Summary

The goal of this master thesis project is to model the business, email marketing process and apply the data mining techniques for a start-up e-commerce company called as Viata online pharmacy. The reason for this research is to establish better customer relationship through future emails by applying data mining techniques, to achieve efficient, effective retention emails marketing campaigns in the near future for the company. This will support the future retention email marketing campaigns to be more data driven policy decisions. This thesis can be divided broadly into three parts and is composed of seven chapters as follows:

Part I: The 1st part explains the company's business model in general, the raw data availability and describes in detail the email marketing process. Having a complete understanding of the company's business and overview of the business model components, this thesis focuses on the retention email marketing component. Thus, this part presents in-depth the complete retention email process, analysis and results of the sale order generated from the retention email marketing campaigns of the company.

Part II: Based on the detail analysis and results presented from the part one, appropriate data mining algorithms are applied on the retention email data set by using R software. For the prediction of the coupon usage from the retention email data set, two algorithms are applied namely decision trees and random forest in R software. We then present the comparison of the two prediction algorithms results for the retention email dataset. The last algorithm called as apriori algorithm is applied to the retention email dataset, to perform market basket analysis to generate interesting association rules, which can be used by the company in to bundle products in the future email campaigns.

Part III: Finally, to close the loop of the thesis, the complete process is given to apply the appropriate data mining algorithms in the retention email campaign. Following with this, a test process model had been developed to test the retention email campaign idea through data driven and policy driven decisions.

Each chapter deals with various aspects of this master thesis as explained below:

Chapter one, gives the introduction and defines research questions of the company.

Chapter two, discusses the literature review on the subject data mining. This chapter gives an overview of the data mining concepts.

Chapter three, explains the research methodology used in this thesis. It explains, how the literature review had been carried out and what type of raw data were available in the company.

Chapter four, explains the company's complete business model.

Chapter five, explains the company's general email marketing process, retention plan decision modelling notation and the results of the retention email order analysis to identify and apply data mining algorithms.

Chapter six, explains the applied algorithms in R software and presents the comparison between the prediction algorithms applied on the retention email dataset i.e. random forest and decision tree algorithms. It also explains association rule mining application on retention email dataset by using R and gives the results of the market basket analysis from the retention email data set.

Chapter seven, presents the process for the application of the data mining algorithms in the retention email process and general test model process build to test the future campaigns in retention email for the company. Finally, the last chapter is closed with the conclusion.

Preface

I hereby present my master thesis, “Viata online pharmacy”, Retention email process modelling and application of data mining techniques using R”. This thesis had been written in order to fulfil the requirements of the degree “Master of Management, specialization in Management Information Systems” at University of Hasselt situated in Hasselt, Belgium, Europe.

The thesis had been carried out in an e-commerce company called as “Viata online pharmacy”. There are more than 50 online companies in Belgium, which offers pharmaceutical products online. Nevertheless, in pharmaceutical business, currently the company is in the top three in online market and growing rapidly with new customers every week. Working in the company for more than a year, under the direct supervision of entrepreneurs, gave me an enriched experience in the field of pharmaceutical industry online web shop, online marketing strategies, logistics and warehouse and the experience of data science analyst.

I am heartily thank-full to the entrepreneurs of the company, Benoit Van Huffel and Gianni De Gaspari for giving me an opportunity to work on their project for my master thesis. My sincere gratitude to the professor Dr Koen Van HOOF for his encouraging supervision and supported me to develop an understanding of the data mining concepts. With his constant feedback and precise guidance had helped me to finish my thesis in the right direction.

Also, I am grateful to my parents for their support and to Daniela Paula Barbu, who supported me right from the start of the thesis. Finally, I thank the Viata team, my friends and the people who were involved during my thesis.

Saud Khan

Hasselt, Belgium

May 2017

Contents

List of Figures

List of tables

1 Chapter 1

1.1 Introduction

The field of knowledge discovery plays a key role in many organisations. Small organisations as well as large organisations benefits from their data availability to take better business decisions. Customers are one of the highest priority and property of an organisation to generate constant profit(Cuadros and Domínguez 2014, Ziafat and Shakeri 2014). In this era of internet, the organisations, whose business is through online channel generates digitalized transactions, which results in a boost of customer information stored in the large databases. These large information present in the database can be used to extract useful information, which is useful to solve business challenges of an organisation.(Coussement, Van den Bossche et al. 2014). Data mining techniques are used to mine unknown information from the raw data and these techniques of data mining can be applied in various fields.

