UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Masterthesis

Achieving Customer Loyalty from Email Campaigns by Using Data Mining Techniques

Julia Naber

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization International Marketing Strategy

SUPERVISOR :
Prof. dr. Koenraad VANHOOF

CO-SUPERVISOR :
dr. Gonzalo NAPOLES RUIZ

2016
2017

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Masterthesis

Achieving Customer Loyalty from Email Campaigns by Using Data Mining Techniques

Julia Naber

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization International Marketing Strategy

SUPERVISOR :
Prof. dr. Koenraad VANHOOF

CO-SUPERVISOR :
dr. Gonzalo NAPOLES RUIZ

Preface

I present this thesis, Achieving customer loyalty from email campaigns by using data mining techniques as a fulfillment for the requirements of the Master in International Marketing Strategy program at the University of Hasselt. This work was in collaboration with Viata Online Pharmacy, entrepreneurs Benoit Van Huffel and Gianni De Gaspari are the reason for the opportunity I had. I would like to thank them for providing me with the data and support I needed.

Most importantly, I would like to thank my supervisor Prof. Dr. Koen Van Hoof for his constant support, leadership and his encouraging words. I am forever grateful, for you provided me with the strength and confidence by being a role model. I would like to thank Dr. Gonzalo Napoles Ruiz for his guidance and care from beginning to end. You were both even more supportive from hundreds of miles away.

Fadia, Maher and Jawad. You are the rocks of this thesis. Thank you for being there for me for all these years, through all the ups and downs and never giving up.

After working for a banking software company for a couple of years in email marketing campaigns, this research was the science behind it. It only showed me the necessity of knowing how it works and why. The numbers proved why we need data mining for present and future businesses. I am glad I now have the chance to add value in decision making of email campaigns having known how it would affect the customers and inevitably, the company itself.

Julia Naber

Hasselt, Belgium

August 2017

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 About Viata

Viata is an online pharmacy and drug store based in Europe. On one side, the online shopping is gaining massive popularity among the consumers, and the same trend has also shifted towards e-pharmacy. With numerous developments and technological advancements in the e-commerce domain, the need for e-pharmacy is also evolving in Europe. Viata guarantees to offer 100% bona fide and genuine medical services and pharmaceutical products, with doorstep delivery. Apart from these, the company also offers healthcare devices, nutritional supplements, wellness products, etc.

The web address of the company is www.viata.be. The web portal supports English, French, Dutch and German languages. The website has a wide range of products, roughly about 20000 verities, with proper product details. Viata also provides chat support for the customers to assist them regarding recommendation and information of various products related queries and other clarifications. The overall business model of Viata is divided into 3 parts, marketing, web store and logistics (viata-shop.com, 2017).

This thesis will concentrate on the use of data mining techniques on clients' information, RFM investigation and discovering the customer satisfaction with the help of data classification.

1.2 Problem Statement

Predicting loyalty could be beneficial in focusing on the audience with the demographics that has the highest loyalty, making sure the company invests in them and changes the advertisements and price offers to fit to the taste of the targeted group. Moreover, predicting loyalty of groups could tell us why others are not loyal and presumably give an indication of how the company could use different marketing strategies to reach out to those other groups. Doing so could help in reallocating all budgets for email marketing campaigns correctly and unlocking hidden profitability from each targeted group. Brand loyalty is an important note too, this means that companies that want to decrease client churn must maintain their brand identity for all customers despite having different segments, however, predicting loyalty here could help in making the

correct small changes to the brand identity to fit most groups without losing their image and hence loyalty of existing customers. Also, in order to segment the customers as loyal or non-loyal, an organization has to do lots of analysis based on transaction history of consumers. RFM (Recency, Frequency and Monetary) analysis is crucial to find customers loyalty towards the company. This research intends to explore various algorithms that may help in classifying the customer's loyalty accurately, so that it could be used in future for target marketing.

1.3 Research Goals

The main goal of this research is to analyze customers' loyalty in VIATA Company.

Specific goal 1. To characterize the available data

The first in data-oriented research should be the proper characterization of available data in order to evaluate the algorithmic challenges to be faced in next steps.

Specific goal 2. To determine the correlation between variables.

The second step will be focused on determining the correlation between variables, which can help us to understand later the observed patterns.

Specific goal 3. To characterize the behavior of loyal and non-loyal customers

The third step will be focused on gathering the customers according to their behavior and next discovering association rules between the variables and the decision class.

Specific goal 5. To predict the behavior of new customers

The above steps allow having a clear picture of customers' loyalty, but it does not efficiently allow predicting whether a new customer will be loyal or not. The next step is concerned with building different prediction models.

Chapter 2 E-Commerce, Challenges and Solutions

2.1 E-Commerce and Statistics

E-commerce business is vast and enormous. In the recent couple of years, it has developed from a 90's fascination, into a strong challenger for retailers, which has the ability to put them out of business. According to Forbes report, the e-commerce industry will cross an average annual transaction of over $2 trillion in 2017 (MichaelLazar, 2017). The eMarketer's report of a similar kind has predicted that the e-commerce transaction will reach a humongous number of $4 by 2020, which represents around 14.6% of total customer spending on retail industry (eMarketer, 2016). As can be measured, a total of 50% growth can be seen annually, by making a comparison on these two reports. This further gives an idea of how monstrous the industry can become, all because of advancing technology (MichaelLazar, 2017).

Such e-commerce statistics should make us realize the truth. Statistics says that around 71% of the shoppers prefer online store to check latest pricings and best offers, which is generally referred as "showrooming" (checking the best price for your mobile phone at the best cost when in a physical store), and "webrooming" (comparing it with multiple web store for lowest price) (MichaelLazar, 2017).

Business Insider (Cooper Smith, 2014) has reported that over 50% of shoppers purchased online more than once in the last year. Upon tallying these numbers, it can be found that over 198 million shoppers in the USA have been benefitted by the technology to help them perform online shopping, last year. It is more than 200 million individual shoppers on the web, which constitutes around 66% of the whole U.S. populace.

According to reports of US Census Bureau, around 75% of the retail sales during the fourth fiscal quarter of 2015 was contributed by the e-commerce (US Census Bureau News, 2017). During that financial year, it contributed to an annual growth of 0.5%. Between 2015 and present, the achieved growth is over ten folds. It is anticipated to repeat the same trend between present and 2020.

2.1.1 E-Commerce in Europe

According to the Interactive Media in Retail Group (IMRG) and Capgemini report of January 2017, the B2C e-commerce in United Kingdom witnessed a sales growth of 16% in 2016. The sales number grew from £115 billion ($175.74 billion) in 2015 to £133 billion ($203.26 billion) in 2016 (Retail & Ecommerce, 2017).

The UK has grown a powerful e-commerce business with other European countries. Generally, the UK has a very strong ecommerce economy in relation to other countries in Europe. As per the report of Ecommerce Foundation (formerly known as Ecommerce Europe) of May 2016, over 30% of overall sales in 2014 are ecommerce. These numbers are twice in comparison to Germany. These digits escalated to 34.5% in 2015, which is still 20% higher than its nearest competitor, France (Retail & Ecommerce, 2017).

2.1.2 Great online potential for the European market

The Europe has over 821 million population, out of which 73.5% use the Internet (Miniwatts Marketing Group, 2017). Hence, there is a huge scope for development of e-commerce in Europe. In support of these, majority of these users are inclined towards online shopping. It is projected that over 66% of them are involved with online shopping in 2015. Some of the popular products are clothe, electronics, sport and leisure, and other services (Eurostat Information, 2016).

One of the predominant factors for improvement of the European retailing is the online retail. The growth rates of e-commerce sector in 2014 and 2015 were 18.4% and 18.6%, respectively. However, this number is slowly diminishing as predicted growth in 2016 is around 16.7% and in 2017, the growth would slow down to 15.7% (retailresearch.org, 2017).

The online retail of Europe is dominated by UK, France and Germany. In spite of having higher online share in their domestic markets, they jointly contribute to around 81.5% of total sales in Europe (Gemma, 2016).

Figure 1: Online Share of Retail Trade

Source: (Gemma, 2016)

In Europe alone, the internet usage and buying trends of people from different countries, vary significantly. The United Kingdom ranks first with over 81% of online buyers in 2015. It is trailed by Denmark with 79% customers, and with 78%, Luxemburg is in third place (Gemma, 2016).

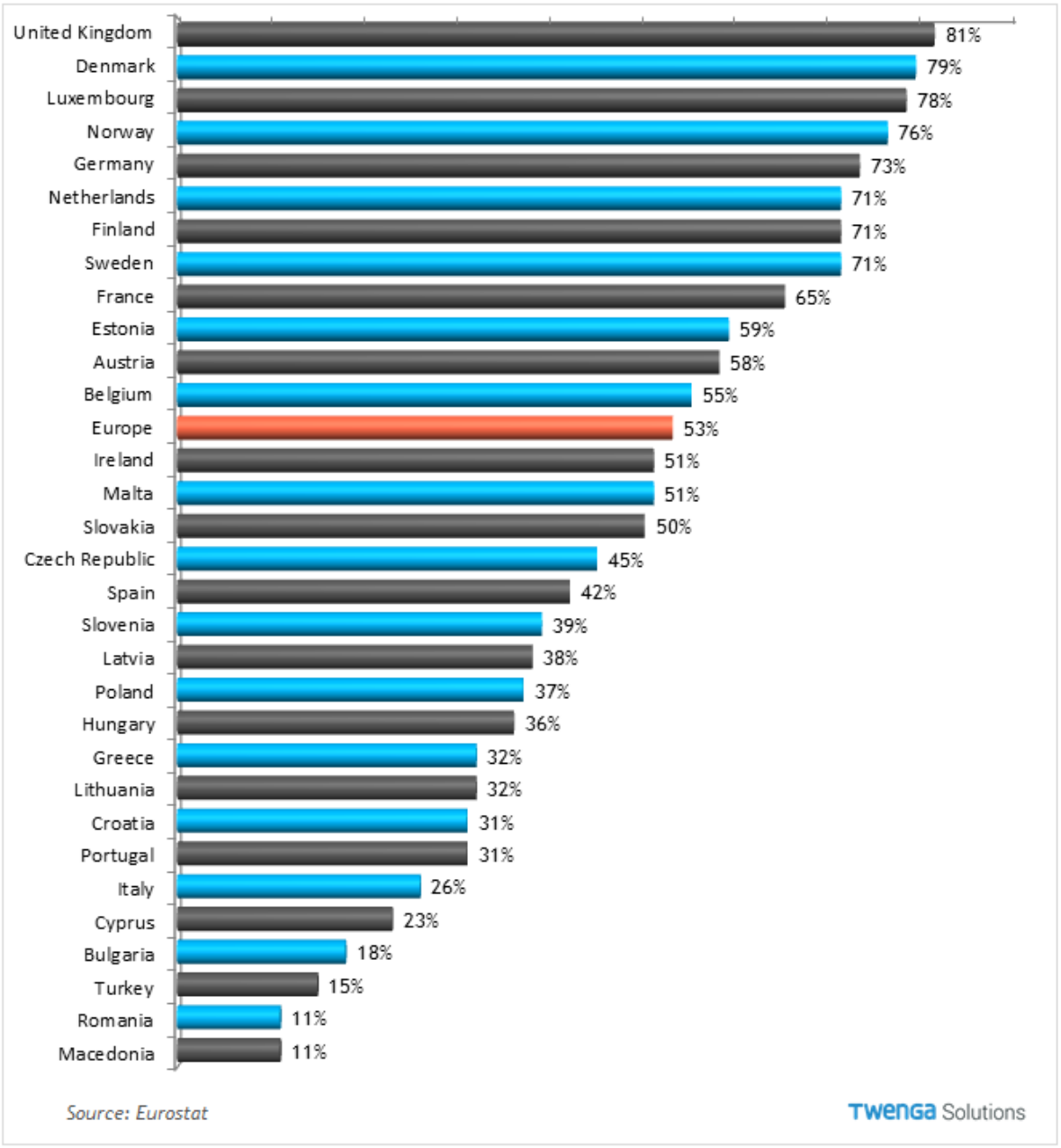

Figure 2: Proportion of individuals who purchased online in the last 12 months

Source: (Gemma, 2016)

2.1.3 E-Commerce in Belgium

Belgium is one of the federal parliamentary democracy in Europe. The capital of Belgium is Brussels. The overall population of Belgium is around 11.3 million. Dutch is the official language of Belgium. Apart from Dutch, French and German is also spoken. With a GDP of €409.8 Billion in 2015, the GDP per capita of Belgium is around €36,500. In the same year, the turnover of B2C e-commerce industry in Belgium grew by €8.2 billion, which accounts to 34.2% growth against last year. Belgium has over 8 million internet users who are over 15 years of age. Among those people, around 6.9 million people have purchased some products online in 2015. On an average, a typical Belgium citizen spends around €1,191 in a year (Ecommerce Foundation , 2016).

According to 2016 reports, the e-commerce industry of Belgium was estimated to have a value of 9.1 million. Compared to 2015, a customer would spend 10% higher in 2016. The overall share of e-commerce in country's retail business has raised from 14% in 2015 to 16% in 2016 (Ecommerce News, 2017).

This finding is supported by the recent BeCommerce Market Monitor. It is an association formed between PostNL, Google and Worldline, which majorly consists of Belgian e-commerce industry. In 2016, the total amount spent by 8.4 million Belgian consumers on online purchases was around 85 million. This suggest that an online shopper has made an average purchases of 10 in that year, with €9.34 per transaction (Ecommerce News, 2017).

In Belgium alone, the e-commerce is gaining more prominence and becoming vital component of the country's retail industry. The total share of online transaction in the country's retail industry, is around 14%, while that number has escalated to 16% in 2016. Among these online purchases, over 63% purchases were made on services, while only 8% was spent on products (Ecommerce News, 2017).

2.1.4 Online Pharmacy in Europe

Online pharmacy and mail ordering is a huge hit among the citizens of USA, which contributes to $1/4^{th}$ of total non-institutional healthcare market. But in Europe, the e-commerce in healthcare sector has commenced its operation from 1998. However, the development and success of this e-pharmacy in Europe is different in different regions (James Dudley, 2012).

A study was conducted by James Dudley Management on 17 different European countries which consists of 67% of European internet users. The findings revealed that around 14 countries among the 17 participants have allowed internet and mail purchases for non-prescription drugs and medicines under native law. In the published report namely, Mail Oder and Internet Pharmacy in Europe - the 2012 edition, it was described that only Germany, Denmark, Norway, Netherlands, Switzerland, Sweden and the United Kingdom have mandated that only prescription based medicines should be sold online (James Dudley, 2012).

In 2013, e-prescribing is introduced in Finland, after which the country was also included in this sector. In few European countries such as France, Austria and Italy, such online transactions are made illegal, but non-prescription medicines can be sold online under EU law (James Dudley, 2012).

2.2 Facts & Figures

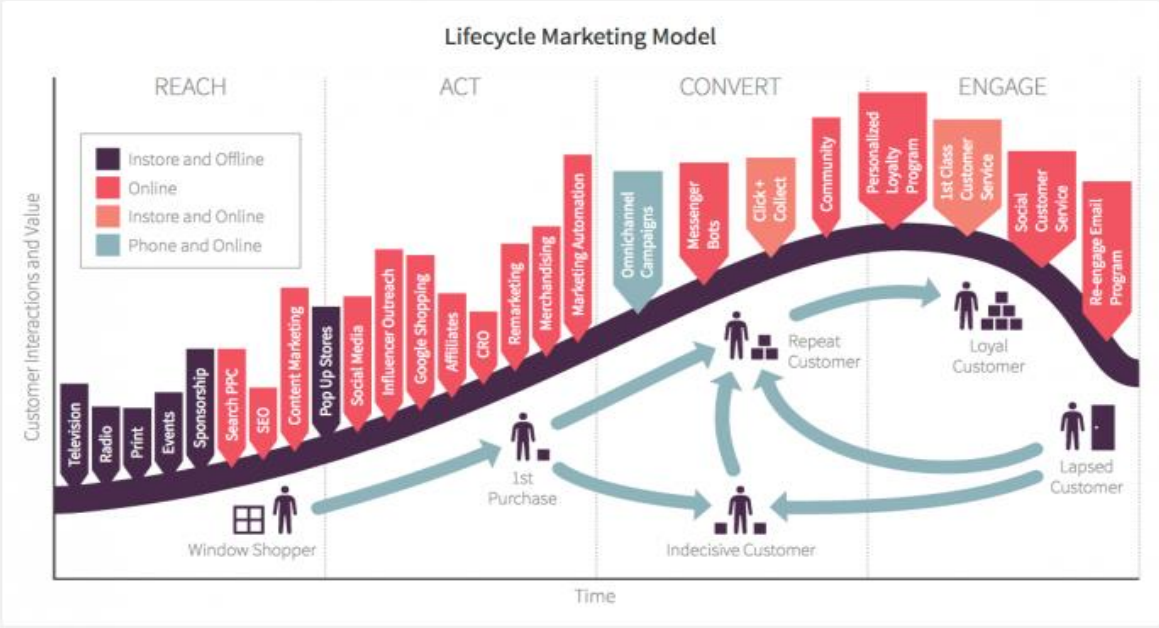

Figure 3: Lifecycle Marketing Model

Source: (Allen, 2016)

1. *Market Prospective:* online retail sales constitute mere 14% of cumulative retail sales in the United Kingdom and even lesser i.e. 8% in the United States – thus scope of augmentation is high (US Census Bureau News, 2017).

2. *Market Prospective:* Online trade by 53% of worldwide internet consumers in 2016 was approx. $1 Billion (US Census Bureau News, 2017).

3. *Market Prospective:* Sales in B2B model of e-commerce are projected to increase B2C sales substantially to reach $6.7 Trillion by 2020 (Venus Tamturk, 2016).

4. *Marketing Tools & Expertise:* Marketing technological tools for e-commerce are at boom and are easily accessible. As a matter of fact, more than 3,500 varieties of Market Technologies are available in 2016 (Scott Brinker, 2017).

5. *Stream of Traffic:* The 3 major streams driving traffic to e-commerce websites are CPC (19%), Email (20%) and Organic (22%). Still social media accounts and online display are less explored (Buras, 2016).

