traditional RFM alone does not deliver the level of accuracy that marketers require. Marcus (1998) noted that the RFM in spite of its simple conceptual framework is too complex and time-consuming for small retailers. Moreover, RFM model only describe what a consumer has done in the past and cannot accurately predict future consumer's behavior (Fader, Hardie & Lee, 2005). Secondly and most importantly, RFM model looks at consumers at a particular point in time and does not take into consideration how the consumer has behaved and in what lifecycle stage the consumer is currently found. Therefore, accurate consumer modeling is very weak unless the consumer's behavior is analyzed over time.

Thirdly, their approach uses only K-means algorithm to segment consumers, makes the research lacks of comparison of algorithms on real datasets to measure the cluster quality. Fourthly, no data transformation is applied prior to applying K-means algorithm to remove noise and outliers in the data. Studies have shown that k-means with high sensitivity clustering data that contains noise and outliers where almost all data points are as far away from the centroid (Sisodia, Singh, Sisodia & Saxena, 2012). Therefore, designing and developing market segmentation data mining model which predicts customer's behavior patterns and recognize customers which tend to churn is vital.

In this research we foresee an enhanced market segmentation modeling methods which are far more advanced and effective than conventional RFM method. By integrating CLTV model with new proposed RFM variants (PQ) (C) into an integrated closed-loop model, that will offer SMR industry highly accurate customer behavior analysis. the (PQ) variants represent the average purchase power per customer and per product. The (C) represents the change of consumer purchase behavior or trend using change rate.

1.4 Research Questions

Based on the problem discussed above. The research questions are as follows:

1. Which RFM factor is the most important in market segmentation?
2. How can a traditional RFM model be enhanced to cater new variables needed for market segmentation?
3. Which clustering algorithms break down the SMR market into more meaningful segments?

1.5 Research Objectives

1. To propose new variants to RFM model to cater the consumer's purchase power and purchase change rate variables and use the modified model for data transformation.
2. To propose new market segmentation approach using classification and regression based on the modified RFM followed by clustering on demographic data.
3. To implement and evaluate Target-marketing strategies based on the output produced by the new market segmentation approach.

1.6 Research Scope

This research focuses on mining changes in customer purchasing behavior based on customer purchasing transactions stored in POS database. The POS database is collected from medium-Size department store retailer based in the State of Kuwait. This research focuses on developing market segmentation data mining model by incorporating CLTV with RFM scoring model.

1.7 Organization of the thesis

This Research consists of five chapters. The first chapter is an introduction that gives a brief background about subject of study. Chapter 2 is a literature review of different methods and models of market segmentation, RFM, CLTV and data mining. Chapter 3 is about the research methodology and introducing our new proposed methods for market segmentation. Chapter 4 is about the results of the analysis and applying the proposed methods and the obtained customer segments will be explored more in detail, also the related strategies for each segment will be discussed. Finally, chapter 5 is the conclusions, contribution and further research recommendation.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In the 1950's, the theory of market segmentation had emerged as a formal component of contemporary marketing practice (Wedel & Kamakura, 2012). Jensen (1996) stated that market segmentation is the way for smaller companies to succeed in big markets. Wind (1978) stated that, if the retailer is to enjoy any level of marketing success, this is through the adeptness and ability to match its capabilities to the needs and requirements of the marketplace.

Central to this matching process is the segmentation of the market. Retailers need to sell ideas, hopes and courtesy before they sell the product. Only then, retailers will get repeated loyal customers (Mathur, 2010). Kimball and Ross (2011) noted that knowing more about customer segments are the bases for successful customized marketing plans specifically to cater the needs of a particular group.

2.2 Market Segmentation

Dipanjan, Satish and Goutam (2011) defined market segmentation as the process to divide customers into similar or homogeneous groups sharing with one or more characteristics such as shopping habits, lifestyle, taste, and food preferences. These characteristics are relevant to marketing and sales, such as demographics, age, location, nationality, gender, interests and spending habits (McCarty & Hastak, 2007).