Data mining plays, fairly a significant role in the field of marketing area. Nevertheless, the marketers trust their practical experience acquired from their years of experience. However to support the marketing business decisions, application of data mining techniques can be used to support the various marketing decisions(Ziafat and Shakeri 2014). There are many case studies and projects done, where the data mining techniques have been used to support the business decision(Chen, Tang et al. 2005, Koh and Tan 2011, Loraine Charlet and Kumar 2012, Radulescu, Băcilă et al. 2012, Zhao 2012).

In order to apply data mining in the organisation, the first and foremost step is to have a complete understanding of the business(Wirth and Hipp 2000).Having a clear overview of the business will help to set up the business goals clearly. There are many algorithms available in the field of data mining, where the algorithms can be applied based on the business objectives to be achieved(Witten, Frank et al. 2016). In recent years, the concepts of data mining techniques have been gaining increasing popularity among the organisations to drive their business decisions. Many organisations have been using data mining techniques to constantly evolve and improve in their respective fields.(Koh and Tan 2011).

The focus of this thesis is to model the company business, to fully understand their business and this will help to identify the application of appropriate data mining algorithms in their business model process for desired business scenario. With the constant growing competition among the online pharmaceutical websites, it becomes important to retain the customers in the organisation to stay profitable and competitive in the business. With the availability of the raw data such as customer data, transaction data and many more data generated from the online business, which are stored in the databases and can be used to further analyse respectively to achieve business goals. Therefore, for an organisation to stay competitive in the business, it is necessary to analyse their raw data constantly to improve, improvise their business to stay in their respective market field.

1.2 Problem statement

In most companies, typically 10% to 20% of the customers create 60% to 80% of the company's profit. To achieve high chances of successful business, identification of high profit customers is important, whereas these customers brings high profit margin and are minimal risk customers, where less risk is involved with these customers. These type of customers is important to retain them in the business. For marketers it is important to identify and drive the other customers in the high profit and low risk cluster (Rajagopal 2011).Therefore, it is important for the marketer to plan an effective and efficient campaigns plan to identify and target the profitable customers.

The online pharmaceutical company focuses to understand their customers better and to establish better customer relationship. To apply data mining techniques in the raw data, it is necessary to understand the business and various business processes involved in the company. Therefore, the following problem statement is formulated as follows:

What is the current business model of the Viata online pharmacy and what are the appropriate data mining techniques that can be used to implement into the retention email dataset process?

This thesis focuses on one component of the business model that is retention email marketing. The problem statement is further elaborated in to sub questions as follows:

1. What are the various decision, process and strategy involved in the retention email marketing of the company?
2. What is the impact of the retention email marketing of the company with respect to sales order?
3. What are the appropriate data mining algorithms to apply in the retention email data set based on the defined business context of the company?
4. Which data mining algorithm have the higher accuracy of prediction for the coupon usage in the retention email data set?
5. What are the interesting product association rules in the retention email data set?

The first question investigates the various email marketing strategies used by the company and the second question analyses the total number of sale orders which were generated from the retention email strategy.

In the next sub questions mentioned below, answers the application of appropriate algorithm identified to specific business requirement and presents the results of best prediction algorithm in the retention email data set. The final question investigates the interesting product association from the retention email dataset.

2 Chapter 2

2.1 Literature review

Data mining is a process of extracting useful potential information from the raw data. Various data mining methodologies are used in marketing and customer relationship management. Companies use these methodologies to increase profit and establish better customer relationship from their customers.(Kashwan and Velu 2013).Customers are one of the most important property of an organization. One of the goals of the organization in marketing department is to understand each customer individually. Organizations analyse customer information in order to strengthen relationship and to deliver appropriate services to the right customer.(Ziafat and Shakeri 2014).