6. *Visit Pattern:* Too much lag in loading time of website repels 39% of the users to use the portal (Adobe, 2015).

7. *Visit pattern:* Non-attractive interface of website may lose 38% of audience while initial browsing (Adobe, 2015).

8. *Analytics in Marketing:* Predictive analysis on user browsing pattern are done by 49% of the best marketing teams in domain to tap potential consumers (Salesforce Research, 2016).

9. *Consumer Awareness:* User online history analysis is must to understand consumer and is actively done by 61% of the high performing marketing teams, 22% of the moderate performing and 6% of the low performing teams (Salesforce Research, 2016).

10. *Conversion Rate Optimisation:* As per a survey 44% of the e-commerce consumers leave the item-filled shopping cart if they see high charges in shipping or any surprising extra charges (Pienaar, 2015).

11. *Conversion Rate Optimisation:* A troublesome returns policy discourages 80% of customers (ReadyCloud, 2016).

12. *Customer Loyalty:* As per a research, to acquire a new consumer is almost 7 times more expensive than retaining an existing customer (Saleh, 2014).

13. *Multiple Channel Movement:* Shoppers who use multiple e-commerce platforms spend 3 times more than one-shop buyers (Allen Robert, 2016).

14. *Customer Satisfaction:* Bad post sales response repels almost 89% of shoppers to buy from any online store (Pienaar, 2015).

15. *Return Policy:* As per market survey almost 30% of the products purchased online are returned back to e-store and consumers expect easy return policy on it (Saleh, 2014).

2.2.1 Recommendation and Personalization System

The overall fulfillment and loyalty of customers are indirectly influenced by the recommended systems through drastic reduction in search cost and enhancing customer utility. Personalization system can gain valuable data of customers, along with increasing customer switching costs. The research study is segregated to two sections, namely content and technology. Eckert J and Hinz O conducted a study on the effect of search and recommendation system for retailers and it was established that the technology has a competitive edge (Hinz O and Eckert J, 2010).

Sin R and Chellappa R came up with a basic model for empirical analysis and they discovered that the utilization of customized service are associated with trust of customers on the vendors (Chellappa R. and Sin R, 2005). They further suggested few additional ideas to create customized services. Liu K, Gao M and others presented a brief outline on the personalization system and proposed few radical strategies and methods for these days (Gao M, Liu K and Wu Z, 2010). Erek K, Repschlaeger J, and others worked on distributed cloud computing to perform empirical observation on customers' inclinations and preferences of startup companies (Repschlaeger J., Erek K. and Zarnekow R., 2013). KieBling W, Holland S implemented data mining on the preference information of customers. The results could be extended to personal customer service, personalized product recommendations and direct marketing (Holland S. and Kießling W., 2004).

Creating individual relationships with every customer can results in greater loyalty on the organization with a reliable client base, which helps in higher productivity (Reichheld FF., 1996) (Reichheld F. & Sasser Jr. W., 1990). (Rauyruen P. & Miller K., 2007) Quoted that "the significance and advantages of pulling in and retaining loyal customers has emerged from a typical

acknowledgment of the fact that customer loyalty results in higher profitability". In simple words, the retaining loyal customers is one of the universal goal of several businesses. (Zineldin M, 2012), implies that a customer comes back to the business again and again to involve in new business transactions and purchases.

The customer can later come back to the company if they need anything and doesn't pay attention to any other businesses (Eriksson, K. & Vaghult, A, 2000). The concentration has moved from serving the mass market with an intention of gaining loyalty of the customers who comes back to the organization in the future. In simple words, there is a transition from transactions to relationship marketing (Gummesson, E, 1987). The fundamental financial results for any company with respect to customer loyalty and their retention are, higher revenues, lower costs, higher profitability, greater market share, and increased customer switching costs (Zineldin M, 2012).

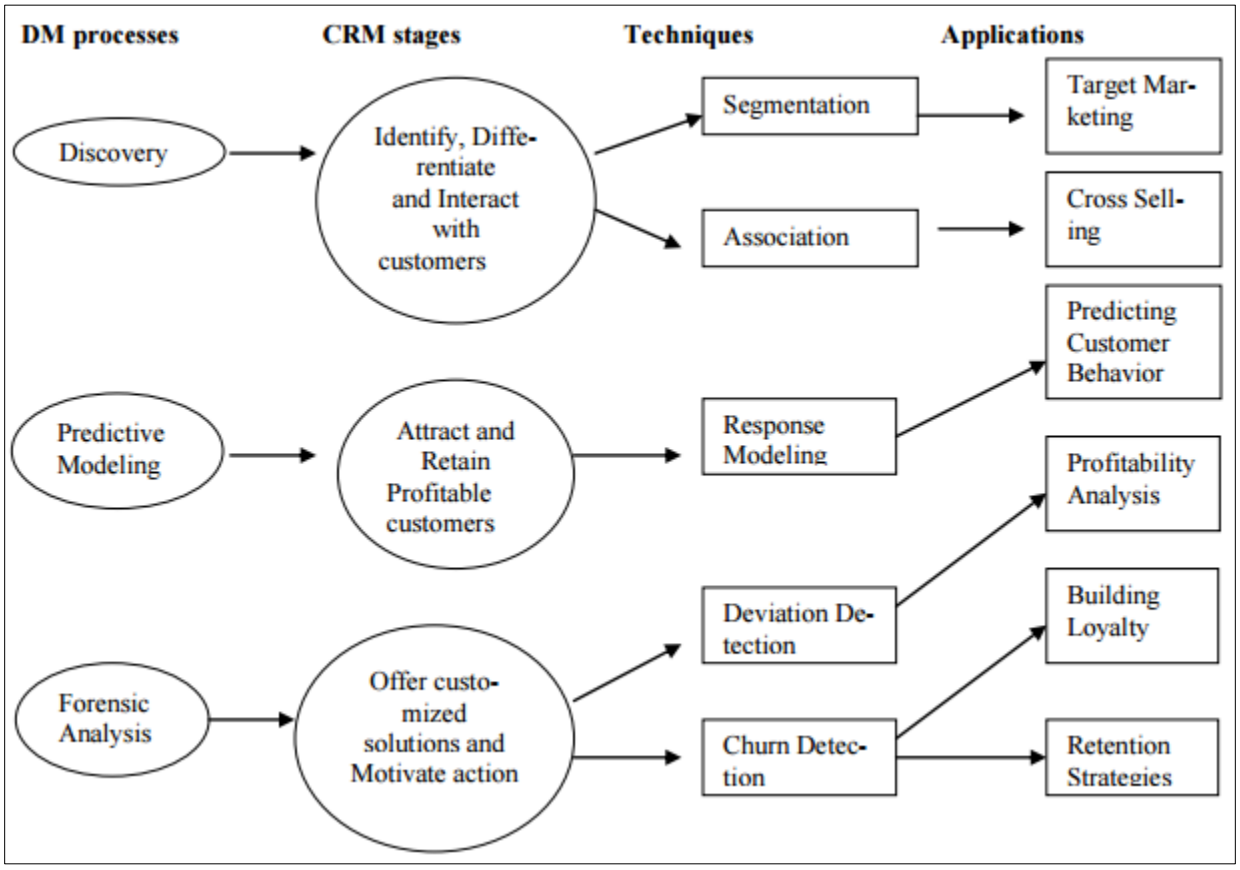

Figure 4: Integrated model of Data Mining for CRM

Source: (SS. Kadiyala & A. Srivastava, 2011)

2.3. Customer Relationship Management

Customer relationship management (CRM) comes as an advantage for businesses to maintain its operations and retain loyal customers, boosting their value, gaining loyalty and retention along with generation of client related policies. The main aim of CRM is to concentrate on wider perspective of integration of customer values, expectation, needs, demands, etc., with an intention to enhance the customer relations. For the purpose of maintaining, supporting, managing and retaining valuable customers, more and more business are implementing CRM (Customer Relationship Management) methodologies which includes the functionalities of data mining, data management and data warehousing, etc. (Cunningham C., Song Il-Yeol. Chen, Peter P., 2004). As per (Yeh IC, Yang KJ, Ting TM, 2009), the man concentration of CRM is to obtain and retain the highly valuable customers thought persistent insight of their values. Alternately, RFM (Recency, Frequency, and Monetary) model provides effective tools for customer segmentation (Newell, F, 1997).

2.3.1 Definitions of Recency, Frequency and Monetary

RFM Model is a segmentation technique which is widely popular due to its extended support of all three purchase related parameter, such as frequency, recency and monetary (Wei Jo-Ting, Lin Shih-Yen, Wu Hsin-Hung, 2010), that are consolidated into a three-digit RFM cell value. Altogether, it covers over five equivalent quintiles that sums up to around 20%.

Recency corresponds to average interval gap between last purchase and analyzed time. It is generally expressed in terms of number of days (days, months, and year) from the date of last purchase. Recency is generally expressed as the most significant measure of the three, as the latest products are probable to be purchased more frequently. Nevertheless, this is not a general expression, but instead, it varies depending on the product type and company.

Frequency refers to total number of purchases or transactions made by the customer over a specific duration of time. Larger the frequency, higher will be the loyalty towards customers, which translates to increased demand for the products.

Monetary measure denotes the aggregated sum of money that a customer spends over a duration of time. It also denotes average cost of all the purchased products (Hughes, A.M, 1996). (Marcus

C., 1998)Addressed the issue of co-linearity of frequency and monetary, by utilizing the average purchase amount. This highlights the customer who buy expensive products.

2.3.2 Recency, frequency and Monetary

RFM (Recency, Frequency and Monetary) model comes as a boon to decision makers to easily recognize those customers who are more valuable to the company, so that they can come up with effective marketing strategy (Wei Jo-Ting, Lin Shih-Yen, Wu Hsin-Hung, 2010). Nevertheless, such model has a major drawback. As it considers the previous purchase history, the model is restricted to existing customers and not new prospect customers. The inexistent purchase history of new customers makes the model miss out on expansion of customer base.

It is hard to distinguish and classify the existing customers based on the profitability. However, it is much harder to recognize new prospecting customers who can turn out to be valuable. As stated by (Zeithaml V. A, 2000), “creating a profile of the prospective customers needs a blend of comprehending the current profitable customers, depicting statistic and psychographic factors that can forecast the profitability and designing and testing key methods to gain new and qualify customers”. Nonetheless, there is not much experimental work on this context (Zeithaml V. A, 2000) (Reinartz W., Kumar, V, 2003).

RFM investigation technique is commonly used for recognizing reactive customers in marketing promotions. This helps in enhancing the overall response rates. It is very popular nowadays, and they are widely used (Wei Jo-Ting, Lin Shih-Yen, Wu Hsin-Hung, 2010). However, the significance of applying RFM scoring to a client database and evaluating customer profitability is less known (Aggelis V., Christodoulakis D, 2004). In the dawn of new technology, data mining techniques are tuned for RFM implementation, and they are actualized in several areas such as, automotive, computers, electrical, electronics, etc.

2.3.3 Recency, Frequency and Monetary analysis, Customer Lifetime Value and Customer Profitability

When it comes to taking decision, the priority of customer relationship management is always the profitability of customers. The RFM (Recency, Frequency and Monetary) has been generally implemented in numerous aspects of customer relationship management, especially in direct marketing and advertising (McCarty, JA, Hastak M., 2007) (Wei Jo-Ting, Lin Shih-Yen, Wu Hsin-Hung, 2010). Essentially, it can be considered as an analytical method which helps the direct marketers to perform better segmentation of their customers. Likewise, the RFM is a behavior-based model that assists in analysis of the customers' behavior and construct a prediction model based on these behavior (Hughes, A.M, 1996) (Yeh IC, Yang KJ, Ting TM, 2009). Despite availability of all the more statistical refined techniques, a survey by (Verhoef, Peter, C., Franses P. H., Hoekstra, Janny C., 2002) has demonstrated that cross tabulation is more frequently used for direct marketing, than RFM model. There are a few explanations behind the preference of cross tabulation over RFM model. RFM model is easier to create and maintain. Moreover, this technique can be easily understood by managers and decision takers (Marcus C., 1998). It is essential for any technique to distinguish the customers efficiently, as guided by the decision makers.

Other surveys indicate that RFM is more popular methods for analysis of customer values. In the benefit of this, it can derive the characteristics and behavior of any customer, with minimal input directions. Additionally, in terms of consuming behavior, the RFM is quite helpful in evaluating the quality of customer relationship. Maintenance cost of older customers is far less than procurement cost of new ones. Hence, several organizations use RFM methods to carry out data mining on the known customers to help them understand the customer behavior and draw more customers based on these findings (Cheng CH., Chen YS, 2009).

A few researches have concentrated on several variants of RFM model. The "quintile strategy" arranges the customer in descending order with respect to loyalty, from best to worst (Miglautsch, J. R., 2000). The overall amount of customers in every quintile remains unchanged at 20%, even though the size of each segment varies, as RFM cells vary in size. The hard-coding technique assigns scorings to their customers based on their behavior, so that numerous customers can be present in a quintile (McCarty, JA, Hastak M., 2007). Practically speaking, direct marketing

managers must choose certain criteria for the RFM segments, which comes from their experience and wisdom.

For example, the marketers can choose the coding intervals for recency as 0-3 months, 3-6 months, 6-12 months, 12- 24 months, higher than 24 months etc., which can be coded as 5, 4,3,2,1, respectively. RFM scoring using weighted approach is also introduced (Miglautsch, J. R., 2000). It involves allocating weights to RFM cells and then the final RFM score can be generated through simple addition of these weights. The equations for addition of weights in the RFM cells are distinctive. According to (Miglautsch, J. R., 2000), the equation for scoring techniques is given as (R x 3) + (F x 2) + (M x 1). (Hughes, A.M, 1996) Proposes to utilize similar weights, so that the cells (5, 3, and 5) would result in an overall score of 13. However, this technique is dictated by product of type of industry. As the assigning of initial weights depends on the experience and knowledge of the marketer, this technique is also known as judgment based RFM (McCarty, JA, Hastak M., 2007).

2.3.4 Applications of Recency, Frequency and Monetary Model

The idea of RFM is based on 80/20 Pareto principle, which states 80% of overall sales are contributed by just 20% of customers. This model is implemented on the RFM values for file segmentation containing existing customers. However, it is not suitable of new customers as there is apparent lack of purchase information of prospect customers (McCarty, JA, Hastak M., 2007). Combination of data mining systems and RFM investigation offers valuable data on existing and new customers.

The best way to predict a customer's future buying trends can be done by observing their past purchase history. With the help of RFM, the existing profitable customers are discovered, which are then translated into profitable customers of the future. For expressing CLV, numerous researches have used the RFM model (Liu DR, Shih YY, 2005) (Miglautsch, J. R., 2000) (Sohrabi B, Khanlari A, 2007). Customer lifetime value (CLV) are generally exploited to discover profitable customers in the existing client database and to develop some techniques to give higher preference to those customers (Irvin, S., 1994). It is described as the current value of upcoming profits based

on the customers' relation with the company (Sohrabi B, Khanlari A, 2007). CLV is defined differently in various researches, but the difference is quite low.

Few of the definitions are given as follows,

- The overall discounted profit generated by a customer over her time on the house list.
- Achievable profit from the customer, excluding cost of relationship management.
- The net present value of all upcoming contributions to overhead of cost and profit.
- The net present value of the course of provisions to profit which is a consequence of customer transactions.
- The net present value of all upcoming contributions to profit and overhead anticipated by the customer.

Therefore, CLV can be easily computed to determine the profitable customers, as estimating RFM model is quite significant for assessing customer lifetime value (Liu DR, Shih YY, 2005). Apart from its advantages and performance, this technique is used to estimate customer profitability. (Sohrabi B, Khanlari A, 2007) Developed a conceptual mode, as shown in Figure 5, for customer segmentation based RFM model using their transactional data. They additionally presented few techniques to retain such customers.

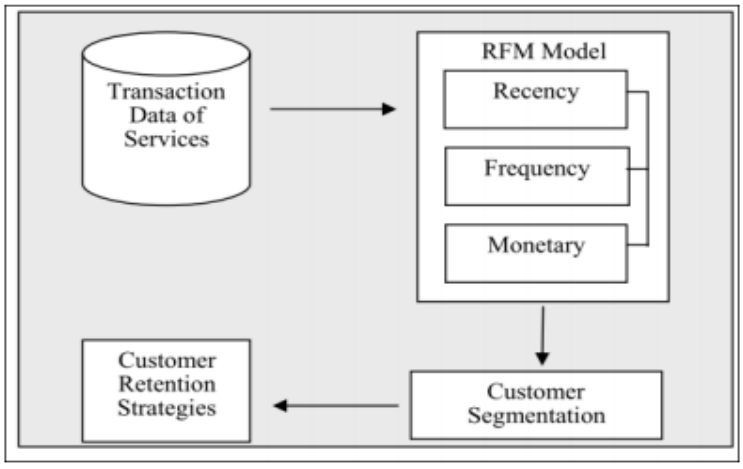

Figure 5: Conceptual framework of research by Sohrabi and Khanlari (2007)

The computation of CLV involves numerous equations and calculations. The normalization of F (frequency) and M (monetary) values is done using form $x' = (x - x^s)/(x^L - x^s)$, as these two parameters are directly linked to CLV. The recency (R) value can be normalized using the cost form, $x' = (x^L - x)/(x^L - x^s)$. x' is the normalized RFM values, and x is the original RFM values. x^L is the largest RFM value, whereas, x^s is smallest RFM value of all customers. Numerous other studies have also used this approach (H.M. Chuang, C.C. Shen, 2008). In certain studies, the normalized RFM values are multiplied by relative weights of RFM model of each customer, which are denoted as w_R, w_F and w_M (Liu DR, Shih YY, 2005).

Chu Chai et al. (2008) used the RFM model for investigation of customer behavior. The author also uses LTV model, or Lifetime Value model. A high performance genetic algorithm was proposed to choose customers behavior based on this LTV model. Razieh et al. (2012) used the clustering algorithms and RFM model to conduct customer segmentation. The implemented utilized the customer value to depict profitable and loyal customers from a grocery store. The demographics, characteristics and other attributes were considered to describe and identify profitable customers.