Market segmentation methods have great importance in empowering retailers to precisely reach a consumer with specific needs by dividing the market into similar and identifiable segments to help marketers to bring together individuals with similar choices, needs and interests on a common platform (Kashwan, K. R. 2013; Kolyshkina, I. et al., 2010). Companies begin to look for customer service as a market differentiator and many companies also started to segment customers for service delivery (Milgramm, 2011).

2.2.1 Market Segmentation Bases

(A) Demographic segmentation

According to Jobber and Ellis-Chadwick (2012) demographic segmentation is the most used variable when segmenting a market. It gives a precise customer purchase profile and it focuses on measurable criteria of consumers and their households. Furthermore, this segment is primarily descriptive in terms of gender, race, age, income, lifestyle and family status (Cleveland, Papadopoulos & Laroche, 2011).

(B) Geographic Segmentation

Geographical segmentation classifies customers according to their location and when a broad segment of customers has different preferences based on their location.

This strategy is often used as a starting point in Market Segmentation and mainly used by retail or origination in the service sector, with limited resources. However, geographic segmentation is not used by itself, rather in conjunction with other Market Segmentation factors (Goyat, 2011).

(C) Behavioral Segmentation

Behavioral segmentation divides customers based on their attitude toward products. Many marketers believe that the behavioral variables such as occasions, benefits, user status, usage rate, buyer-readiness stage, loyalty status and attitude are the best-starting points for constructing market'ssegments (Cleveland, Papadopoulos & Laroche, 2011).

2.2.2 Market Segmentation Data Sources

There are a number of methods for collecting, analyzing and processing data. In general, these methods are divided between qualitative and quantitative approaches (Bernard, 2011). In most cases, a combination of qualitative and quantitative methods is the best practice (Elby, 2015).

Table 2.1: Illustrates the difference between Quantitative and Qualitative methods

Methods	**Characteristics**
Quantitative	a. Use data in the form of numbers. b. Require variables to be predetermined. c. Data collected through methods such as questionnaires (closed-responses), record keeping and population surveys.

Qualitative	a. Use data in the form of words. b. Do not require pre-determined variables and can be used for open-ended or exploratory questions. c. Data usually collected through methods such as: observation, interviews, questionnaires, focus groups, case studies and document analysis.

2.3 Customer Value Analysis

The value of a customer changes throughout their established relationship with an organization (Safari, 2015). Customers vary extensively in a range of attributes, including products preferences, sensitivity to price, cost-to-serve, retention rates and responses to marketing strategies. As a result of these and other factors, customers differ widely in the value they represent to an organization (McDougall et al., 1997).

To build and improve upon traditional segmentation, retailers have been trying to identify profitability segments of customers, that differ in current and future profitability to the retailer. The most common analysis of customer profitability is the customer pyramid (Zeithaml, Rust & Lemon, 2001).

Often retailers spend the large sum of their marketing budget on new customers and non-profitable customers. On the contrary, the Customer pyramid approach goes beyond traditional market segmentation because it tracks revenues and costs for segment of customers, therefor, capturing customers financial true worth to retailer (Curry, A., & Curry, J. 2002). Thurs, feasible marketing strategies can therefore be achieved like, preserving the Top 20% customers and developing the lowest 80% customers to foster lower and high potential segment of customers towards the top 20%.

2.3.1 Customer *Lifetime* Value (CLTV) Model

The literature has generally defined CLTV as a quantitative measurement of the amount of sales the customer will spend with a particular retailer over their lifetime (Dwyer, 1997). Safari (2015) defined CLTV as the current value of all future profits obtained from a customer over his or her lifetime relationship with the retailer.

Abe (2009) noted that the past few years have seen an explosion of researches into customer lifetime value. This has followed by an increased focus on customer relationship management where retailers consider their interactions with customers over the entire duration of the customer lifetime to evaluate strategies and improve sales and profitability. This is where CLTV proves its effectiveness by using a metric to evaluate the actions of the firm (Borle, Singh & Jain, 2008).