Data mining are classified into two categories called as descriptive and predictive. Descriptive models extract useful patterns and explore its property and characterize the properties of the data from the database. Whereas predictive models try to predict future trends or behaviour depending on variable presented in the dataset(Witten, Frank et al. 2016). It is important for an organisation to follow the CRISP methodology for the application of data mining projects. The first step is to understand the organisation business in depth, as this will help to target the business areas to apply data mining projects. The second step is to have general understanding of the data available in the organisation. This will give further insight of the raw data, which can be used to define the specific data required to apply data mining algorithms. The third step is the most time-consuming depending upon the size of the raw data. This step focuses on the data preparation such as removing missing values, arranging the data and converting the data values in to factor. The next step is to identify the data mining algorithms to be run on the data set and then finally to evaluate the outcome of the results. After following the CRISP methodology, the organisation can take actions on how to deploy the outcomes of the data mining project. Figure 2.1 gives the overview of the steps taken for the data mining CRISP methodology, which is cross industry standard process(Wirth and Hipp 2000).

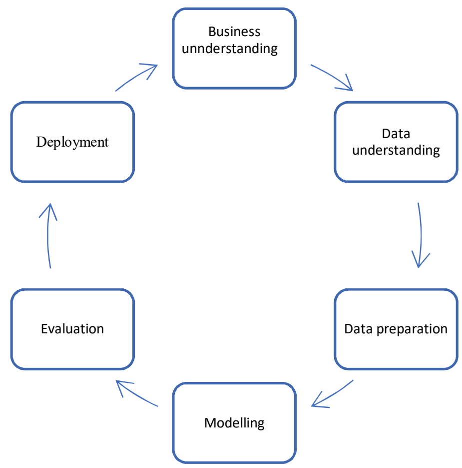


Figure 2.1: Crisp methodology

2.2 Data mining algorithms

Data mining algorithms such as association rules, sequential pattern, neural networks, clustering and classification are commonly used in many business scenario. Association rule and sequential pattern are used to discover association, similar patterns and correlation relationships in a data set. Association rule algorithm deploys one of two common approaches called as breadth first search (BFS)approach and depth first search(DFS) approach.(Ziafat and Shakeri 2014).

Association rule algorithm is widely used in the supermarkets and retail stores to find interesting product relation in their market basket data set. This technique is also known as market basket analysis. In many organisations, the marketing decisions on promoting bundled products are decided through association rules and combining practical experience of the marketers. Organisations can segment their customers in to different clusters and based on their product basket, association mining can be applied in order to find interesting relationships between products(Cuadros and Domínguez 2014, Kaur and Kang 2016).This will help the marketers to take decision on the product placement, pricing, profitability and promotions(Chen, Tang et al. 2005, Loraine Charlet and Kumar 2012). Association rule mining is based on the K-Apriori algorithm. The output of the algorithm gives support, confidence and lift to evaluate the interesting rules. There are several papers available online to find out how does the K-Apriori algorithm works(Torgo and Torgo 2011, Loraine Charlet and Kumar 2012, Kashwan and Velu 2013, Kaur and Kang 2016).

Prediction algorithms such as decision tree and random forest are used to predict the probability based on the desired variable given. Prediction can be broadly defined as the classification or numerical prediction. Numerous organisations had applied these kind of prediction algorithms. However, this part of the data mining process is one of the hard part. Based on the results, flawed business decision could be taken if the predicting algorithms are not properly deployed. One such example of application of prediction algorithm used is in the strategy based video games(Weber and Mateas 2009). The general idea behind the decision tree algorithm is based on divide and conquer. It starts to build the tree based on the highest variable information gain and keeps on splitting further depending upon the variable information gain(Varma and Rao). Another example of typical application of the decision tree algorithm are used in health care industry. Whereas based on patients attributes given , to classify whether the patient will suffer from diabetes or not, the algorithm will be able to predict the patients chances to suffer from diabetes or not. This prediction information and practical experience of the doctor will help to take necessary decisions and advice accordingly to the patients.(Koh and Tan 2011). Figure 2.2, gives the overview of general data mining algorithms used.