Several researches on the CRM employs data mining algorithms to analyze the attributes and behavior the customer (Fletcher et al 1993; Bortiz et al.1995; Lau et al 2003; Langley et al 1995). These studies have illustrated that the data mining techniques can identify and promote hidden knowledge or information present in the vast customer database. The main objectives of any data mining technique is to describe, predict and convey the knowledge/information (Fayyad et al. 1996). CRM uses information technology (IT) to interact with customers. It helps in understanding needs of customers and serving those needs helps in enhancing their lifetimes (Verhoef 2003). On the other hand, segmentation is a simple task of segregating the customers in to distinctive groups depending on their behavior and characteristics, which is one of main part of the CRM module (Verhoef 2003).

Numerous business problems can be addressed with the help of data mining, six different types of models are developed, such as time series, regression, classification, association analysis, clustering and sequence discovery (Thearling K. 1999). Predictions are done using classification and regression. Behaviors are described using association and sequence discovery. Clustering is helpful for both description and prediction. Classification is one of significant type of predictive

modeling (Weiss and Indurkhya 1998; Dunham 2003). The data are classified into groups in the supervised learning using classification technique, which are reclassified before the investigation of information (Dunham 2003).

Deriving the pattern of data is the main objective of descriptive data mining. Association rules, cluster analysis, sequence discovery and summarization are some of the well-known examples of descriptive data mining (Dunham, 2003). Conversely, the future values and variable are predicted based on the previous values (Fayyad et al. 1996; Tan et al. 2001). It follows predictive modeling approach, which is a process of developing target variable as a dependent on descriptive variable (Berger and Barbieri M.M 2004).

2.3.5 Supervised learning techniques for Customer Relationship Management

Every CRM element is sustained by various data mining models. The background details on the supervised learning techniques are given as follows:

Association modeling is generally implemented for cross selling programs and market basket analysis (Mitra et al. 2002).

A study by Nan-Chen Hsieh et al. (2009) on evaluation of bank databases showcases a two-stage model of consumer behavior analysis. The unique aspect of this technique is a cascade involving Self-Organizing Map (SOM) neural network system. This methodology segregates the customers into identical groups of customer objects. The relevant information is then portrayed using a decision-tree simplified method. The decision tree inducer and the statistical summarized data are utilized to differentiate the groups of customers, soon after SOM identifies the profitable customers. For the purpose of determining relevant classification rules, tree simplified mechanism is implemented.

Siavash Emtiyaz et al. (2011) conducted an investigation on the semi-supervised learning technique. The main concentration of the study was focused on the management and analysis of data warehouse and information related to customer. The concept of semi-supervised learning is to utilize structural information present in the unlabeled data, and not just from the labeled training data. A feed-forward neural network is used for semi-supervised method. A multi-layered back propagation algorithm is used for training purpose. It helps to predict the class for unfamiliar and unidentified customers.

Keyvan (2012) implemented three different supervised classifiers namely bayesian network (Bayes and Price 1763) decision tree and neural networks. The objective of this study was to predict the behavior of customers in an Iran based insurance company. The experimental results showed that the decision tree algorithm outperformed other algorithms.

Rekha (2011) worked on Decision Tree-Based classifier and Naïve Bayesian classifier to perform prediction and analysis of fraudulent claim in automotive insurance policies. The investigation of performance is based on the confusion matrix.

2.3.6 Unsupervised Data Mining techniques for Customer Relationship Management

Generally, the unsupervised learning doesn't depend on predefined classes or class-labeled training cases (Han J. and Kamber M 2008). Alternatively, clustering is a type of training which includes observation instead of training by examples. One of the significant forms of data mining techniques in customer relationship management is the customer clustering. It is involved with analysis of transactions and purchase history of customer, by keeping track of their buying habits and impose new business development strategies.

On the other hand, the unsupervised data mining strategies doesn't involve any manual intervention, but it relies on training example. One of the best example is the self-organization map clustering, which is a prominent methods of experimental investigation, when no priori classes are discovered. Self-organization map clustering or simply known as SOM, is a graphical tool and involves powerful abilities to deal with missing or skewed information, non-linear relationships, etc.

Ruey-Shun Chen et al. (2005) made exemplary utilization of data mining techniques to effortlessly determine the ongoing purchase trends of customers and track the patterns of behavioral variation. The data on this experimental study was based on the credit card summary of the customer data during 2003 and 2004. The research concentrated on the spending habits of the customers which also considered all the purchases by a single user using multiple credits cards. The selected customers are later classified into groups, known as clusters with the help of RFM model. The RFM model assists in gaining profitable customers. Therefore, data mining utilizing is implemented using association rules to evaluate the spending habits of purchasers.

Horng et al. (2007) worked on mapping of star products with prospective customers. The prospective customers are identified by performing clustering analysis on the personal information of loyal customers. The future interests of prospective customers on the star product are predicted using association rule analysis with the help of purchase history of loyal customers

The significance and vitality of Data mining methods are strongly endorsed by Jayanthi Ranjan (2011), particularly in context with managing Customer Relationship Management (CRM). It is done by estimating the unknown and obscure data obtained from a genuine insurance agency. The customer satisfaction level was carefully analyzed by the author using clustering technique, for the purpose of profitability.

Dilbag singh et al.(2012) undertook comprehensive research on the conceptual mapping of numerous undertaking of insurance risk management and presented few data mining strategies for risk prioritization, risk analysis, risk identification, risk monitoring, planning, etc.

2.4 Customer Segmentation and Data Mining Techniques

Customer segmentation can be described as the process of organizing the customer database through classification into different classes. Alternately, it involves assigning each and every customer to a distinctive group, based on their characteristics (Rinta-Runsala and Bounsaythip 2001). The objective of segmentation is to understand the existing customers and use this information to gain new customer, lower the operating costs, boost the service and increase profits. Segmentation can offer multidimensional perspective of the customer for enhanced service rendering. It can be described as assigning customers with identical attributes into certain groups. The attributes can be based on demographics, topographical or behavioral characteristics, and marketing them as a group (Sheth and Parvatiyar 2001). Therefore, in a group, the members' express similar requirements, which are not always consistent. Segmentation needs aggregation, ordering and analyzation of customers' information. The identification of customer loyalty can be done through good segmentation techniques, as it directly translates to overall profitability.

Segmentation can be considered as a task of creating significant customer groups which is dictated by personal characteristics and attributes (Trappey et al. 2009).

According to Greengrove (2002) there are two segmentation approaches;

1. Needs based segmentation: the needs and requirements of the users should be well described.
2. Characteristics-based segmentation: segmentation is performed on the grounds of behavior and attributes of the users.

Twedt (1964) recommended the segmentation techniques based on sale volumes should concentrate on customers, who perform higher transactions. This technique is termed as “heavy half theory”. It suggests that only half of the total customers can contribute to over 80% sales.

In 1970’s, the rationality of the multivariate approaches were questioned in their methods to detect the variables that influence the deals (Green and Wind 1973). This inspired the design of powerful models of consumer behavior (Blattberg et al. 1978). Around 10 year later, a generalized psychographic segmentation technique was developed by Mitchell (1983), which segregates the markets based on lifestyle, social class and individual attributes. Nevertheless, the practical implementation of this technique was quite difficult (Morgan and Piercy 1993, Simkin and Dibb 1997).

Rinta-Runsala and Bounsaythip (2001) indicated that segmentation can also be considered as a technique to have a direct communication with the customers. Segmentation process is depicted by the characteristics of the customers groups. The segmentation becomes less effective due to diversity in customer’s lifestyle, buying trends, spending habits, age, income, etc. So, the marketing segmentation are generally based on the perceived customer behavior with respect to previous purchase history. The results are then analyzed using data mining techniques. Kiang et al. (2006) conducted a survey on the various data mining techniques for segmentation in this context.

In numerous researches, the segmentation models are chosen to determine customer profitability (Hwang et al. 2004; Kim et al. 2003; Kim and Street 2004; Shin and Sohn 2004; Kuo et al. 2006; Woo et al. 2005).

Green and Helsen (1991) have postulated new market segments based on computer driven clustering techniques using customer data obtained from various surveys. The segmentation is upheld by the rate of significance given to the product characteristics.

Han and Min (2005) worked on clustering of customers based on their movie preferences. The data was collected through numerous movie rating that are individually pleased different customers. These ratings describe the different views of customer to a particular movie.

Chu Chai et al. (2008) used RFM model to identify customer behavior. Later, Lifetime Value (LTV) model was used to perform customer segmentation. Genetic algorithm was used to create an efficient technique to choose RFM characteristic of customers using LTV model. The fitness value of genetic algorithm is the customer lifetime value. This technique has several advantages. It helps in identification of valued customers for marketing campaigns. The correlation between customer values and campaigns are also considered.

Razieh et al. (2012) implemented customer segmentation techniques, through clustering algorithms and RFM model on the customer value, to define and identify profitable customers. The data was taken from a grocery store. A blend of demographical and behavioral attributes are used to find the loyalty.

2.5.1 Classification techniques for customer segmentation

Supervised classification is also known as Classification analysis. It is a technique of determining a model which defines and recognizes the data classes. The main aim of this technique is to exploit those model to carry out prediction of class of objects whose class label is not properly known (Han and Kamber 2008). Data can be easily organized using such classification schemes. The classification makes use of given class labels to arrange the objects present in the data collection in a proper order. In atypical classification technique, the user must divide the data into segments, which is followed by creating separate non-overlapping groups. It is essential that he user must possess some knowledge on the data to efficiently divide the data into groups. Therefore, the classification techniques helps is recognizing those characteristics, which suggest the suitable group for an object. Classification models can be easily created using data mining by analyzing already classified data and determining a predictive pattern (Westphal C. and Blaxton T. 2005).

In general, classification techniques utilize a set of training examples, where the participating objects are already associated with some class labels. Then a model is built based on the learnt information of the classification algorithm. With the help of this model, new objects can be easily classified. In general, the classification involves two steps. In the first step, a classification model will be generated on the basis of training examples. In the second step, this model will be applied to carry out classification for new data sets (David Hand et al. 2008).

Classification process is implemented for numerous CRM applications. Kim et al. (2006) used decision tree algorithm to perform classification of customers and introduced a technique with respect to customer lifetime value. Baesens (2004) used Bayesian network and identified the slope of the customer lifecycle. The author also contends that the Bayesian network classifiers provides high performance for CRM based application for determining the slope of the customer lifecycle.

Sheu et al. (2009) explored possible connection between customer loyalty and important influential factors. This task was performed using decision trees. This study helps in implementing decision trees to determine the connection among different demographics of customers, purchase costs, buying trends, etc.

2.5.2 Clustering techniques for customer segmentation

Clustering can be characterized as a method of segregating and grouping an array of physical or conceptual objects into classes of categories of similar items (Han and Kamber 2008). Clustering is generally referred as unsupervised classification, as the classification process is not directed by class labels.

The main idea of clustering technique is to increase the similarity between objects of similar class, and reduce the similarity between objects of dissimilar classes, which are called as intra-class similarity and inter-class similarity, respectively. Based on this principle, there are numerous clustering approaches.

Even though clustering is found to be identical to classification, the definition of classes are given in the former, and the clustering algorithm must perform the same. Generally, it is important to fine tune the clustering technique by discarding the variables which are used to group instances.

The main reason is that these variables are found to be meaningless or irrelevant to the user. After completion of finding of clusters, the classification of new data can be performed, followed by the database segmentation. K-means clustering (Lloyd, 1982) and Kohonen feature maps (Kohonen, 1982) are some of the popular clustering techniques.

Clustering should not be mistaken for segmentation. Clustering is just a method of grouping the data into different groups without any pre-defined characteristics of the objects, while the segmentation involves identifying groups that have similar characteristics (Two Crows Corporation 1999).

Clustering offers good means of grouping data into groups, known as clusters. All the objects in a cluster will possess similar characteristics. Such clustering techniques can be implemented to put all the customer with common characteristics into a single group, so that making business decisions becomes easier (David Hand et al. 2008).

The studies on Clustering is also termed as unsupervised learning, which is a process of classification with an unknown agenda. Which suggests that the class types are not known. The objective of the algorithm is to perform segmentation of objects, and group them into disjoint classes which are uniform to given inputs (Han and Kamber 2008). The studies on Clustering did not have any dependent variables. Clustering is most vital process of data mining technique, which focuses on identifying groups and patterns with the underlying data. Clustering problem involves dividing the data sets into cluster in such a way that the data objects of similar kind falls in same cluster (Guha et al. 1998).

Clustering technique is predominantly exploited to realize object profiles based on the variables of the objects. The objects refers to documents, customers, users, features. Hruschka (1996), Weber (1996) Ozer (2001), employed clustering scheme to segment markets and customers. Some of the popular clustering techniques are Kohonen self-organizing map (Kohonen, 1982) and K-means clustering (Lloyd, 1982).

Samira et al. (2007) has fruitfully implemented segmentation of customers of TPO (Trade Promotion Organization) in Iran. The implementation of this algorithm used proposed distance function that gives the idea of dissimilarities between export containers of various countries using association rule. Every cluster is analyzed using RFM model for determining best technique for

promoting each segment. Factors utilized as criteria for segmentation are, “the type of group-commodities”, “the value of the group commodities”, and “the correlation between export group-commodities”.

Pramod et al. (2011) extends the implication of clustering scheme to perform segmentation of customers in a retail store. The investigation revealed that the K-Means clustering (Lloyd, 1982) helps the retailers to understand customer needs and take fast decisions to offer good and customized services to the customers.

Huang et al. (2009) made a successful attempt to analyze the customer value, using Fuzzy C-means clustering, K-means clustering (Lloyd, 1982) and bagged clustering algorithm. It was performed on a hunting store in Taiwan. The experimental results have indicated that the bagged clustering algorithm was better compared to other.

Hosseini et al. (2010) performed classification of customer loyalty based on RFM values, using K-means clustering (Lloyd, 1982). Chen and Cheng (2009) performed segmentation of customer values on the basis of RFM value, using rough set theory and K-means clustering (Lloyd, 1982). Chen et al. (2009) recognized buying trends with the help of sequential patterns.

Migueis V.L et al. (2012) presented a strategy for effective segmentation of customers through analysis of their purchase history and types of products purchased. The author performed segmentation customers of a European retail business with respect to their buying habits and lifestyle. Based on the obtained findings, the authors have presented few marketing strategies which are specific to each segment of customers. However, the main aim being able to gain new customers and increase sales and profits of the organization. Accordingly, VARCLUS algorithm was introduced, which would directly mingle with SAS tool. The algorithm was also extended to perform clustering based on type of products purchased by the customers too.

Chapter 3 Research Methodology

3.1 Research Methodology

As per the definition by a noted researcher Jibendu (Jibendu Kumar Mantri, 2008), research methodology is a mechanism to solve the problems involved in research in an organized way. Due to its characteristic of carying out of research in systematic way, it is considered as a science.

To achieve knowledge on this domain and to solve the problem statement, numerous methodologies were used, such as, through analysis of literature available online or offline, e-books, journals, articles, support from supervisor, data collection methods, etc.

In this research analysis was done on e-commerce dataset provided by Viata Company. I used RFM analysis model to find loyalty of customers in customers' transactions dataset. The main objective of this study is to do customer segmentation using RFM model and classification of new customers as loyal or non-loyal using various pre-defined algorithms of data mining.

3.1.1 Data Mining & WEKA

Data mining is an old and effective way to carry out analysis on a dataset which has multiple attributes and a huge number of records. There are lots of references available on the internet regarding data mining. Data mining techniques are really interesting to work with when there are lots of algorithms available in single tool such as WEKA and comparisons are to be made among the techniques. WEKA tool is most commonly used for the simulation purpose and provide in-built support for most of the data mining algorithms.

3.1.2 Understanding Data

Dataset used in this research was provided by a Belgium based pharma e-commerce company called Viata and collected through its website database. All of the transactions carried out by customers are stored initially in a RDBMS tool which was later converted to CSV and ARFF format to perform analysis. From this database, we have designed a data warehouse that comprises of a wide variety of products, descriptive information on each customer and transactional data.

The transactional data consist of 114,212 customers who have made purchases from the Viata website from January 2016 to December 2016.

3.1.3 Approach

Quantitative techniques were used to carry out this research so that measurable stats can be achieved. Subsequent to the idea of using RFM analysis, a deep literature review was conducted to gain good knowledge in this area. After reviewing various literature, suitable assumptions were established. Subsequently, the data was properly shaped out as a result of the complete course of cleaning and transformation of it. Next, the selected data mining classification algorithms Decision Tree (J48) (Quinlan, 1993), Decision Tables, K-nearest neighbor (IBK), Random Forests (Breiman, 2001) and Bayesian Networks (Bayes and Price 1763) were applied to the data using WEKA tool to explore the prediction rates using a 10-fold cross-validation. Finally, the results were analyzed, interpreted and conclusions were made. High-level research approach is illustrated in the figure below.

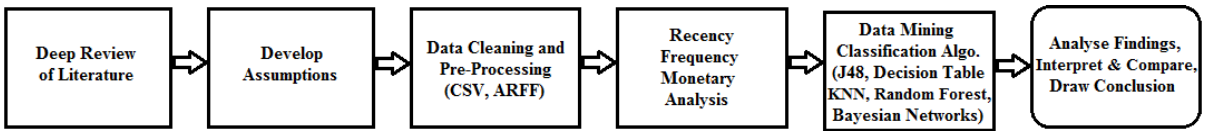

Figure 6: High level approach

3.1.4 Data pre-processing

Data pre-processing is the most significant, complex and time taking task in the whole process of data mining. In this phase, cleansing and formatting of data is done so that it could be used for the RFM analysis. Subsequently, selection of required fields are done so that the attributes which are essential for RFM are present and others can be kept as parameters to compare characteristics and behavior while doing test analysis with single order dataset. Important attributes for RFM in our dataset were client_id, date and revenue.

By utilizing transaction dataset and clustering records by customer id, Recency and Frequency can be derived, also Monetary can be analyzed by aggregating revenue by individual customer. Formulas to derive these are discussed in next section.

3.1.5 Recency, Frequency and Monetary Model

To derive RFM model and do analysis on new data to classify a customer as loyal or not, we should have certain criteria to consider.