2.3.2 Components of CLTV

CLTV has three main components in order to calculate CLTV. These components are customer acquisition, customer expansion and customer retention (Gupta et al., 2006). However, it is very important to consider COGS (Cost of Goods Sold) and acquisition cost to square off the real CLTV.

$$\boldsymbol{CLTV = (Average\,Value\,of\,Sales) \times (Number\,of\,Repeat\,Transactions) \times (Average\,Retention\,Time)}$$

The basic model to calculate CLTV is shown in Figure 2.1

$$CLV = \sum_{t=1}^{n} (r)^t \frac{P_t}{(1+d)^t}$$

Figure 2.1: Customer Lifetime Value (CLTV) formula

The CLTV formula above is more of a proxy for an average customer stick around for X period of time and pay Y during this period of time. Where, ***t*** represents a period of time, where (t=1) represents the first year, and (t=2) represents the second year. The **n** represents the total number of periods the customer will stay with the retailer before he or she finally churns. **The r represents** the month over month retention rate/possibility. $\mathbf{P_t}$ is the profit the customer will contribute or generate to the Retailer in the Period **t, finally, d** is the churn rate.

$$Retention\ rate = \left(\frac{(CE - CN)}{CS}\right) \times 100$$

Figure 2.2: Retention rate formula

Customer's loyalty can be calculated using the above Retention Rate formula. where CE represents the number of customers in the end of the chosen period, where CN is how many new customers have been acquired in the chosen period and CS stands for the number of customers at the start of the chosen period.

2.3.3 CLTV Retention and Churn Rates

Churn rate and retention rate are primary components of the future CLTV. Where CLTV is an estimation of the average profit a customer is expected to generate before he or she churn (Borle, Singh & Jain, 2008).

According to Lejeune (2001) the concept of retention and churn is often correlated with the industry life-cycle. When the industry is in the growth phase of its life-cycle, sales increase exponentially; Therefore, the number of new customers exceeds the number of churners, but for products in the maturity stage of their life-cycle, companies put the focus on the churn rate reduction.

According to Fader and Hardie (2010) previous studies have examined the concept of customer churn from different points of view. For customer retention, it continues to be an important topic to marketing researchers and analysts. Much of the recent research on customer retention has linked retention rates and churn probabilities to forecasts of customer lifetime value balancing, the allocation of resources among retention and other marketing efforts.

Borle, Singh and Jain (2008) noted that CLTV formula assumes that retention and churn happen linearly over the lifetime of a customer with the retailer.

2.3.4 CLTV Benefits and Limitation

The importance of the CLTV originates from the fact it outlines the net amount of customer contributes over their life with the retailer (Safari, 2015). Also CLTV treats a customer, who was previously perceived as one of many, is now taken as a unique entity that requires specialized and individual service, that makes the customer, and each relationship has its value described by the CLTV model (Rožek & Karlíček, 2014).

Furthermore, the CLTV assumes that customers are not equally profitable and resources have to be allocated accordingly to their expected lifetime value in order to maximize profitability (Toporek, 2016).

According to Zeithaml, Rust and Lemon (2001) and Drew et al. (2001), CLTV helps limiting the acquisition spending, deciding the promotion policy, better valuation of lists in direct mailing and the nature of data to collect. But despite the above benefits of the CLTV, some researches have indicated limitations in CLTV. The most apparent fact is the uncertainty of CLTV calculation which determines customer's profitability to the retailer, therefore this uncertainty makes the CLTV model hard to implement without any further investments (Kotler & Keller, 2011).

According to Abdolvand, Albadvi and Koosha (2014) CLTV has limitation to accurately predict customer purchase behavior due to the lack of empirical implementations and the lack of integration between customer data and marketing efforts, those limitations are the main obstacle in utilizing its potential in improving business strategies.

2.3.5 Market Segmentation based on CLTV Model

Marcus (1998) introduced CLTV Matrix and was initially developed from the RFM method for small-business retailers. The CLTV represented a greater improvement over traditional RFM analysis. The frequency of purchase and average purchase amount are used for the segmenting customers. The easiness to understand quadrant identifiers was its main advantage.