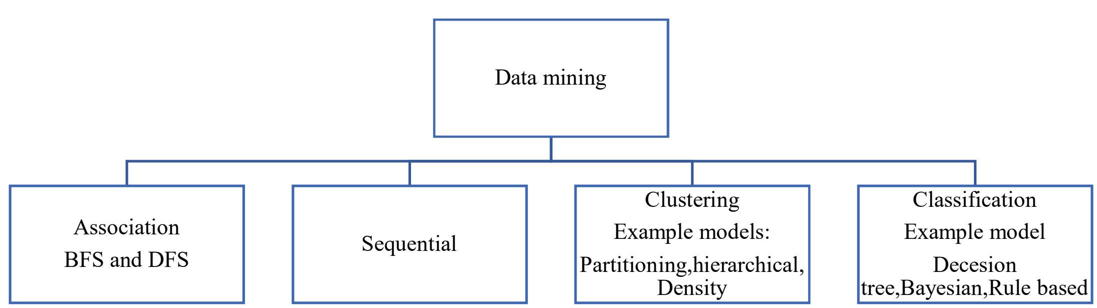


Figure 2.2: Overview of data mining algorithms

2.3 CRM Data mining

A popular technique used by the marketing professionals to analyse their customers is called RFM analysis. RFM stands for recency, frequency and monetize. This technique is one of the most widely used in marketing strategies to segment the customers(Radulescu, Băcilă et al. 2012). Many case studies had been published in the business area such as tourism sector, telecom sector with respect to application of RFM analysis. Based on the variables of (R), (F) and (M), data mining techniques have been applied on RFM analysis to take better decision by marketing managers. (Khajvand, Zolfaghar et al. 2011, Radulescu, Băcilă et al. 2012, Dursun and Caber 2016).

Clustering and association rule are the most popular data mining algorithm in customer relationship management. The goal here is to identify pattern in existing data. One of the common technique to divide the customers in to various segments is called customer segmentation. It is a process of dividing customer into meaning homogenous groups based on attributes and characteristics. The type of segmentation depends on the business objective to be achieved. The type of segmentation which are commonly used are value based, behavioural, propensity, loyalty, demographic and needs based. (Cuadros 2014, Hosseini and Ziaei Bideh 2014, Ziafat and Shakeri 2014).

3 Chapter 3

3.1 Research Methodology

To achieve the knowledge on the subject matter and to answer the objective of problem statement various methodologies had been used. These methodologies range from in depth literature review, practical field experience, data collection and many more.

Literature review was primarily done through the Hasselt university online scientific database. Google scholar was also used in the process of literature review. Several keywords were used to find the scientific articles in the field of data mining. The keywords such as "RFM", "Marketing strategies", "E commerce", "Data mining application", "Prediction algorithms", " Market basket analysis"," Association rules", "Business process", "Data mining techniques" and many more relevant key words had been used to gain enough insight on the data mining literature. Several case studies were read for the application of data mining techniques in organisations. Lastly, to strengthen the practical application of data mining algorithms various classes had been followed from the Hasselt university and online courses offered in the field of practical data mining.

Apart from this, the practical experience from the company was developed overtime through working and gaining an experience from the company database. This helped to build a solid foundation to understand the availability of the raw data and the company business model. The data gathered for this project was purely live data used and several software's were used to gather data accordingly to solve project goals. The list of software used to gather the data are listed as follows:

- Business intelligent tool – Qlik view software.
- Open source Viata Enterprise Resource Process system(ERP).
- Copernica email marketing software.
- Google analytics

For the analysis of the data, design of the process and application of data mining algorithms, list of software were used in the process:

- R studio open source software
- IBM SPSS software
- Microsoft Excel 2016

- Signavio business modelling software.

After having a hands-on experience with the data, while working in the company, gave me enough insight to understand the data and to define the process of the company. The time span of the data used in this project is from 2014 till end of the December 2016.The next section describes clearly the data source of Viata online pharmacy.

3.2 Data collection

The online pharmaceutical company uses an open source ERP system for their business. In order to make marketing-decisions, the company uses the business intelligence tool called as “QlikView” to analyse their sales, customer’s profile and products. Since 2014, the e-commerce company has stored raw data from their customers in their database. To mention a few variables such as, unique client ID, name, age, sex, geographical location, and many more variables which could be extracted from their database. The list of variables considered for the retention email marketing is given in the appendix 1.

Several variables can be extracted and defined as per the need of company business challenges. With the availability of raw data, it is possible to apply various and appropriate data mining algorithms.

3.2.1 Data A: Raw data

The data generated from the website such as customer data, purchase data are stored in the company’s database. The order received from the website, the customer data is stored and passed to ERP system for further processing of the order. The data such as logistics data, packed time, shipped time are generated through ERP systems. Finally, the raw data from the company website database and ERP system are exported to the Qlikview business intelligent tool. The combine data from both sources are available in Qlikview and can be easily extracted for further need based on business requirements. Figure 3.1, shows an overview of the data source A.