About the RFM legend attached to clients, which ranges from 1-5:

Recency: Most recent order

5 <= 3 weken

4 > 3 weken <= 2 maanden

3 > 2 maanden <= 4 maanden

2 > 4 maanden <= 9 maanden

1 > 9 maanden

0 Geen order

Frequency: Number of orders in last 12 months

5 > 8

4 6 - 7

3 4 - 5

2 2 - 3

1 1

0 Geen order

Monetize: Average spending per order in last 12 months

5 > 65 euro

4 > 55 <= 65 euro

3 > 45 <= 55 euro

2 > 30 <= 45 euro

1 <= 30 euro

0 Geen order laatste 12 maanden

Highest sales: Highest spending per order in last 12 months

5 > 150 euro

4 > 75 <= 150 euro

3 > 45 <= 75 euro

2 > 10 <= 45 euro

1 <= 10 euro

0 Geen order laatste 12 maanden

Start: When first order placed

5 > 24 maanden

4 > 12 <= 24 maanden

3 > 6 <= 12 maanden

2 > 3 <= 6 maanden

1 <= 3 maanden

0 Geen order

Another important aspect is how to label customers (the decision class in the classification problem). VIATA considers that a customer is loyal if he/she has high recency and frequency values. As an initial step, we defined two loyalty categories: no_loyal and loyal, which is again defined as follows:

*IF recency*frequency > 16 THEN*

user = loyal

ELSE

user = no_loyal

3.2 Research Steps

Information of clients with a single order, with two orders, with three orders and so on are essential. If the maximal number of orders is too high, we segment the orders, for example: clients with a single order, clients with orders from 2-5, more than 5 orders, etc.

The hypothesis is that we can obtain better prediction rates if we include clients with more orders. If not, then the result is even better because it implies that we can make good predictions based on demographic factors. Algorithms using a 10-fold cross validation were used to have a good prediction rates.

Two datasets are essential: **viata-loyalty_clean.arff** with the information of customers with more than one order and **viata-loyalty_one-order.arff** with the information of customers with a single order.

The plan is to explore the prediction rates using a 10-fold cross-validation for the following classifiers: Decision Tree (J48) (Quinlan, 1993), Decision Tables, K-nearest neighbours (IBK, which stands for Instance Based K), Random Forests (Breiman, 2001) and Bayesian Networks (Bayes and Price 1763). For this first experiment, **viata-loyalty_clean.arff** dataset will be utilized for training and testing.

As a second experiment, we will train the above classifiers using the **viata-loyalty_clean.arff** dataset, and next test them using the **viata-loyalty_one-order.arff** dataset. In other words, we will learn from the data of consolidated customers (first dataset) in order to predict whether fresh customers (second dataset) will be potentially loyal to VIATA.

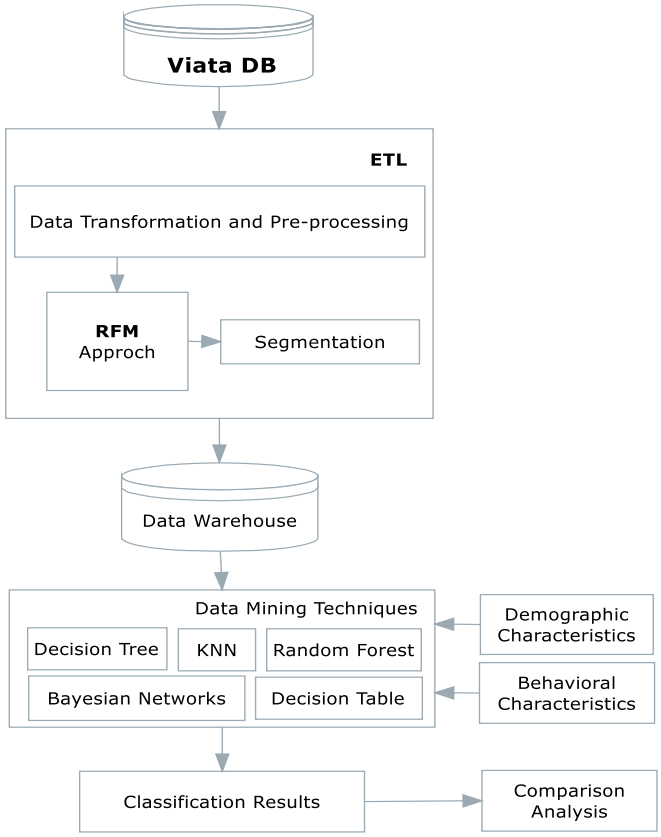

Figure 7: Research Process

In the second experiment the prediction rates are not so important, but the number of customers that are labelled as loyal. This is tricky because new customers (second dataset) are automatically labelled as non-loyal by definition, so we are expecting that the trained classifier fails in label some cases as non-loyal as a way to identify potentially loyal customers. This modelling belongs to a branch of Machine Learning called Semi-Supervised classification. In summary, the number of predicted loyal customers vary depending on the classifier's accuracy, which is expected.

Chapter 4 Main Results and Discussion

This chapter deals with the results obtained after executing few tests on clean and single order dataset of consumers using various classifiers and analysis on those results. To obtain the results, WEKA tool was used to process the dataset through classifiers. Basic steps to perform the test were, first, RFM is applied to measure loyalty of the customers in the dataset which contain all the transaction records. After finding the loyalty of each customer, classifier algorithms are used to train the system with this dataset along with loyalty attribute. Subsequently, a new dataset containing just single record of each customer is tested to analyses the pattern and behavior. As a result, percentage of customers classified as loyal or non-loyal is obtained.

4.1 Classifiers' Prediction Rates using a 10-Fold Cross-Validation

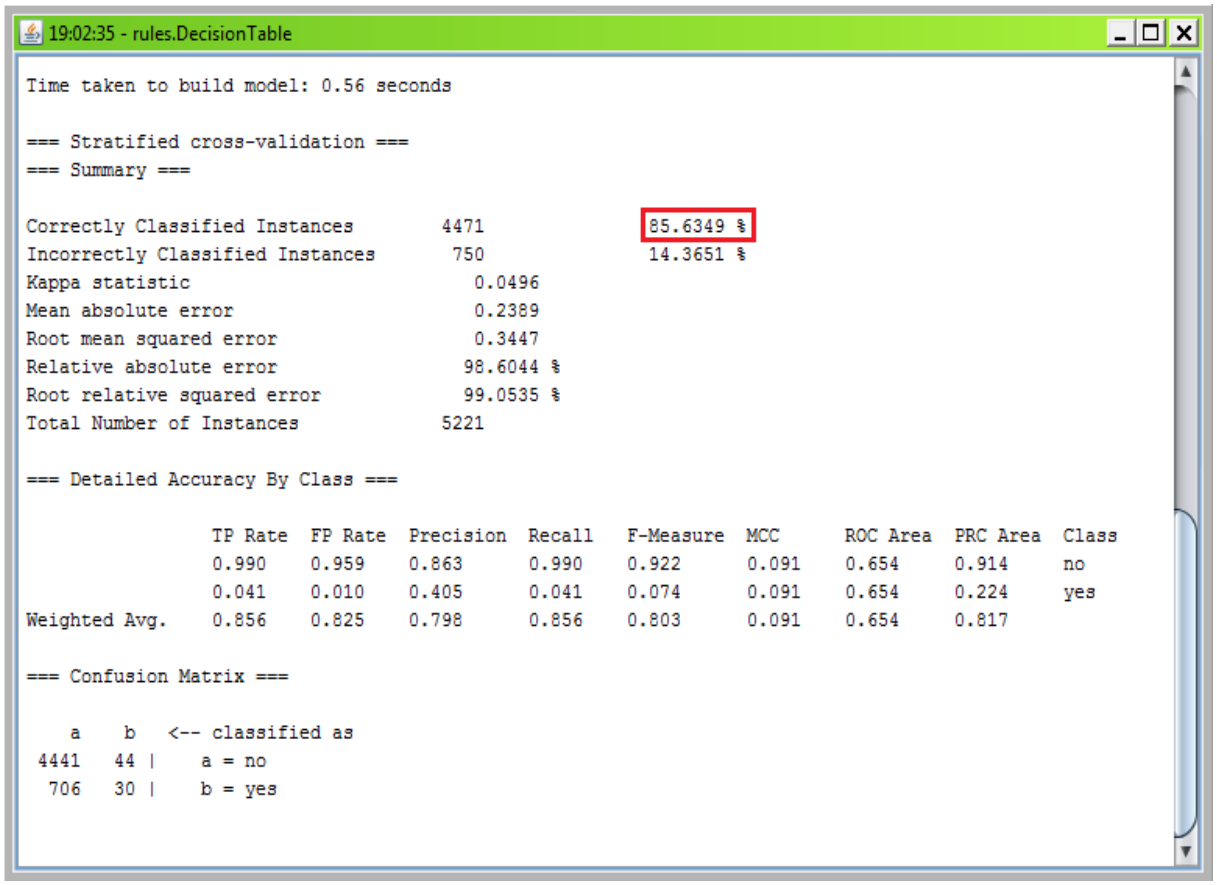

Figure 8: Decision Table Prediction Rate

In the above figure 8, prediction rate of Decision Table is demonstrated using 10-fold cross validation. It is highlighted in the figure that percentage of correctly classified instances is 85.6349 and number of instances are 4471. Total no. of incorrectly classified instances is 750 that account for 14.3651%.

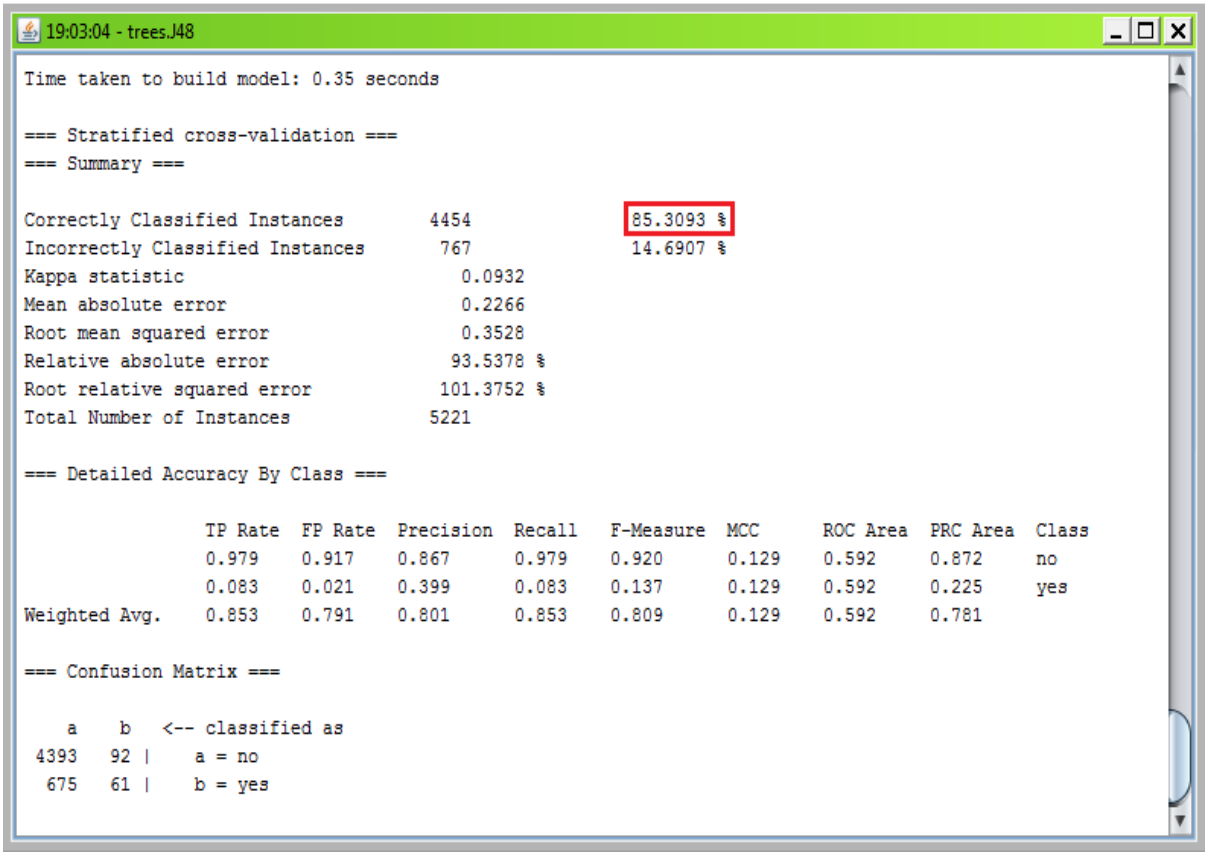

Figure 9: Decision Tree (J48) Prediction Rate

In the above figure 9, prediction rate of Decision Tree (J48) (Quinlan, 1993) is demonstrated using 10-fold cross validation. It is highlighted in the figure that percentage of correctly classified instances is 85.3093 and number of instances are 4454. Total no. of incorrectly classified instances is 767 that account for 14.6907%.

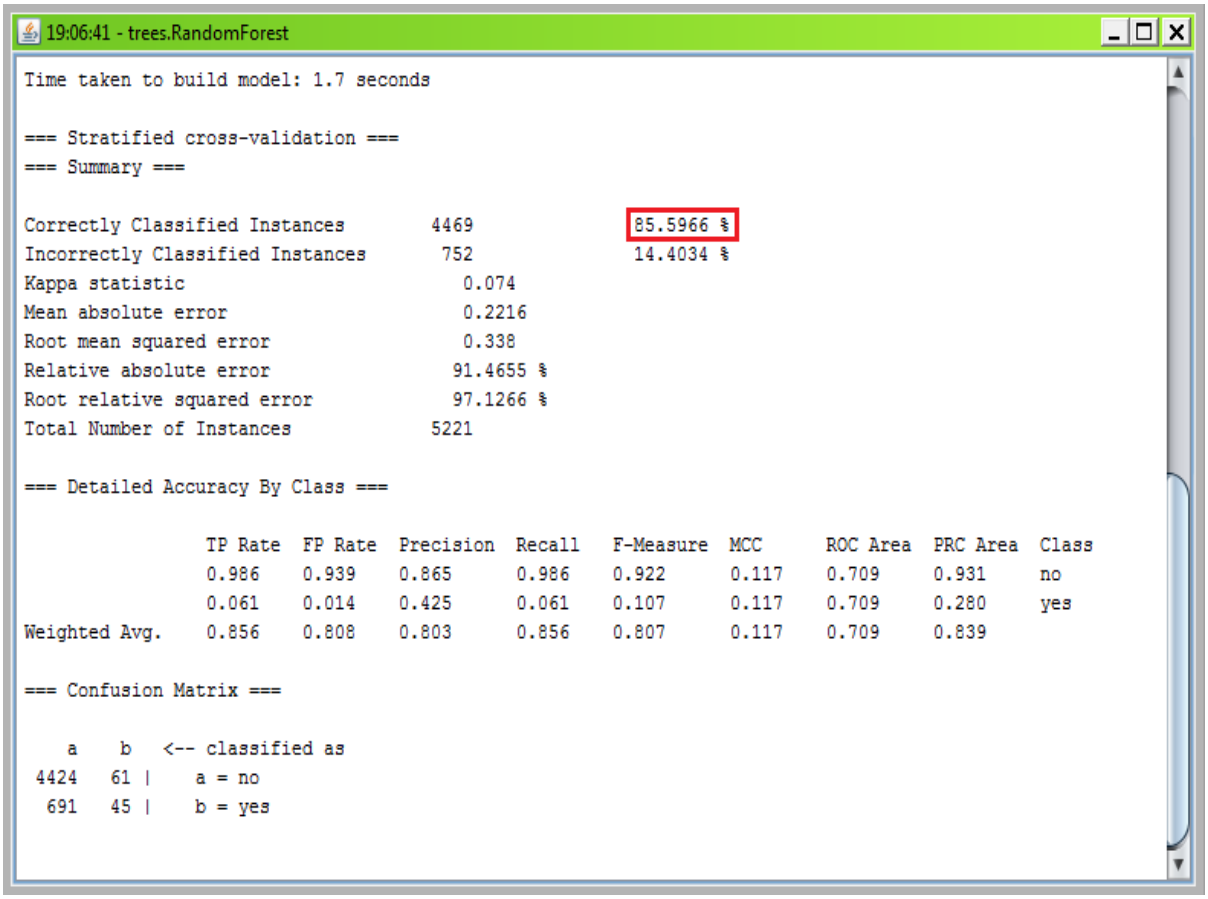

Figure 10: Random Forest Prediction Rate

In the above figure 10, prediction rate of Random Forest (Breiman, 2001) is demonstrated using 10-fold cross validation. It is highlighted in the figure that percentage of correctly classified instances is 85.5966 and number of instances are 4469. Total no. of incorrectly classified instances is 752 that account for 14.4034%.