Marcus (1998) approach, is the calculation of the average values for the number of purchases **(F)** and the average amount spent per customer. Three data values must be collected in CLTV Matrix. The second step is segmentation process where each customer is allocated to one of the four resulting categories (**quadrants**) based on whether customers are above or below the axis averages

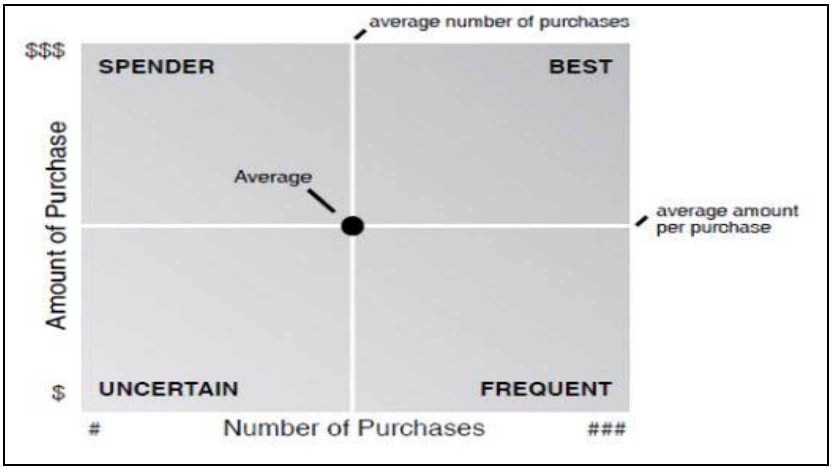

Figure 2.3: Customer Value Matrix. (Source: Marcus,1998)

2.4 Recency, Frequency and Monetary RFM Model

The literature has traditionally defined RFM analysis as the standard approach to assess and understand customer lifetime value and it is quite popular, especially in the retail industry. RFM involves the calculation and the examination of three variables – Recency, Frequency, and Monetary (RFM) (Tsiptsis & Chorianopoulos, 2011).

RFM analysis is based on the famous marketing axiom that "80% of your business comes from 20% of your customers." With these RFM's arranged, groups of consumers can be classified consistent with certain proportion. Once the customers are allocated, then RFM behavior scores can be grouped into segments and their consequent effectiveness is analyzed. The mentioned above forms the basis for future consumer contact frequency decisions (Kruger, 2011; Miglautsch, 2002).

RFM is considered as one of the most important models used for market segmentation that distinguish important customers and identify customer's purchase behavior by three dimensions which are customer's consumption interval, frequency and spent money. **R** symbolizes recency referring to the interval between the time when the latest consuming behavior happens and present. How much the interval is shorter, the **R** is bigger. **F** symbolizes frequency referring to the frequency of consuming behavior in a period of time. **M** symbolizes monetary referring to consumption money amount, to a period of time (Birant, 2011).

2.4.1 Benefits of using RFM Model

RFM model benefits retailers in a number of ways of segmenting the customers according to their attractiveness thus helping decision-makers to decide which customers to give particular offers based on the likelihood and find ways to increase their spending. RFM also helps with targeting lost customers or retain customers by offering them incentives and customized promotion (Wei, Lin & Wu, 2010; Birant, 2011; Miglautsch, 2002). RFM analysis model can be easily understood by decision makers and can increase a company’s profits in the short term (McCarty & Hastak, 2007; Wang, 2010; Khajvand & Tarokh, 2011).

2.4.2 RFM Analysis Scoring Techniques

The idea of RFM scoring is to plan future customer behavior and driving better market segmentation decisions. It is critical to interpreting customer behavior into numbers which can be used through time (Birant, 2011).

It’s very often direct marketers will use static customer selections when initially building their segmentation system, and they define some factors with some thresholds, if these limits keep fixed, the results will be poorer and poorer over time (Miglautsch, 2002).

2.4.3 RFM Scoring Based on Quintiles

Quintiles scoring is the most common scoring method used to arrange customers in ascending or descending order (Best to Worst). Customers are divided into five equal groups or Quintiles. The best group receives a score of 5; the worst receives a score of 1 (Dash & Mishra, 2010).