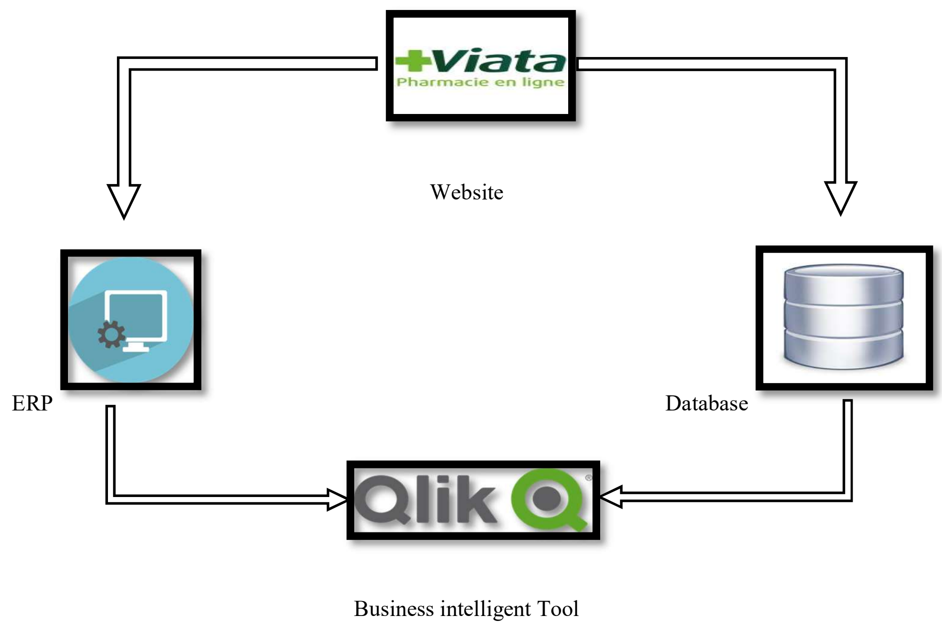


Figure 3.1: Data source, Data A

3.2.2 Data B: Anonymous data

The online marketing and advertisement data acquired from the online marketing techniques such as search engine optimization, search engine advertisement as well as user online behaviour can be extracted from google analytics tool. Google analytics tool captures the advertisement clicks statistics and many more statistics of the various online advertisements which are used for online marketing. However, google analytics tool is completely independent and the data here is anonymous. This means for an example google analytics tool does not give any information of the user, who interacted with the google search engine instead it shows how many users interacted on advertisement on google search engine. However, Data source B had been taken consideration for this master thesis project. Figure 3.2, shows the overview of the data source B.

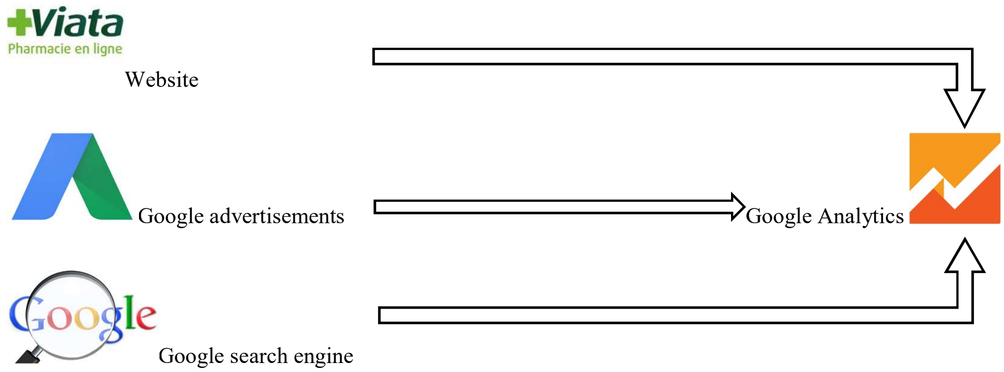


Figure 3.2: Data source, Data B

3.2.3 Data C: Copernica data

The focus of this project is mainly on retention email marketing. To do the analysis and comparison based on the research questions, we require data to be extracted from Copernica software. This software is used for email marketing purposes. The data such as list of emails sent out for specific campaign and list of users who reacted on the emails sent, such data statistics can be obtained from the software. However, the data extracted from the software does not have any relationship with other data source A and B. This data is here is also completely independent from other two data sources A and B. The only way to match the data is to extract the email address of the customers from the software and then to compare the list of emails acquired from the Qlik view which is data source A. With this manual process all the variables associated with the list of email addresses can be linked and complied for further required analysis. Figure 3.3, shows the overview of the data source C.