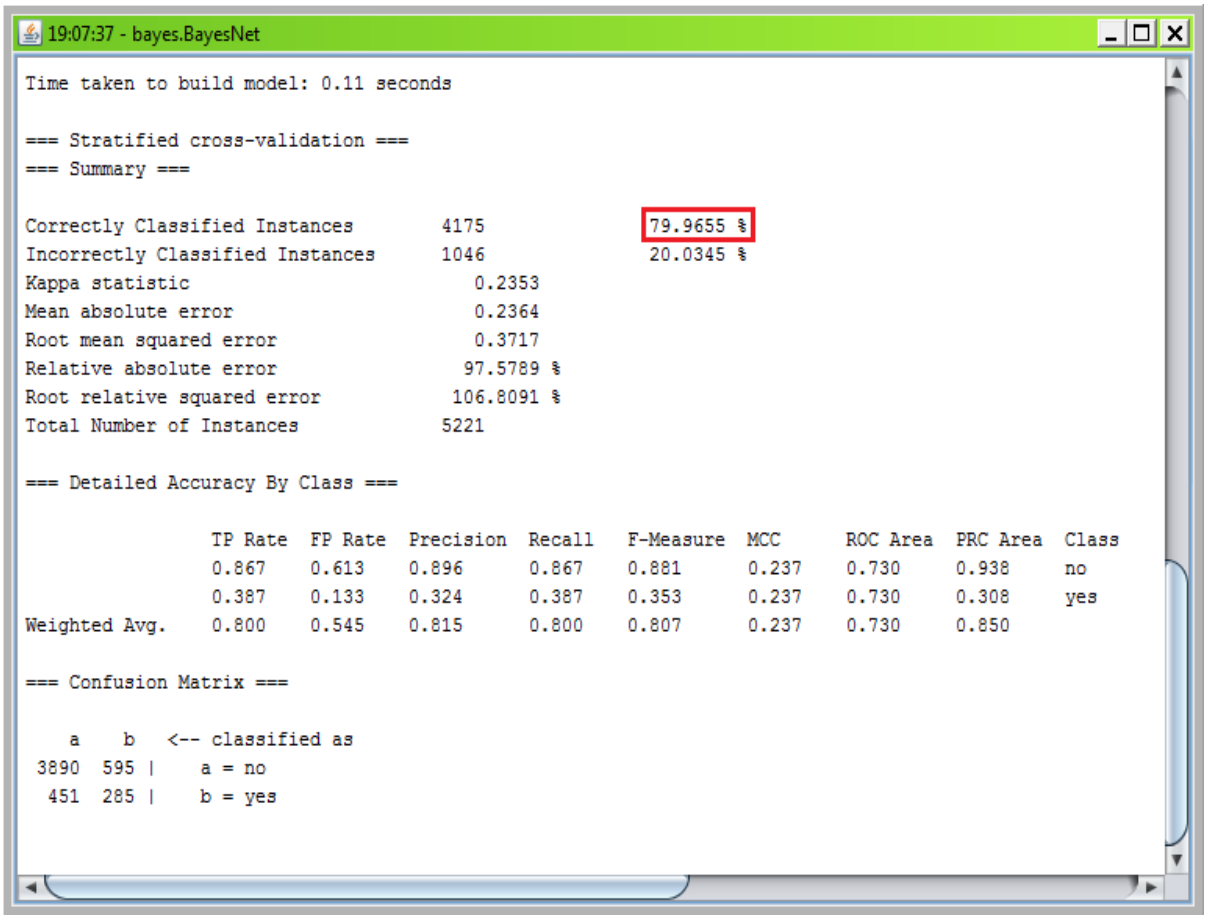

Figure 11: Bayesian Network Prediction Rate

In the above figure 11, prediction rate of Bayesian Network (Bayes and Price 1763) is demonstrated using 10-fold cross validation. It is highlighted in the figure that percentage of correctly classified instances is 79.9655 and number of instances are 4175. Total no. of incorrectly classified instances is 1046 that account for 20.0345%.

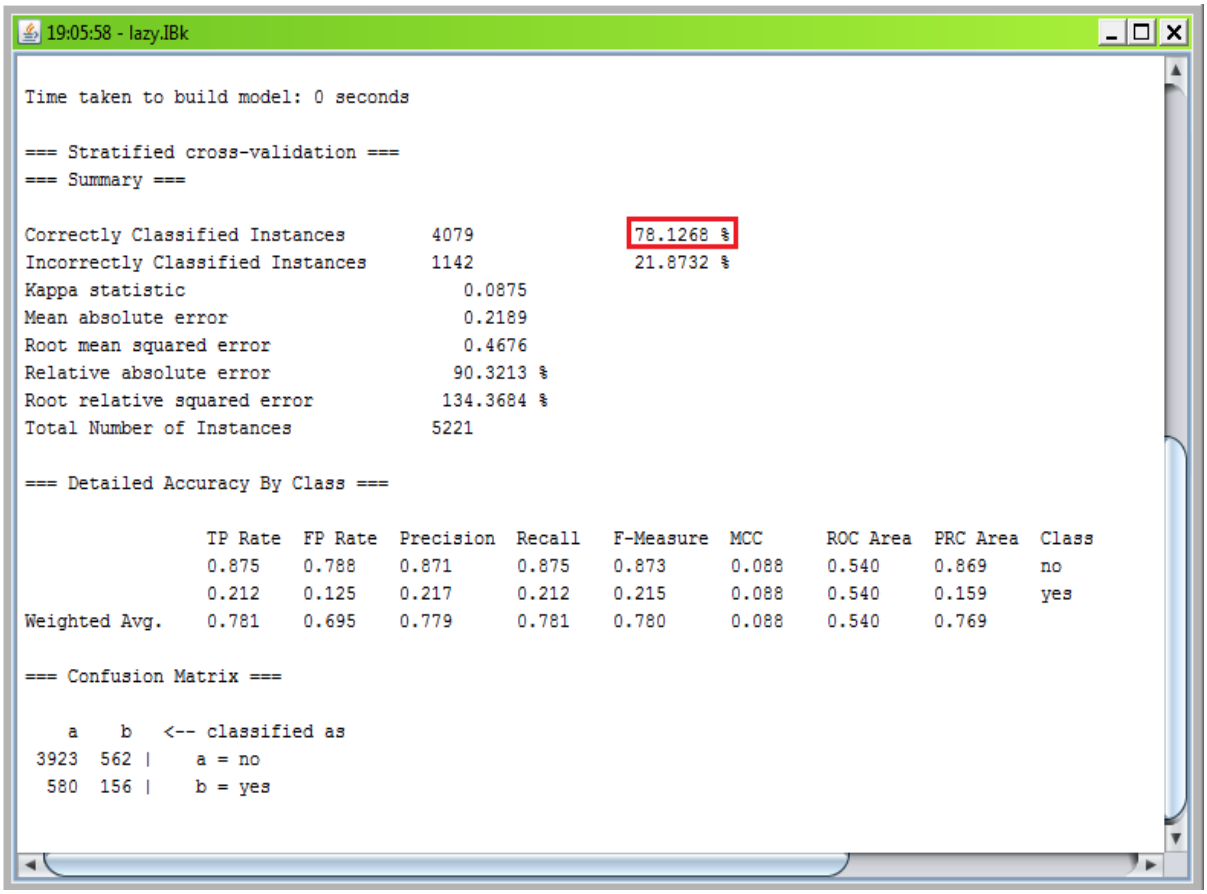

Figure 12: k-nearest neighbors (IBK) Prediction Rate

In the above figure 12, prediction rate of k-NN (IBk) (Altman, 1992), is demonstrated using 10-fold cross validation. It is highlighted in the figure that percentage of correctly classified instances is 78.1268 and number of instances are 4079. Total no. of incorrectly classified instances is 1142 that account for 21.8732%.

4.2 Tables of the Classifier Accuracy Summary

Table 1: Classifiers Prediction Rate Summary

Classifiers' Prediction Rates (10-Fold Cross-Validation) Summary					
	Descision Table	Decision Tree(J48)	Random Forest	Bayesian Network	kNN (IBk)
CCI	4471	4454	4469	4175	4079
CCI (%)	85.6349	85.3093	85.5966	79.9655	78.1268
ICI	750	767	752	1046	1142
ICI (%)	14.3651	14.6907	14.4034	20.0345	21.8732
KS	0.0496	0.0932	0.074	0.2353	0.0875
MAE	0.2389	0.2266	0.2216	0.2364	0.2189
RMSE	0.3447	0.3528	0.338	0.3717	0.4676
RAE (%)	98.6044	93.5378	91.4655	97.5789	90.3213
RRSE (%)	99.0535	101.3752	97.1266	106.8091	134.3684
CCI - Correctly Classified Instances, ICI - Incorrectly Classified Instances, KS - Kappa Statistics, MAE - Mean Absolute Error, RMSE - Root Mean Squared Error, RAE - Relative Absolute Error, RRSE - Root Relative Squared Error					

In the above table 1, prediction rate summary of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using 10-fold cross validation. Various parameters are considered for summary namely, CCI (Correctly Classified Instances), ICI (Incorrectly Classified Instances), KS (Kappa Statistics), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), RAE (Relative Absolute Error), RRSE (Root Relative Squared Error).

Table 2: Classifiers Accuracy by 'NO' Class

Classifiers' Prediction Rates (10-Fold Cross-Validation) Accuracy by 'NO' Class					
	Descision Table	Decision Tree(J48)	Random Forest	Bayesian Network	kNN (IBk)
TP Rate	0.99	0.979	0.986	0.867	0.875
FP Rate	0.959	0.917	0.939	0.613	0.788
Precision	0.863	0.867	0.865	0.896	0.871
Recall	0.99	0.979	0.986	0.867	0.875
F-Measure	0.922	0.92	0.922	0.881	0.873
MCC	0.091	0.129	0.117	0.237	0.088
ROC Area	0.654	0.592	0.709	0.73	0.54
PRC Area	0.914	0.872	0.931	0.938	0.869

In the above table 2, accuracy of classifier algorithms by “NO” class is demonstrated using 10-fold cross validation. Various parameters are considered for it namely, TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area and PRC area.

Table 3: Classifiers Accuracy by 'YES' Class

Classifiers' Prediction Rates (10-Fold Cross-Validation) Accuracy by 'YES' Class					
	Descision Table	Decision Tree(J48)	Random Forest	Bayesian Network	kNN (IBk)
TP Rate	0.041	0.083	0.061	0.387	0.212
FP Rate	0.01	0.021	0.014	0.133	0.125
Precision	0.405	0.399	0.425	0.324	0.217
Recall	0.041	0.083	0.061	0.387	0.212
F-Measure	0.074	0.137	0.107	0.353	0.215
MCC	0.091	0.129	0.117	0.237	0.088
ROC Area	0.654	0.592	0.709	0.73	0.54
PRC Area	0.224	0.225	0.28	0.308	0.159

In the above table 3, accuracy of classifier algorithms by “NO” class is demonstrated using 10-fold cross validation. Various parameters are considered for it namely, TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area and PRC area.

4.3 Classifiers' Prediction Rates using a 10-Fold Cross-Validation Graphs

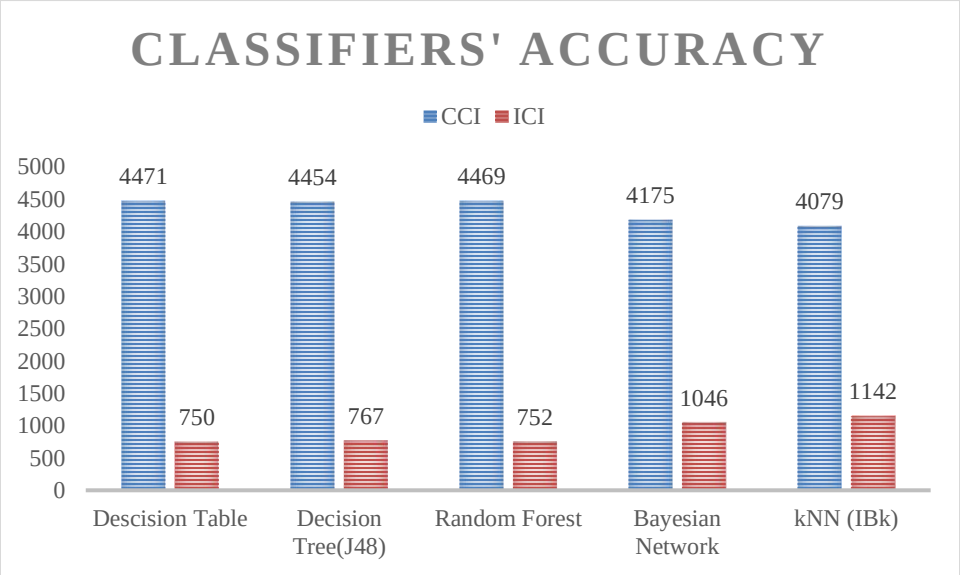

Figure 13: Classifiers' Accuracy (Instances) Graph

In the above figure 13, classification accuracy of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using 10-fold cross validation. It was found that Decision Table has least no. of ICI (Incorrectly Classified Instances) i.e. 750 and kNN has maximum no. of ICI i.e. 1142.

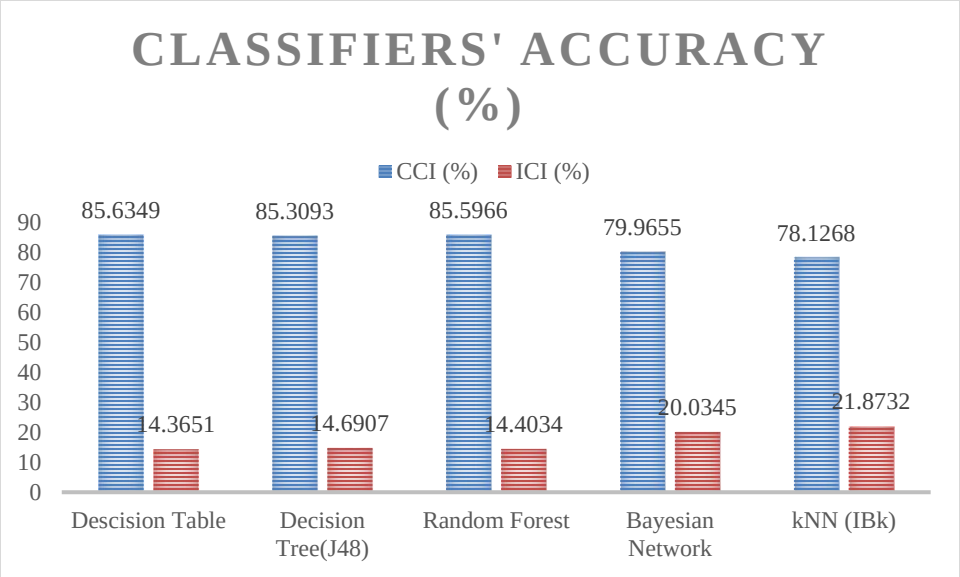

Figure 14: Classifiers' Accuracy (%) Graph

In the above figure 14, classification accuracy (%) of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using 10-fold cross validation. It was found that Decision Table has least no. of ICI (Incorrectly Classified Instances) i.e. 14.3651% and kNN has maximum no. percentage of ICI i.e. 21.8732.

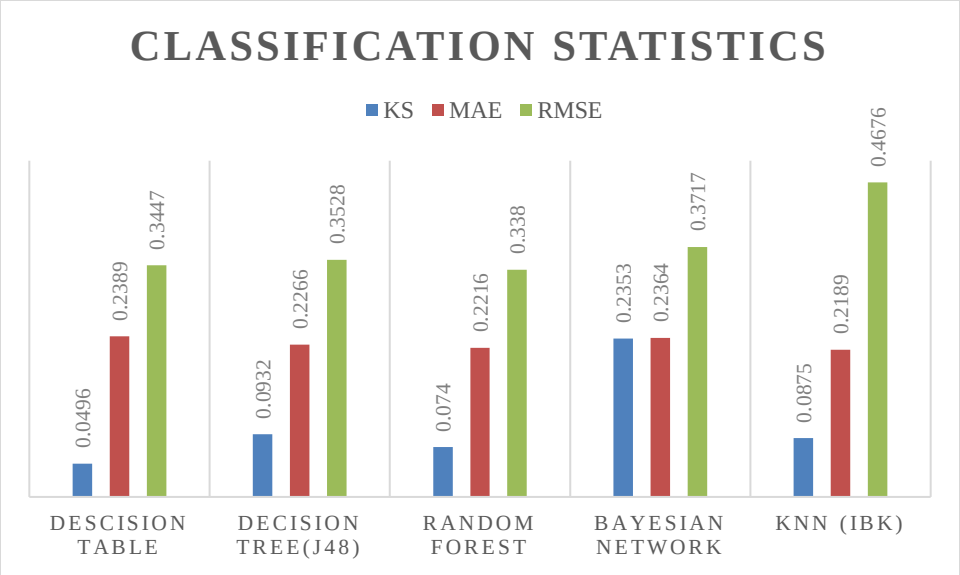

Figure 15: Classification Statistics Graph

In the above figure 15, classification statistics (KS - Kappa Statistics, MAE -Mean Absolute Error, RMSE - Root Mean Squared Error) of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using 10-fold cross validation. It was found that Decision Table has minimum KS i.e. 0.0496, KNN (IBk) (Altman, 1992), has minimum MAE i.e. 0.2189 and Random forest (Breiman, 2001) has minimum RMSE i.e. 0.338.

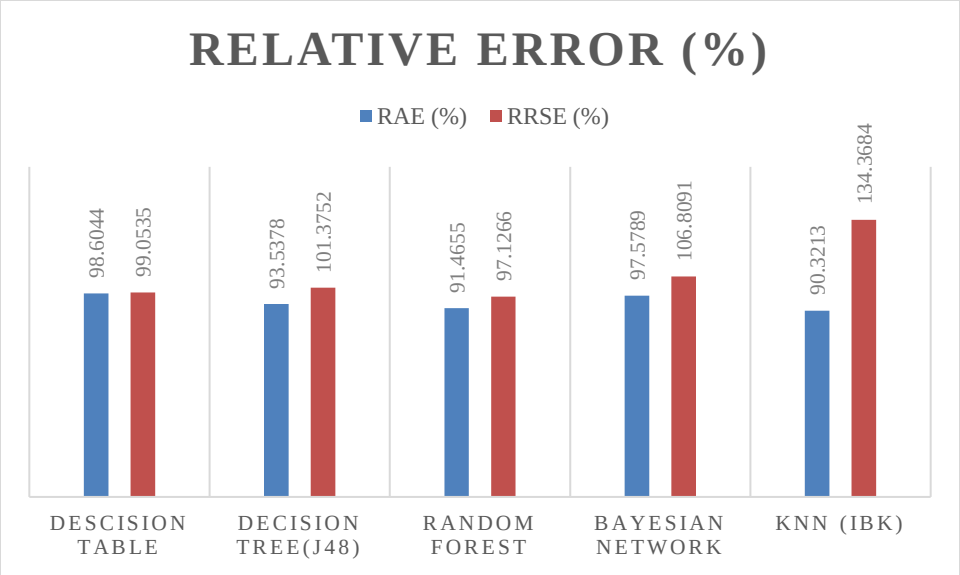

Figure 16: Relative Error (%) Graph

In the above figure 16, percentage of relative error (RAE - Relative Absolute Error, RRSE - Root Mean Squared Error) of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using 10-fold cross validation. It was found that kNN (IBk) (Altman, 1992), has minimum RAE i.e. 90.3213 and Random Forest (Breiman, 2001) has minimum RRSE i.e. 97.1266.