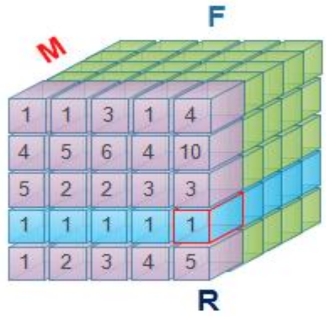

Figure 2.4: RFM distribution: 125 RFM values: Source: (Birant, 2011).

2.4.4 RFM Weighted Scores

Birant (2011) and Dan Ross (2005) stated that RFM values can be added together. Scoring is not explicitly discussed, but they offer a formula for creating a single RFM value. their method includes adding average order and Frequency per year.

An alternative would be adding together the RFM scores discussed above. The best customers would have a composite score of 15 (5+5+5), and the worst customers would have a minimum score of 3 (1+1+1). The RFM score is the weighted average of its individual components and is calculated as shown in Figure 2.4:

Combining R, F, and M Components to Derive a Continuous RFM Score

RFM score = (recency score × recency weight)
+ (frequency score × frequency, weight)
+ (monetary score × monetary weight).

Finally, these scores can be re-scaled to the 0–1 range according to the following formula:

$$\text{Rescaled RFM score} = \frac{\text{RFM score} - \text{minimum RFM score}}{\text{Maximum RFM score} - \text{minimum RFM score.}}$$

Equation 2.1: RFM Scoring Formula source: (Tsiptsis & Chorianopoulos, 2011).

2.4.5 RFM with CLTV

Several studies employ RFM model to calculate CLV (Liu & Shih, 2005 ; Sohrabi & Khanlari, 2007). Liu and Shih (2005) developed a novel product recommendation methodology that combined group decision-making and data mining techniques by utilizing AHP, clustering and association rule mining techniques.

They have proposed a technique to apply RFM to evaluate CLTV. Four methods were compared in their study, namely weighted-RFM method, non-weighted RFM method, the non-clustering method and the typical collaborative filtering (CF) method.

Sohrabi and Khanlari (2007) have used K-means clustering technique to develop a CLTV model by determining customers' CLTV and segmentation by taking into account the RFM measures.

2.4.6 Market Segmentation based on RFM Method

A large number of studies have considered RFM method for market segmentation and highlighted the importance of RFM variables.

Aggelis and Christodoulakis (2005) studied the RFM scoring of active e-banking users, the paper used clustering techniques as one of the methods of data mining to organize observed examples into clusters based on a pyramid model.

2.4.7 RFM Analysis Model Limitation

Although the RFM model is a vital analysis tool for organizations to design and develop customized marketing strategies, it also has its own limitation (Miglautsch, 2002). First, RFM analysis aims to identify only valuable customers. Therefore, it only focuses on those scored best customers and provides less meaningful scoring on recency, frequency and monetary when most consumers (Wei, Lin & Wu, 2010). Second, RFM analysis has inability to prospect for new customers, as RFM analysis only focuses on the organization's current customers (McCarty & Hastak, 2007). Also, RFM analysis is not foreseen as a precise quantitative analysis model, but the importance of each RFM measure is different among other industries (Yeh et al., 2009).

Yeh et al. (2008) proposed a comprehensive methodology to select targets for direct marketing from a database by extending the RFM model to new proposed RFMTC model and adding two parameters, namely: time since first purchase and churn probability.

2.5 Data Mining

The development of modern information technology has generated massive amounts of data from various databases, data warehouses and other repository information, constructing those transaction data requires proper mechanisms to convert it into knowledge, using this knowledge retailer can make better business decision. (Sisodia, Singh, Sisodia & Saxena, 2012; Dawei, 2011).

Data Mining is seen as a powerful analytical tool for the retail industry (Tufféry, 2011). It is used to provide the analysis of information on product, sales, customer buying habits, data and identify naturally occurring clusters of behavior, which then form the basis of segments (Azevedo, 2014).