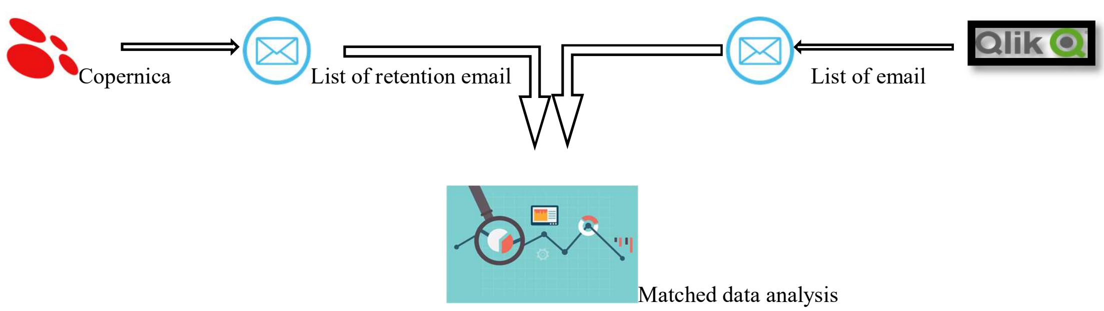


Figure 3.3: Data source, Data C

4 Chapter 4

4.1 Viata online pharmacy business model

The online pharmacy web shop serves their customer with better lifestyle, healthy living and to improve patient's quality of life. It provides customer satisfaction through recommendation on the products and the fastest next day delivery option to their customer. The e-commerce company has three core business components which are logistic, web shop and marketing. The important business component is marketing and the focus of this master thesis project will be towards the marketing component. Figure 4.1, shows the Viata online pharmacy business components.

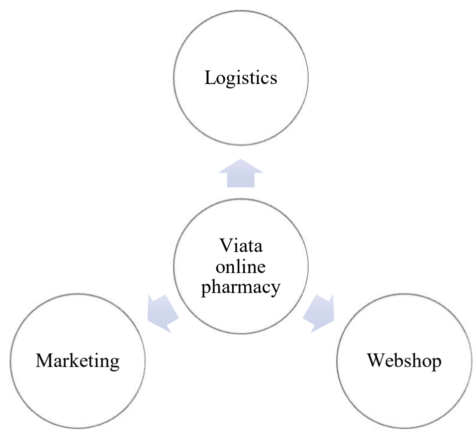


Figure 4.1: Viata online pharmacy business components

4.1.1 Product supplier and category

There are broad range of products offered from the e-commerce company, such as general medical products, beauty items, sports nutrition, medical household products, baby nutrition, fat burners, medicines, and many more category of products. Majority of the products are supplied through various wholesalers in Belgium and few products are supplied through the e-commerce company itself.

4.1.2 Customer relationship

The company recognizes the importance of their customers' needs and satisfaction. With a team of trained pharmacists and beauticians, it offers support to their customers such as product description, recommendation on health and beauty products, fitness and medical advice. Moreover, it maintains a semi transactional relationship with its customers. For instance, a customer who ordered the products for the first time and second time, are given a hand written welcome card with their package. The welcome card is synchronized accordingly with the client chosen language. Currently, it offers hand written customer card in French, German, Dutch and English.

The company offers various promo codes to their customers on regular basis and it pays special attention to their loyal customers, wrong orders and late deliveries. The customers who received the wrong order or late delivery are compensated by offering a gift to that customer. Loyal customers are identified based on their frequency of purchase and recency of their purchase.

4.1.3 Web shop

The company's website can be reached through www.viata.be. It is offered in Dutch, French, English and German language. It offers more than 20,000 products and each product has a detailed information. Customers can chat and call to "Viata" team for the recommendation and information on different products and various general inquiries. The customers can choose different options of online payments such as Visa, MasterCard, Ban contact, Ideal, Euro check and PayPal. They can also choose delivery services from DPD, DPD parcel shop and PostNL.