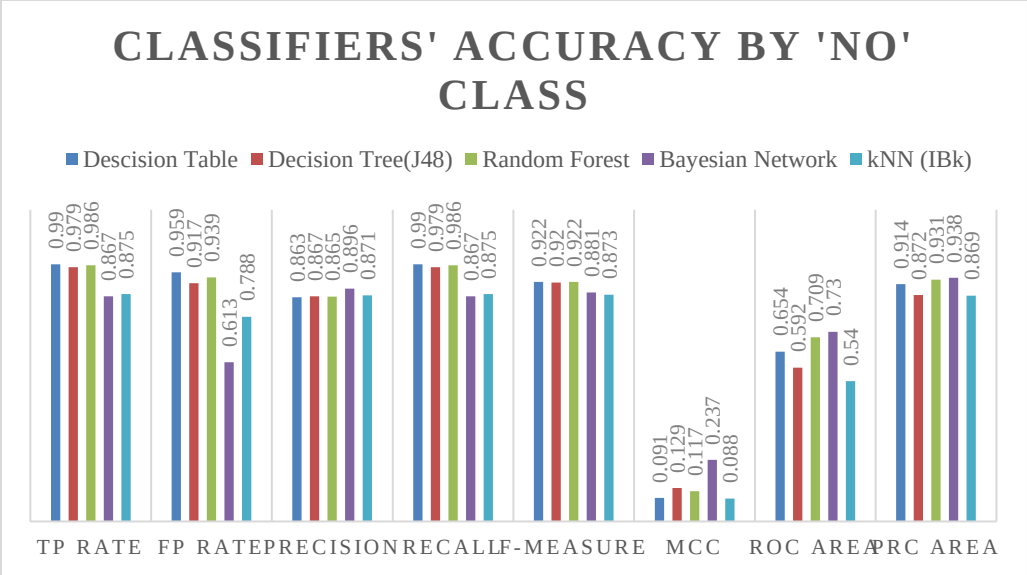

Figure 17: Classifiers' Accuracy by 'NO' Class Graph

In the above figure 17, accuracy of classifier algorithms by “NO” class is demonstrated using 10-fold cross validation. Various parameters are considered for it namely, TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area and PRC area.

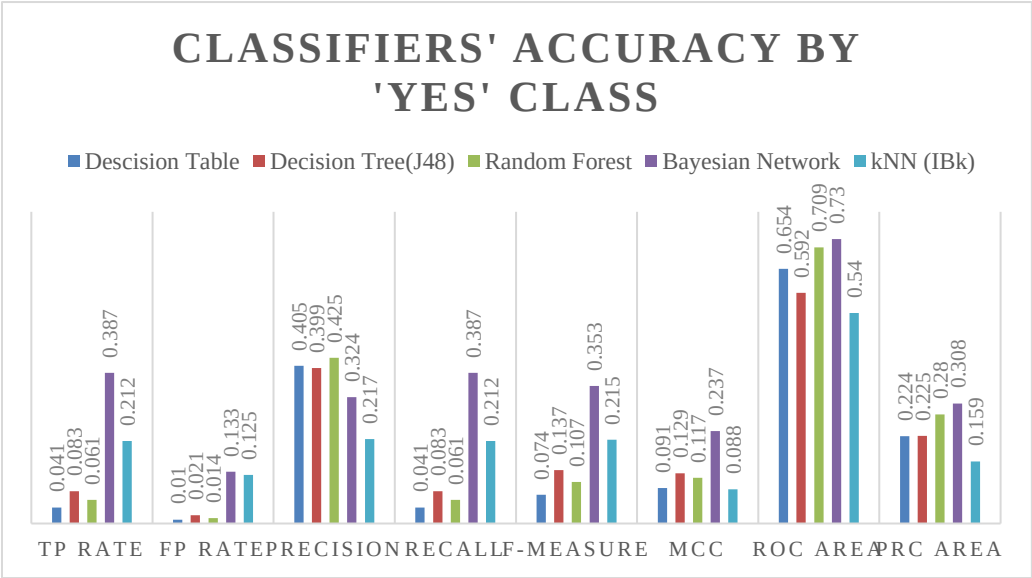

Figure 18: Classifiers' Accuracy by 'YES' Class Graph

In the above figure 18, accuracy of classifier algorithms by “YES” class is demonstrated using 10-fold cross validation. Various parameters are considered for it namely, TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area and PRC area.

4.4 Test Set Evaluation Summary

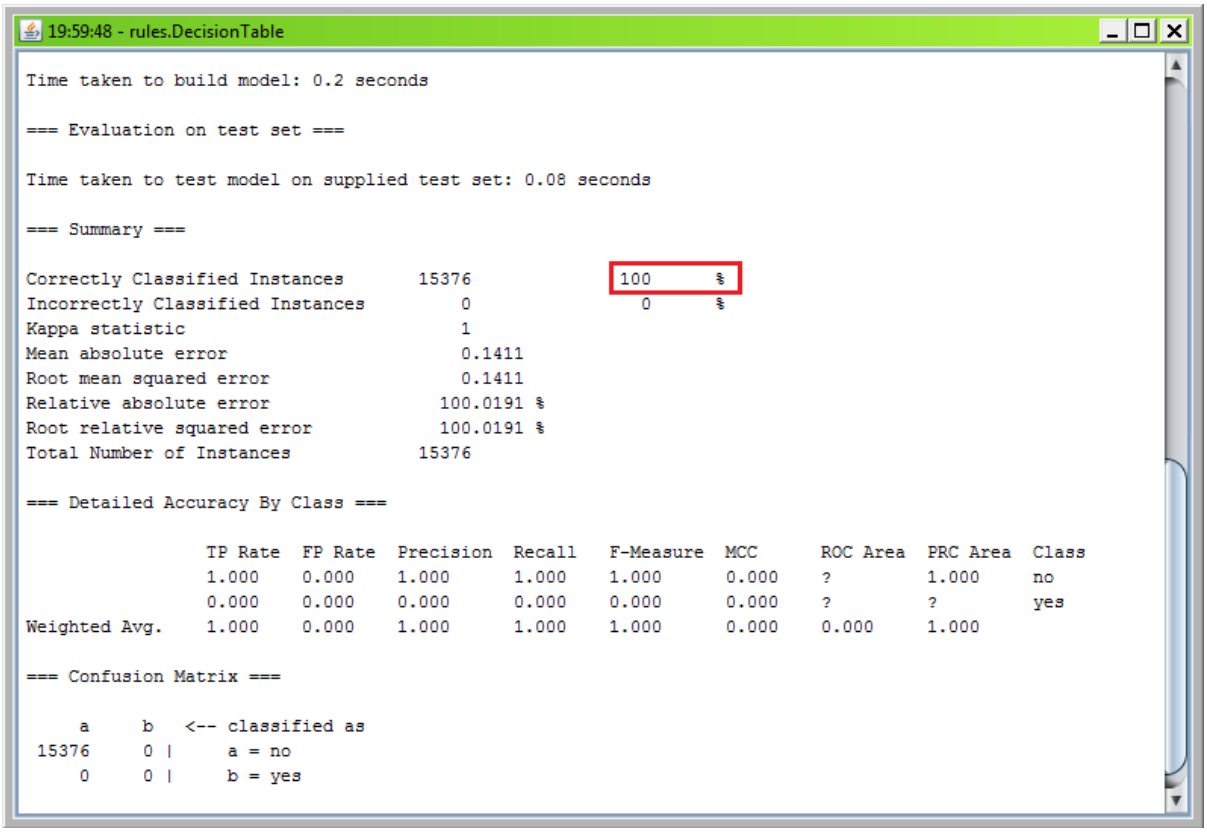

Figure 19: Decision Table Test Set Evaluation Summary

In the above figure 19, Test set evaluation summary of Decision Table is demonstrated. It is highlighted in the figure that percentage of correctly classified instances is 100 and number of instances are 15376. Total no. of incorrectly classified instances is 0 that account for 0%. This indicates that Decision Table is best in class for this experiment.

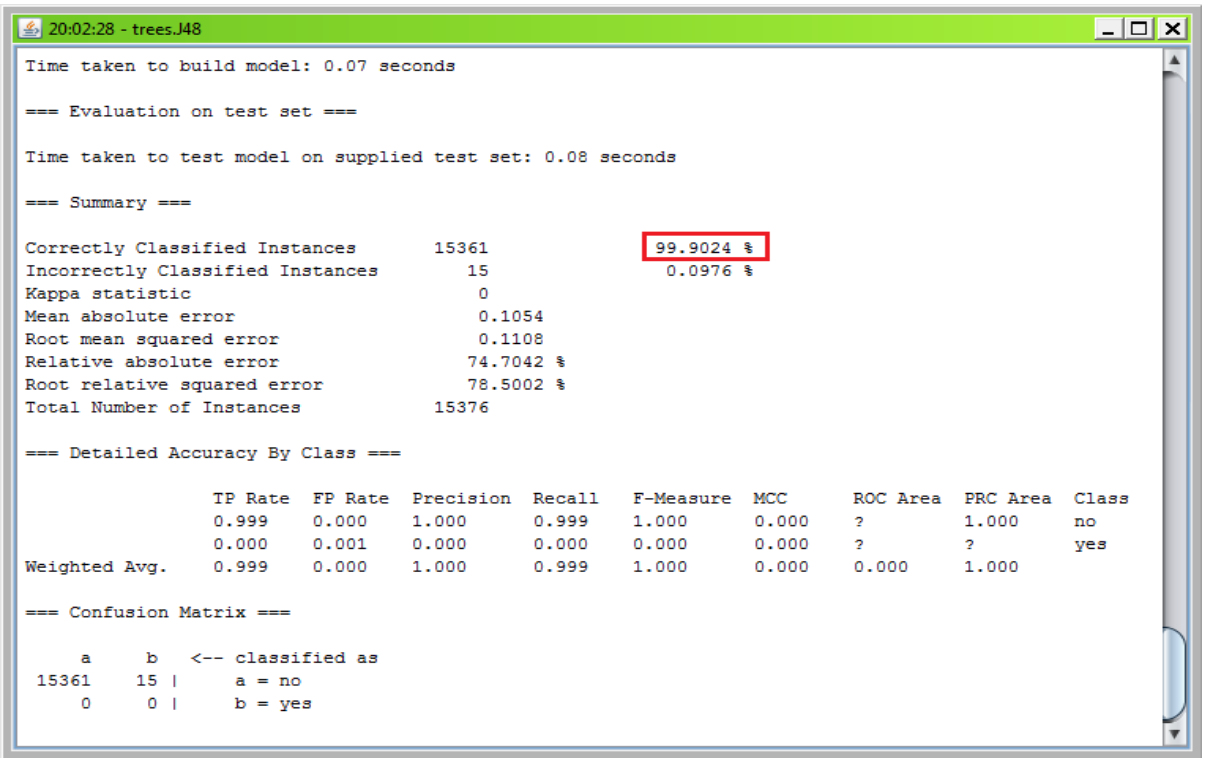

Figure 20: decision Tree (J48) Test Set Evaluation Summary

In the above figure 20, Test set evaluation summary of Decision Tree (J48) (Quinlan, 1993) is demonstrated. It is highlighted in the figure that percentage of correctly classified instances is 99.9024 and number of instances are 15361. Total no. of incorrectly classified instances is 15 that account for 0.0976%.

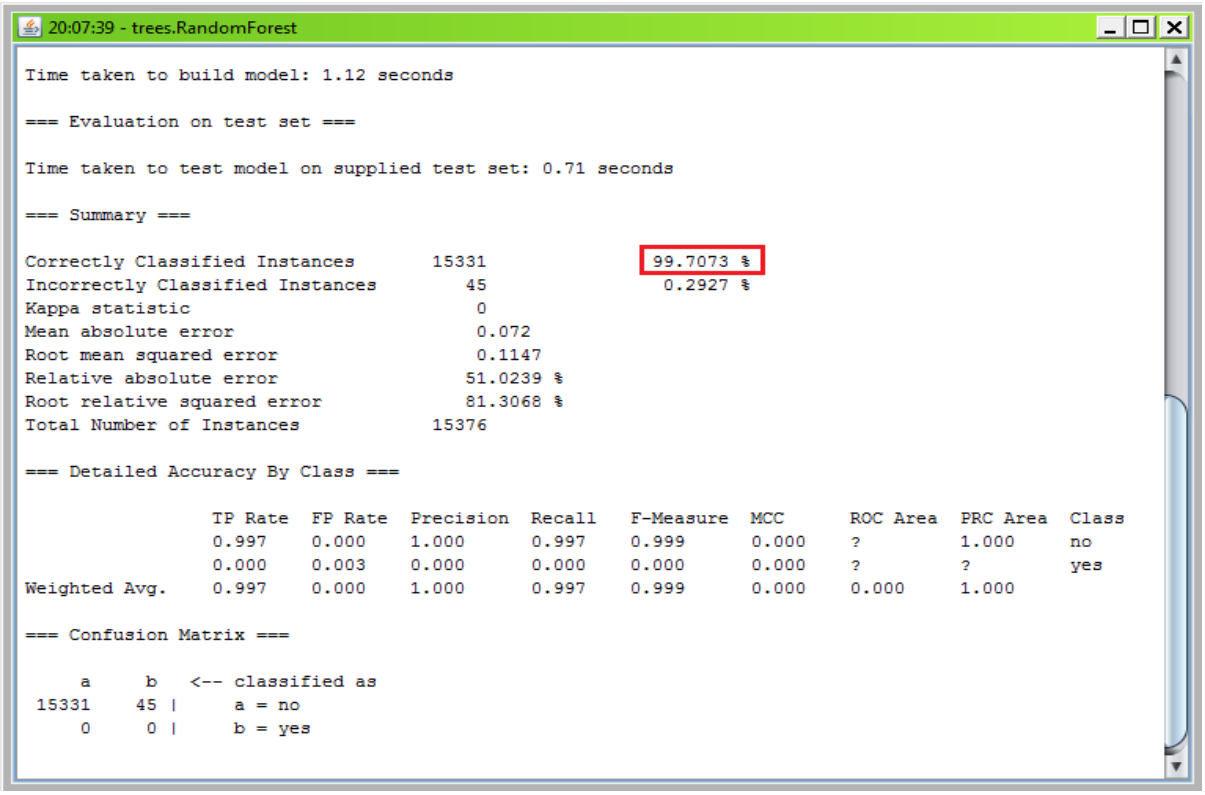

Figure 21: Random Forest Test Set Evaluation Summary

In the above figure 21, Test set evaluation summary of Random Forest (Breiman, 2001) is demonstrated. It is highlighted in the figure that percentage of correctly classified instances is 99.7073 and number of instances are 15331. Total no. of incorrectly classified instances is 45 that account for 0.2927%.

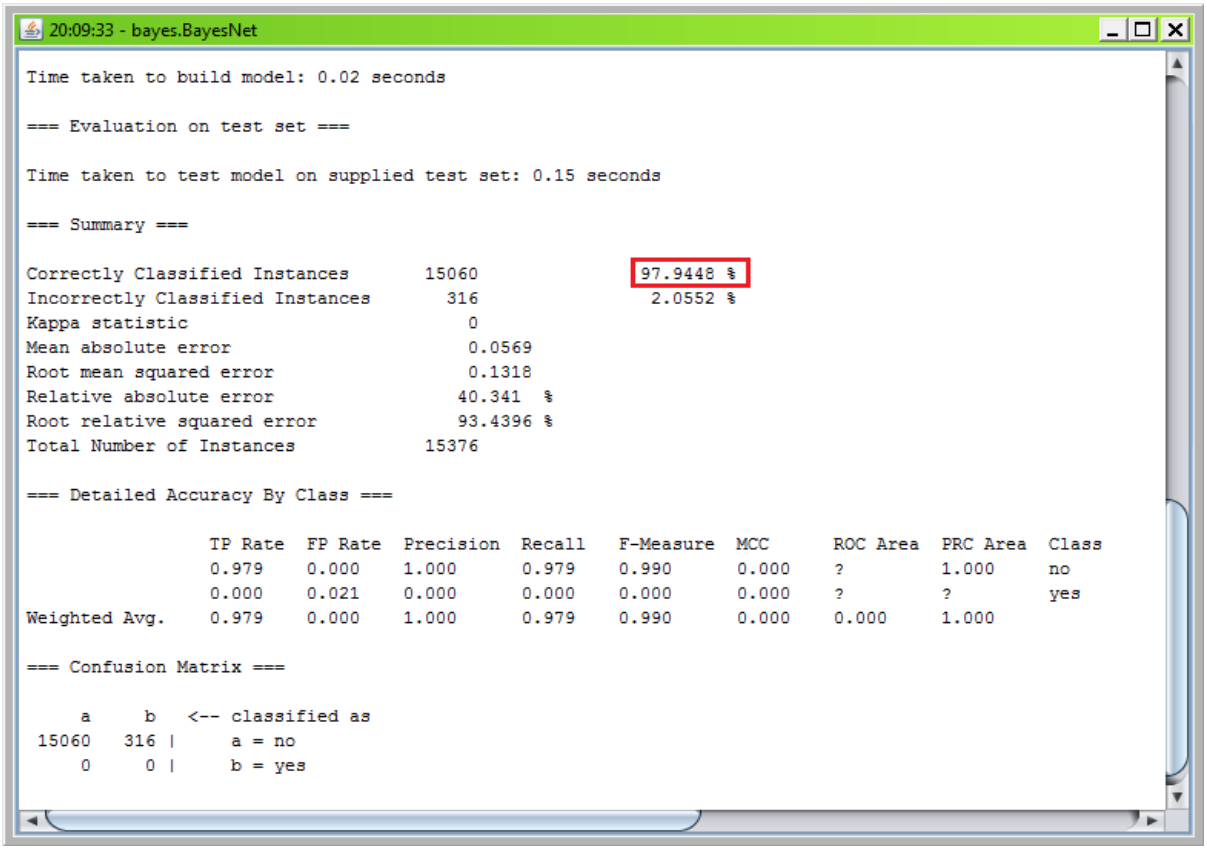

Figure 22: Bayesian Network Test Set Evaluation Summary

In the above figure 22, Test set evaluation summary of Bayesian Network (Bayes and Price 1763) is demonstrated. It is highlighted in the figure that percentage of correctly classified instances is 97.9448 and number of instances are 15060. Total no. of incorrectly classified instances is 316 that account for 2.0552%.