Data mining involves the inferring algorithms that explore and analyze large quantities of data, develop mathematical models and discover significant patterns (implicit or explicit) which are the essence of useful knowledge (Chen, Sain & Guo, 2012). Data mining main aim is to discover valuable patterns from a large collection of data for users, (Gunaseelan & Uma, 2012).

There are two primary data mining process goals which are verification, and discovery. Verification is verifying the user's hypothesis about the data while discovery is automation of finding unknown patterns shown in Figure 2.6.

In summary, data mining is the process of trying to make sense of data.

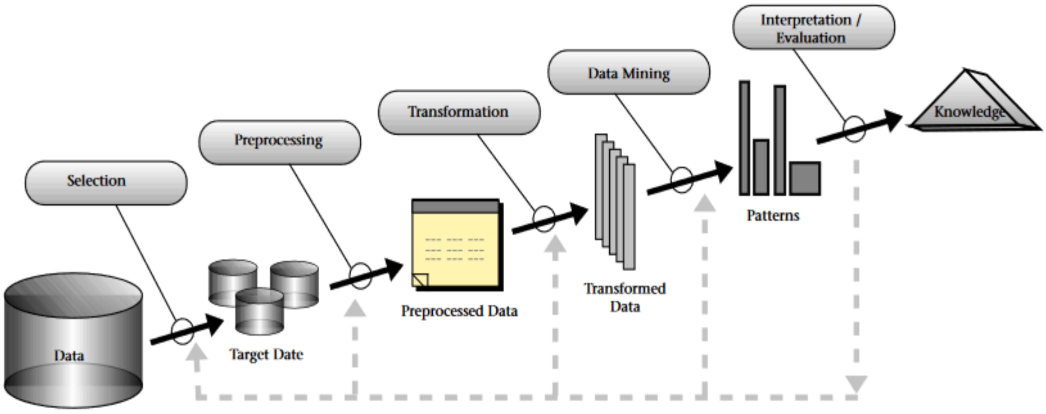

Figure 2.5: Knowledge Discovery in Database Processes Source: (Frawley, Piatetsky-Shapiro & Matheus, 1992)

There are five algorithm techniques that apply to data mining methods. We will briefly cover all 5 algorithms, but we will mainly focus on the relevant market segmentation algorithms which are clustering algorithms.

Outlier detection: Outlier or anomaly detection is used to detect fraud or risks within critical systems (Collins, 2013).

Association rules learning: The process of discovering interesting and unexpected rules from large data set is known as association rule mining (Ghosh, Biswas, Sarkar& Sarkar, 2010; Rajagopal, 2011). Apriori algorithm is the classic algorithm of association rules, which enumerate all of the frequent item sets (Yabing, 2013). The association aims to establish relationships between items, which exist together in a given record. The most common algorithms are K-means, Kohonen, 2 step and a priori, GRI.

Classification: Classification rule mining aims to discover a small set of regulations in the database that forms a perfect classifier (Dhanabhakyam & Punithavalli, 2011). The most common algorithm is CART and C5.0.

Regression: Regression is an empirical statistical method and is extensively used in business, the social, the biological sciences, behavioral sciences, climate prediction, and many other areas. Linear regression is one of the most common data mining technique for predicting the future value of a variable based on the linear relationship, (Zhao, 2011). The most common type of algorithm is neural networks, CART, Regression, and GLM.

2.5.1 Data Mining Methods for Discovering Patterns

The majority of previous studies have used various mathematical models to segment customers without considering the correlation between customer purchase behavior and POS data. (Lefait & Kechadi, 2010). The traditional mathematical models are difficult to predict the segmentation patterns, (Kashwan & Velu, 2013). This approach resulted in overlooking some useful information and far less use of the potential benefits of increased data-gathering capabilities.

Predictive, data mining technique to create a model to predict the future values based on the past and current data set values. The various Predictive data mining techniques are (Venkatadri, Sastry & Reddy,2012).

Descriptive, data mining technique to describe the general properties of the existing data (Zaiane, 1999). Descriptive data are the techniques to organize the data, based on