The website maintains a blogging section on health tips and it is active in various social media platforms. Customers register their details in the website which makes easier for them and the company to have easy and efficient way to order, communicate and to keep a track of their parcels. Figure 4.2, shows general process of web shop customer journey.

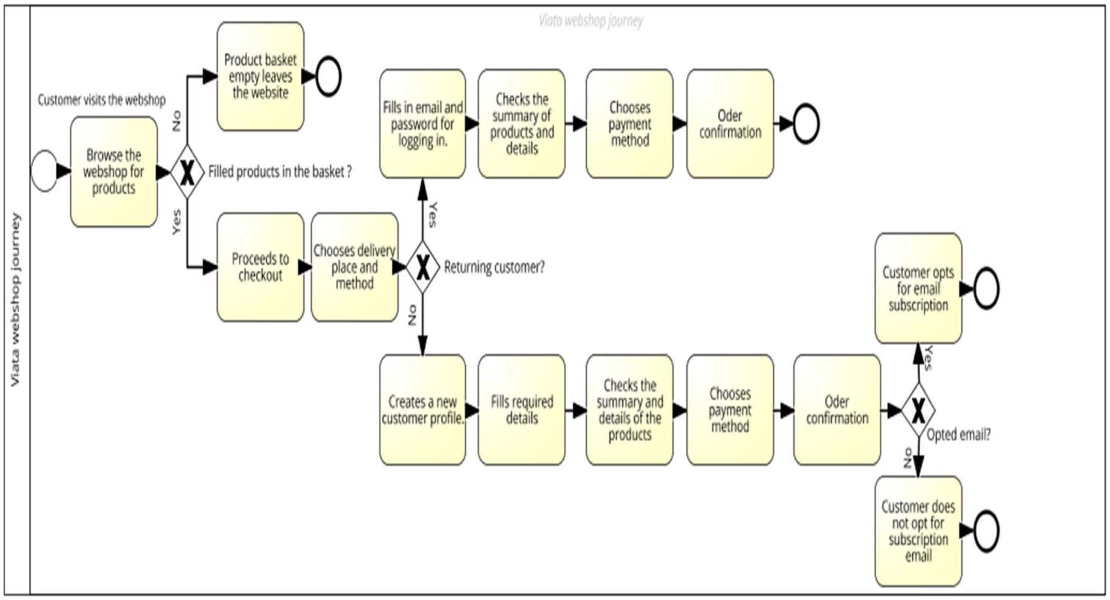


Figure 4.2: Customer web shop journey

4.1.4 Logistics and delivery

The company organizes the delivery of the orders in an efficient and organized way. It can deliver the orders next day in Benelux region with a reasonable cost. However, if there is a change in the standard procedure, this will most likely to end up with the additional cost. Extra care is taken with the customers' order to be processed, packed and delivered to the appropriate regions. The logistics process involves the following step: receiving the order, check the contents of products, label the products with the stickers, classifying the customer whether he or she is the first time or returning customer, check the contents of the order, pack the products with bubble wrap where ever required and place it to the final delivery karts. Finally, at the end of the day all the orders are completed and ready to be shipped through DPD, DPD parcel shop and PostNL. DPD and DPD parcel shop delivers across European regions, whereas PostNL only delivers in Netherlands.

Since the company offers pharmaceutical products, it pays special attention to the customers' orders. Each order is processed and checked carefully. Therefore, the company has a fairly high accuracy of the correct products delivered to their customers. Figure 4.3, gives an overview of the logistics process.