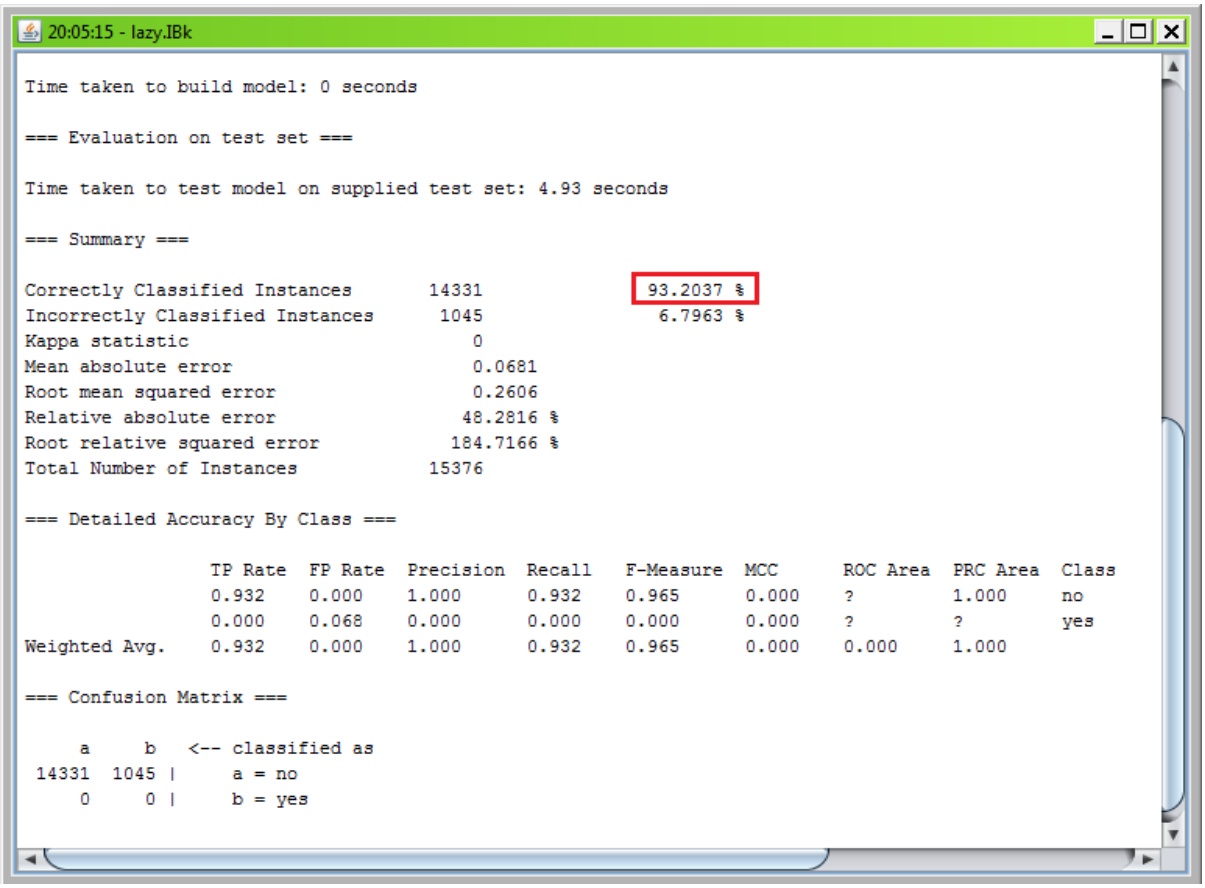

Figure 23: IBK Test Set Evaluation Summary

In the above figure 23, Test set evaluation summary of IBk is demonstrated. It is highlighted in the figure that percentage of correctly classified instances is 93.2037 and number of instances are 14331. Total no. of incorrectly classified instances is 1045 that account for 6.7963%.

4.5 Table of Test Set Evaluation Summary

Table 4: Test Set Evaluation Summary

Test Set Evaluation Summary					
	Descision Table	Decision Tree(J48)	Random Forest	Bayesian Network	kNN (IBk)
CCI	15376	15361	15331	15060	14331
CCI (%)	100	99.9024	99.7073	97.9448	93.2037
ICI	0	15	45	316	1045
ICI (%)	0	0.0976	0.2927	2.0552	6.7963
KS	1	0	0	0	0
MAE	0.1411	0.1054	0.072	0.0569	0.0681
RMSE	0.1411	0.1108	0.1147	0.1318	0.2606
RAE (%)	100.0191	74.7042	51.0239	40.341	48.2816
RRSE (%)	100.0191	78.5002	81.3068	93.4396	184.7166
CCI - Correctly Classified Instances, ICI - Incorrectly Classified Instances, KS - Kappa Statistics, MAE - Mean Absolute Error, RMSE - Root Mean Squared Error, RAE - Relative Absolute Error, RRSE - Root Relative Squared Error					

In the above table 4, test set evaluation summary of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated. Various parameters are considered for summary namely, CCI (Correctly Classified Instances), ICI (Incorrectly Classified Instances), KS (Kappa Statistics), MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), RAE (Relative Absolute Error), RRSE (Root Relative Squared Error).

4.6 Test Set Evaluation Summary Graphs

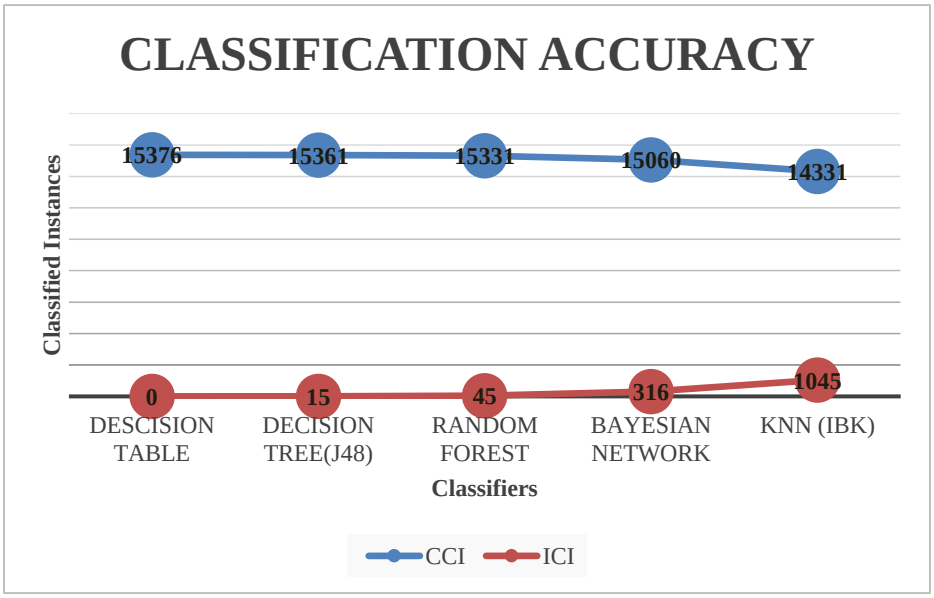

Figure 24: Test Set Classification Accuracy Graph

In the above figure 24, classification accuracy of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using test set evaluation. It was found that Decision Table has least no. of ICI (Incorrectly Classified Instances) i.e. 0 and kNN has maximum no. of ICI i.e. 1045.

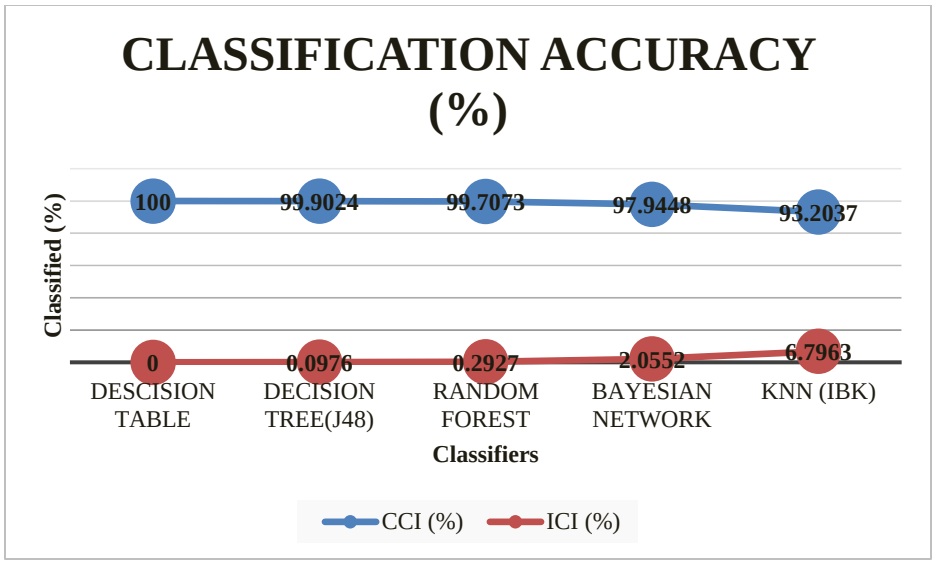

Figure 25: Test Set Classification Accuracy (%) Graph

In the above figure 25, classification accuracy percentage of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using test set evaluation. It was found that Decision Table has least no. of ICI (Incorrectly Classified Instances) i.e. 0% and kNN has maximum percentage of ICI i.e. 6.7963.

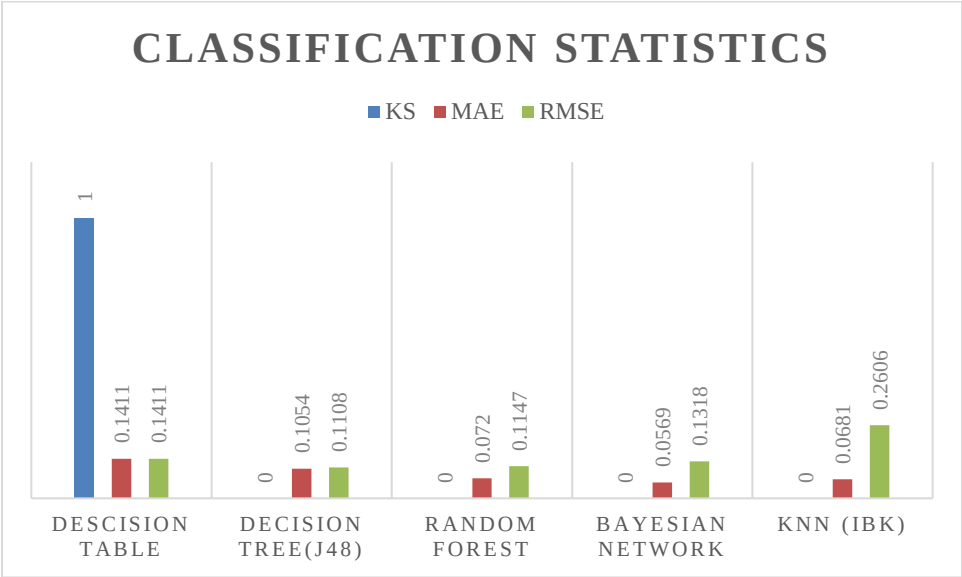

Figure 26: Test Set Classification Statistics Graph

In the above figure 26, classification statistics (KS - Kappa Statistics, MAE -Mean Absolute Error, RMSE - Root Mean Squared Error) of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using test set evaluation. It was found that Decision Table has maximum KS i.e. 1, Bayesian network (Bayes and Price 1763) has minimum MAE i.e. 0.0569 and decision tree (J48) (Quinlan, 1993) has minimum RMSE i.e. 0.1108.

.

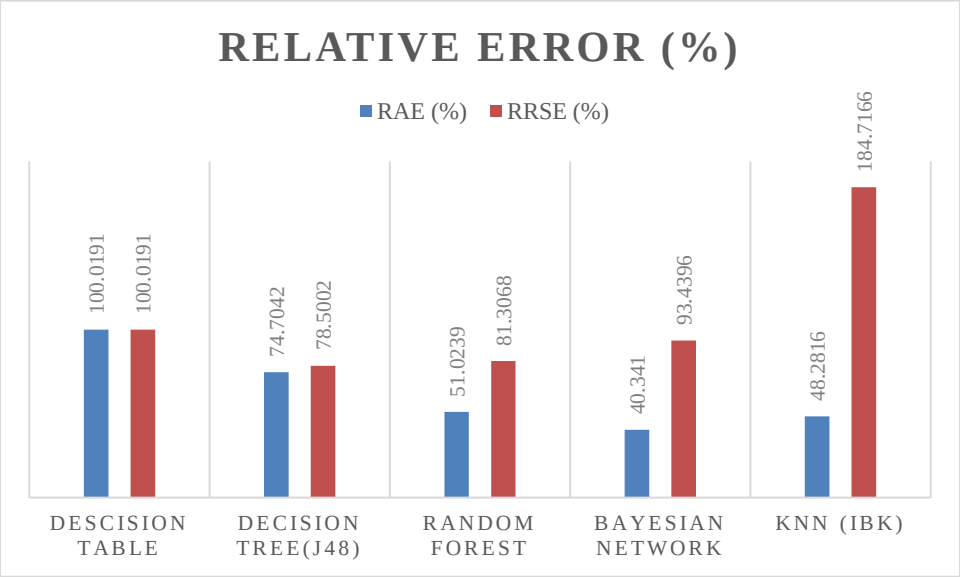

Figure 27: Test Set Classification Relative Error (%) Graph

In the above figure 27, percentage of relative error (RAE - Relative Absolute Error, RRSE - Root Mean Squared Error) of various classifier algorithms namely, decision table, decision tree, Random Forest (Breiman, 2001), Bayesian network (Bayes and Price 1763) and IBk is demonstrated using test set evaluation. It was found that Bayesian Network (Bayes and Price 1763) has minimum RAE i.e. 40.341 and Decision Tree (J48) (Quinlan, 1993) has minimum RRSE i.e. 78.5002.

4.7 Findings

After analyzing the classification accuracy and other statistics of various classifiers namely, Decision Tree (J48) (Quinlan, 1993), Decision Tables, k-nearest neighbors (IBK), Random Forest (Breiman, 2001) and Bayesian Network (Bayes and Price 1763), I found kNN (IBK) to be most helpful in identifying loyal customers to Viata. kNN has maximum no. ICI (Incorrectly Classified Instances) i.e. 1045 out of 15376 instances on Test Set Evaluation Summary. It illustrates, that there is a scope for the company to do target marketing on 1045 loyal customers from one-order test dataset.

Chapter 5 Conclusions

E-commerce is on boom these days and online pharma companies are not exception in any ways. Viata is one of the major online pharmacy and drug store based in Europe. This thesis completely deals with RFM (Recency, Frequency and Monetary) analysis, done for Viata Company. RFM has been generally implemented in numerous aspects of customer relationship management, especially in direct marketing and advertising. The major goal of this research was to analyze the loyalty of customers who transacted through Viata portal. Throughout the research, WEKA tool was used for the analysis which comes with ready in-built integration of most of the data mining algorithms. Consumers' dataset with multiple transactions was used as training set to classify and test single order customer's loyalty. Various analyses were done to find the performance of different data mining algorithms namely, Decision Tree (J48) (Quinlan, 1993), Decision Tables, k-nearest neighbors (IBK), Random Forest (Breiman, 2001) and Bayesian Network (Bayes and Price 1763). Also, results were plotted in graphical format for easy understanding. K-NN, to its best, offered 6.7963% incorrectly classified instances i.e. found 1045 customers to be loyal to the company. This analysis is helpful in segmentation of loyal and non-loyal customers and to build a strategy for marketing and advertisements considering various sets of consumers depending upon their loyalty. For future works, email campaigning and targeted advertisements may appear into scene.

References

1. Adobe. (2015, 10). Summary of Key Findings. Retrieved from adobe.com: http://wwwimages.adobe.com/content/dam/Adobe/en/max/2015/pdfs/state-of-content-oct.pdf.
2. Aggelis V., Christodoulakis D. (2004). Customer clustering using RFM analysis. Proceedings of the 9th WSEAS International Conference on Computers. Wisconsin, USA: World Scentific and Engineering Academy and Society.
3. Allen Robert. (2016, 03 18). Ecommerce trends to watch 2016. Retrieved from smartinsights.com: http://www.smartinsights.com/ecommerce/ecommerce-strategy/ecommerce-trends-watch-2016-infographic/
4. Allen, R. (2016, 11 24). 38 Indispensable E-commerce stats to inform your 2017 multichannel sales strategy. Retrieved from smartinsights.com: http://www.smartinsights.com/ecommerce/ecommerce-strategy/37-indispensable-ecommerce-stats-to-inform-your-2017-strategy/
5. Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.
6. Baesens B., Verstraeten G., Dirk V. D. P., Michael E. P., Kenhove, V. K. and Vanthienen J. (2004), “Bayesian network classifiers for identifying the slope of the customer-lifecycle of long-life customers”, European Journal of Operational Research,Vol.156, pp.508-523.
7. Barbieri M.M, and Berger.O.J. (2004), “Optimal predictive model selection”,Annals of Statistics,Vol.32, No.3, pp.870-897.
8. Bayes, T.; Price, Mr. (1763). "An Essay towards solving a Problem in the Doctrine of Chances". Philosophical Transactions of the Royal Society. 53: 370–418.
9. Blattberg R., Buesing T., Peacock P. and Sen S. (1978), “Identifying deal prone Segment”, Journal of Marketing Research, Vol.15, No.3, pp.369-377.
10. Bortiz J. E. and Kennedy D.B. (1995), “Effectiveness of neural network types for prediction of business failure”, Expert Systems with Applications, Vol. 9, pp.503-512.
11. Bounsaythip, Catherine and Esa Rinta-Runsala (2001), “Overview of data mining for customer behavior modeling”, VTT Information Technology, Vol.18, pp.1-53.
12. Breiman, Leo (2001). "Random Forests". Machine Learning. 45 (1): 5–32.

13. Buras, M. (2016, 05 26). Custora E-Commerce Pulse Report: Q1 2016. Retrieved from custora.com: http://blog.custora.com/2016/05/custora-e-commerce-pulse-report-q1-2016/
14. Chellappa R. and Sin R. (2005). Personalization versus privacy: an empirical examination of the online consumer's dilemma. Information Technology Management, 06, 181-202.
15. Cheng CH., Chen YS. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert System Applications, 36, 4176-4184.
16. Chu Chai Henry Chan (2008), "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", Expert Systems with Applications,Vol.34, pp.2754-2762.
17. Cooper Smith. (2014, 06 12). The Surprising Demographics Of Who Shops Online And On Mobile. Retrieved from businessinsider.in: http://www.businessinsider.in/The-Surprising-Demographics-Of-Who-Shops-Online-And-On-Mobile/articleshow/36449798.cms
18. Cunningham C., Song Il-Yeol. Chen, Peter P. (2004). Data warehouse design to support customer relationship management analysis. Proceedings of the 7th ACM international workshop on Data warehousing and OLAP, (pp. 14-22).
19. David Hand, Heikki Mannila and Padhraic Smyth (2008), "Principles of data mining", PHI Learning Privated Limited,Third Edition.
20. Dibb S. and Simkin L. (1997), "A program for implementing market segmentation", Journal of Business and Industrial Marketing, Vol.12, No.1, pp.51-65.
21. Dilbag singh,Pradeep Kumar (2012), "Conceptual Mapping of Insurance Risk Management to Data Mining", International Journal of Computer Applications, Vol. 39, No.2, pp.13-18.
22. Dunham (2003), "Data Mining Introductory and Advanced Topics", Prentice Hall, 2003,First edition.
23. Ecommerce Foundation . (2016). Ecommerce Facts & Figures of Belgium. Retrieved from ecommercewiki.org: https://www.ecommercewiki.org/Global_Ecommerce_Figures/Europe/Belgium
24. Ecommerce News. (2017, 03 14). Ecommerce in Belgium: €9.1 billion in 2016. Retrieved from ecommercenews.eu: https://ecommercenews.eu/ecommerce-belgium-e91-billion-2016/

25. eMarketer. (2016, 08 22). Worldwide Retail Ecommerce Sales Will Reach $1.915 Trillion This Year. Retrieved from emarketer.com: https://www.emarketer.com/Article/Worldwide-Retail-Ecommerce-Sales-Will-Reach-1915-trillion-This-Year/1014369
26. Eriksson, K. & Vaghult, A. (2000). Customer Retention, Purchasing Behavior and Relationship Substance in Professional Services. Industrial Marketing Management, 29(4), 363-372.
27. Eurostat Information. (2016, 12). E-commerce statistics for individuals. Retrieved from ec.europa.eu: http://ec.europa.eu/eurostat/statistics-explained/index.php?title=E-commerce_statistics_for_individuals&oldid=266007#30.C2.A0.25_of_online_shoppers_bought_or_ordered_goods_or_services_from_sellers_in_other_EU_countries
28. Fayyad U., Piatetsky-Shapiro G., Smyth P. and Uthurusamy R. (1996), "Advances in Knowledge Discovery and Data Mining", First Edition, AAAI/MIT Press.
29. Fletcher D. and Goss E. (1993), "Forecasting with neural networks: An application using bankruptcy data. Information and Management", Vol.3, pp.205-213.
30. Gao M, Liu K and Wu Z. (2010). Personalization in web computing and informatics: theories, techniques, applications, and future research. Information System Front, 12, 607-629.
31. Gemma. (2016, 06 07). E-commerce in Europe 2016: Facts & Figures. Retrieved from twenga-solutions.com: https://www.twenga-solutions.com/en/insights/e-commerce-europe-2016-facts-figures/
32. Green P.E. and Wind Y. (1973), "Multiattribute decisions in marketing", Dryden Press, First Edition.
33. Greengrove and Kathryn (2002), "Needs-based segmentation: principles and practice",USA,International Journal of Market Research,Vol.44, pp.405-421.
34. Guha S., Rastogi R. and Shim K. (1998), "CURE: An efficient clustering algorithm for large databases", In Proceeding of the 1998, ACM SIGMOD International Conference on Management of Data Engineering, Seattle, WA USA, pp. 512-521.
35. Gummesson, E. (1987). The New Marketing - Developing Long term Interactive Relationships. Long Range Planning, 20(4), 10-20.

36. H.M. Chuang, C.C. Shen. (2008). A study on the application of data mining techniques to enhance customer lifetime value based on the department store industry. Proceedings of 7th International Conference on Machine Learning and Cybernetics, (pp. 168-173). California, USA.
37. Helsen K. and Green P.E. (1991), "A computational study of replicated clustering with an application to market-segmentation", Decision Sciences, Vol.22, No.5, pp.1124-1141.
38. Hinz O and Eckert J. (2010). "The impact of search and recommendation systems on sales in electronic commerce. J. Bus. Inf. Syst. Eng, 02, 67-77.
39. Holland S. and Kießling W. (2004). User preference mining techniques for personalized applications. Wirtschaftsinf, 46, 439-445.
40. Horng-Jinh Chang, Lun-Ping Hung and Chia-Ling Ho (2007), "An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis", Expert Systems with Applications,Vol. 32, pp.753-764.
41. Hosseini M., Anahita M., Mohammad R. G. (2010), "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty", Expert Systems with Applications,Vol. 37, pp.5259-5264.
42. Hruschka H. (1996), "Market definition and segmentation using fuzzy clustering methods", International Journal of Research in Marketing, Vol. 3, No.2, pp.117-134.
43. Huang S.C., Chang, E.C. and Wu H.H. (2009), "A case study of applying data mining techniques in an outfitter's customer value analysis", Expert System with Applications, Vol.36, No.6, pp.5909-5915.
44. Hughes, A.M. (1996). Boosting Response with RFM. Marketing Tools, 3(3), 4-5.
45. Hwang H., Jung T. and Suh E. (2004), "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry", Expert Systems with Applications, Vol. 26, No.2, pp.181-188.
46. Irvin, S. (1994). Using lifetime value analysis for selecting new customers. Credit World, 82(3), 37-40.
47. James Dudley. (2012). Can e-commerce in Healthcare Succeed in Europe? (J. D.-M. Europe, Editor) Retrieved from james-dudley.co.uk: http://james-dudley.co.uk/news/press-releases/can-e-commerce-in-healthcare-succeed-in-europe/

48. Jayanthi Ranjan, Raghuvir Singh and Vishal Bhatnagar (2011), “Analytical customer relationship management in insurance industry using data mining: a case study of Indian insurance company”, International Journal of Networking and Virtual Organisations,Vol. 9, No. 4, pp. 331-366 .
49. Jibendu Kumar Mantri, (2008). “Research methodology on data envelopment analysis (4 ed.). Universal-Publishers.
50. Kamber M and Han J. (2008), “Data Mining:Concepts and Techniques”, Second Edition, Morgan Kaufmann.
51. Keyvan Vahidy Rodpysh (2012), “Model to Predict the Behavior of Customers Churn at the Industry”, International Journal of Computer Applications,Vol.49, No.15, pp.12-16.
52. Kiang M.Y., Hu M.Y. and Fisher D. M. (2006), “An extended self-organizing map network for market segmentation-a telecommunication example”, Decision Support Systems,Vol.42, No.1, pp.36-47.
53. Kim J., Suh E. and Hwang, H. (2003), “A model for evaluating the effectiveness of CRM using the balanced scorecard”, Journal of Interactive Marketing, Vol. 17, No.2, pp. 5-19.
54. Kim, Yong Seog, & Street, W. Nick (2004). An intelligent system for customer targeting: A data mining approach. Decision Support Systems, 37, 215–228
55. Kohonen, Teuvo (1982). "Self-Organized Formation of Topologically Correct Feature Maps". Biological Cybernetics. 43 (1): 59–69.
56. Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. Expert Systems with Applications, 30 (2), 313–324.
57. Langley P. and Simon H. A. (1995), “Applications of machine learning and rule Induction”,Communication of the ACM, Vol.38, pp.55-64.
58. Lau H.C.W., Wong, C.W.Y., Hui, I.K. and Pun K.F. (2003), “Design and implementation of an integrated knowledge system”, Journal of KnowledgeBased Systems,Vol.16, pp.69-76.
59. Liu DR, Shih YY. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information Management, 42, 387-400.
60. Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28, 129–137.

61. Marcus C. (1998). A practical yet meaningful approach to customer segmentation. Journal of Consumer Market, 15(5), 494-504.
62. McCarty, JA, Hastak M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of Business Research, 60, 656-662.
63. MichaelLazar. (2017, 03 04). E-commerce Statistics & Technology Trendsetters for 2017. Retrieved from ibm.com: https://www.ibm.com/developerworks/community/blogs/d27b1c65-986e-4a4f-a491-5e8eb23980be/entry/Ecommerce_Statistics_Technology_Trendsetters_for_20171
64. Miglautsch, J. R. (2000). Thoughts on RFM Scoring. International Society for Strategic Marketing. The Journal of Database Marketing, 8(1), 67–72.
65. Miguéis V.L., Camanho A.S., João Falcão e Cunha (2012), "Customer data mining for lifestyle segmentation", Expert Systems with Applications, Vol.39, pp. 9359-9366.
66. Min S. and Han I. (2005), "Detection of the customer time-variant pattern for improving recommender systems", Expert Systems with Applications, Vol.28, No.2, pp.189-199
67. Miniwatts Marketing Group. (2017, 03 31). INTERNET USAGE STATISTICS. Retrieved from internetworldstats.com: http://www.internetworldstats.com/stats.htm
68. Mitchell, A. (1983), The Nine American Life Styles, Warner, New York, NY. [Google Scholar].
69. Mitra S., S.K. Pal and P. Mitra (2002). "Data mining in soft computing framework: A survey", IEEE. Trans. Neural Networks, Vol.13, pp.3-14.
70. Nan-Chen Hsieh and Kuo-Chung Chu (2009), "Enhancing Consumer Behavior Analysis by Data Mining Techniques", International Journal of Information and Management Sciences,Vol.20, pp.39-53.
71. Newell, F. (1997). The new rules of marketing: How to use one-to-one relationship marketing to be the leader in your industry. New York: McGraw-Hill.
72. Ozer M. (2001), "User segmentation of online music service using fuzzy clustering", Journal of cybernetics Omega,Vol.29, No.2, pp.193-206.
73. Parvatiyar, Atul and Jagdish N. Sheth (2001), "Customer relationshipmanagement: emerging practice, process, and discipline", Journal of Economic and Social Research, Vol.2, pp.1-34.

74. Pienaar, A. (2015, 04 22). 75 Ecommerce Facts, Quotes & Statistics. Retrieved from conversio.com: https://conversio.com/blog/75-ecommerce-facts-quotes-statistics-that-will-blow-your-mind/
75. Piercy N. and Morgan N. (1993), “Strategic and operational market segmentation: A managerial analysis”, Journal of Strategic Marketing, Vol.5, No.1, pp.123-140
76. Pramod Prasad and Latesh G. Malik (2011), “Generating customer profiles for retail stores using clustering techniques”, International Journal on Computer Science and Engineering, Vol.3, pp.2506-2510.
77. Quinlan, J. R., (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.
78. Rauyruen P. & Miller K. (2007). Relationship quality as a predictor of B2B customer loyalty. Journal Of Business Research, 60(1), 21-31.
79. Razieh Qiasi, Malihe Baqeri-Dehnavi, Behrooz Minaei-Bidgoli and Golriz Amooee (2012), “Developing a model for measuring customer’s loyalty and value with RFM technique and clustering algorithms”, The Journal of Mathematics and Computer Science, Vol.4, No.2, pp.172-181.
80. ReadyCloud (2016). study-finds-that-convenient-online-product-returns-improve-profits-exponentially. Retrieved from readycloud.com: https://www.readycloud.com/articles/study-finds-that-convenient-online-product-returns-improve-profits-exponentially
81. Reichheld F. & Sasser Jr. W. (1990). Zero defections: Quality comes to services. Harvard Business Review, 68, 105-111.
82. Reichheld FF. (1996). Learning from Customer Defections. (cover story). Harvard Business Review, 64(2), 56-69.
83. Reinartz W., Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. Journal of Marketing, 67(1), 77–99.
84. Rekha Bhowmik (2011), “Detecting auto insurance fraud by Data mining techniques”, Journal of Emerging Trends in Computing and Information Sciences, Vol.2, No.4, pp.156-162.

85. Repschlaeger J., Erek K. and Zarnekow R. (2013). Cloud computing adoption: an empirical study of customer preferences among start-up companies. Electronic Markets, 23, 115-148.
86. Retail & Ecommerce. (2017, 01 20). UK Ecommerce Sales Growth Healthy in 2016 . Retrieved from www.emarketer.com: http://www.emarketer.com/Article/UK-Ecommerce-Sales-Growth-Healthy-2016/1015080
87. retailresearch.org. (2017, 03 28). Online Retailing: Britain, Europe, US and Canada 2017. Retrieved from retailresearch.org: http://www.retailresearch.org/onlineretailing.php
88. Ruey-Shun Chen, Ruey-Chyi Wu and Chen J. Y. (2005), "Data Mining Application in Customer Relationship Management of Credit Card Business", Proceedings of the 29th Annual International Computer Software and Applications Conference, IEEE, Vol.2, pp.39-40.
89. Saleh, K. (2014, 01 10). The Importance of Providing a Great Customer Experience. Retrieved from invespcro.com: https://www.invespcro.com/blog/great-customer-experience/
90. Salesforce Research. (2016). State of Marketing. Retrieved from Salesforce.com: https://secure.sfdcstatic.com/assets/pdf/misc/state-of-marketing-report-2016.pdf
91. Samira Malekmohammadi Golsefid, Mehdi Ghazanfari and Somayeh Alizadeh (2007), "Customer Segmentation in Foreign Trade based on Clustering Algorithms", World Academy of Science, Engineering and Technology, Vol.28, pp. 405-411.
92. Scott Brinker. (2017, 06 07). Top E-commerce Trends to inform your 2017 marketing strategy. Retrieved from smartinsights.com: http://www.smartinsights.com/?attachment_id=78883
93. Shaw M.J., Subramaniam C., Tan G.W. and Welge M.E. (2001), "Knowledge management and data mining for marketing", Decision Support Systems, Vol.31, pp.127-137
94. Sheu J.J., Su Y.H. and Chu K.T. (2009), "Segmenting Online Game Customers - the Perspective of Experiential Marketing", Expert Systems with Applications, Vol.36, pp.8487-8495.
95. Shin HW and SY Sohn (2004). "Multi-Attribute Scoring Method for Mobile Telecommunication Subscribers." Expert Systems with Applications 26(3): 363-368

96. Siavash Emtiyaz and Mohammad Reza Keyvanpour (2011), “Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management”, International Journal of scientific and engineering research,Vol.3, No.3, pp.229-236.
97. Sohrabi B, Khanlari A. (2007). Customer lifetime value (CLV) measurement based on RFM model. Iranian Acc. Aud. Rev., 14(47), 7-20.
98. SS. Kadiyala & A. Srivastava. (2011). Data Mining For Customer Relationship Management. International Business & Economics Research Journal, 1(6), 66.
99. Su-Yeon Kim, Tae-Soo Jung, Eui-Ho Suh and Hyun-Sepk Hwang (2006), “Customer segmentation and strategy development based on customer lifetime value”, Expert systems with applications, pp. 101-107
100. Thearling K. (1999), “Data mining and CRM: zeroing in on your best customers”, DM Direct.
101. Trappey V. Charles, Amy J.C. Trappey, Ai-Che Chang and Ashley Y.L. Huang (2009), “The analysis of customer service choices and promotion preferences using hierarchical clustering”, Journal of the Chinese Institute of Industrial Engineers, Vol.5, pp.367-376.
102. Twedt, Dik Warren (1964), “How Important to Marketing Strategy Is the “Heavy User”? Journal of Marketing, 28 (January), 71-72.
103. Two Crows Corporation (1998), “Introduction to Data Mining and Knowledge Discovery”, Second Edition Patomac, MD
104. US Census Bureau News (2017, 05 16). US Census Bureau. Retrieved from www.census.gov: https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf
105. Venus Tamturk (2016, 07 05). eCommerce Trends and Statistics. Retrieved from cms-connected.com: http://www.cms-connected.com/News-Archive/July-2016/eCommerce-Trends-and-Statistics
106. Verhoef P.C. Peter (2003), “Understanding the effect of customer relationship management efforts on customer retention and customer share development”, Journal of Marketing, Vol.4, pp.30-45.
107. Verhoef, Peter, C., Franses P. H., Hoekstra, Janny C. (2002). The Effect of Relational Constructs on Customer Referrals and Number of Services Purchased From a Multiservice Provider: Does Age of Relationship Matter? Journal of the Academy of Marketing Science, 30(3), 202-212.

108. Viata-shop.com. (2017, 08 02). Retrieved from viata-shop: https://www.viata-shop.com
109. Weber R. (1996), “Customer segmentation for banks and insurance groups with fuzzy clustering techniques”, In J. F. Baldwin (Ed.), Fuzzy Logic, New York, NY: Wiley, pp. 187-196.
110. Wei Jo-Ting, Lin Shih-Yen, Wu Hsin-Hung. (2010). A review of the application of RFM model. African Journal of Business Management, 4(19).
111. Weiss and Indurkhya N.(1998), “Predictive data mining:a practical guide”, Morgan Kaufmann Publishers, First edition.
112. Westphal C. and Blaxton T. (2005), “Introduction to Data Mining and Knowledge Discovery”, Two crows Corporation, Third Edition.
113. Woo Ji Young, Bae Sung Min and Park Sang Chan (2005), “Visualization method for customer targeting using customer map”, Expert Systems with Applications,Vol.28, pp.763-772.
114. Yeh IC, Yang KJ, Ting TM. (2009). Knowledge discovery on RFM model using Bernoulli sequence. Expert System Applications, 5866-5871.
115. Zeithaml V. A. (2000). Service quality, profitability, and the economic worth of customers: what we know and what we need to learn. Journal of the Academy of Marketing Science.
116. Zineldin M. (2012). Relationship management for the future. Lund: Studentlitteratur AB.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Achieving Customer Loyalty from Email Campaigns by Using Data Mining Techniques

Richting: **Master of Management-International Marketing Strategy**
Jaar: **2017**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Naber, Julia

Datum: **22/08/2017**