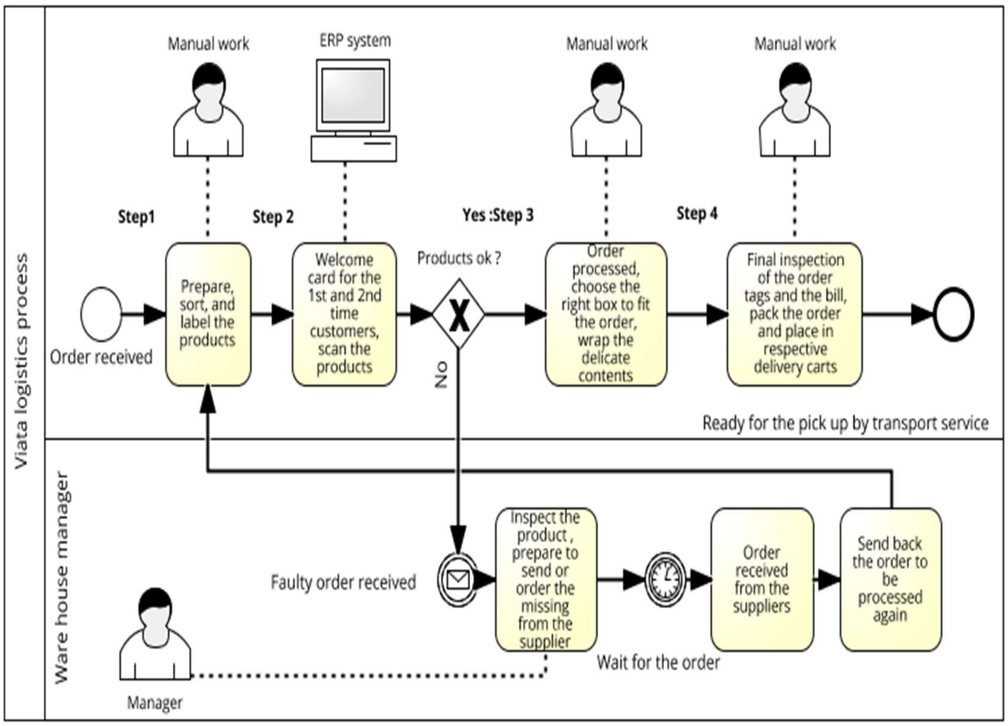


Figure 4.3: Viata logistics process

4.1.5 Online marketing

Marketing is the most important business component of the company. At present, there are more than 50 online pharmacy websites in the Benelux region. To stay competitive in the market, it must innovate and make constant improvements from every angle of the business component of the company. It has to study their customers, analyse their customers' behaviours and to make constant improvements in their online marketing strategy. In the figure 4.4, shows the various strategies used for the online marketing area. The focus of this master thesis project is only on retention email marketing.

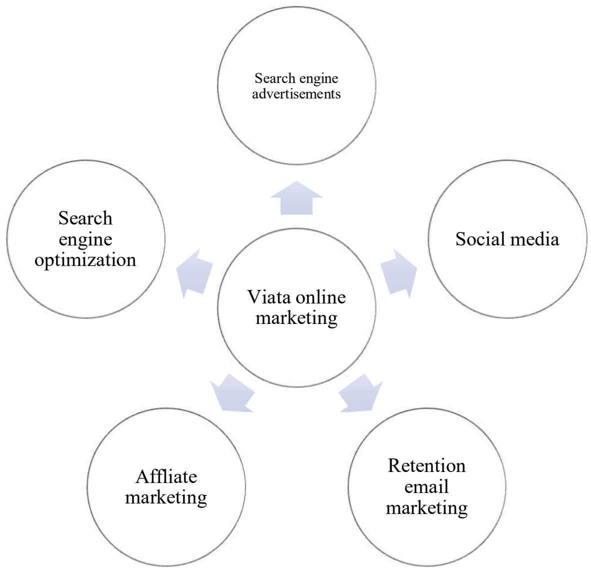


Figure 4.4: Viata online pharmacy marketing strategies

5 Chapter 5

5.1 Product categories and RFMHS

Viata online pharmacy, has eight main product categories for email marketing. They are as follows:

1. Hair loss products.
2. Vitamin products.
3. Joint and muscle ache products.
4. Slimming products.
5. Baby products.
6. Body and facial products.
7. Sports and nutrition products.
8. Generic products, which includes the rest of the products.

The customers are classified based on their expenditure in to one of the above product categories. Based on the market basket of the customer and the highest amount spend in the product category, the customers are offered promo code accordingly. The customers who spend more on purchases are excluded from the free delivery promo code. Whereas customers who spend less they are offered a free delivery promo code. All the promo codes have an expiration date, which is usually 30 – 45 days.

The decision taken for email marketing is based on RFMHS analysis. The decision made here is more policy driven and not data driven. RFMHS stands for recency, frequency, monetize, highest sales and the start period. The complete table of decision made in retention email campaign to offer coupon is given in the appendix 2. In general marketing professionals use RFM analysis for marketing purpose. However, "Viata" has created two extra field in RFM analysis called as H and S. These unique H stands for highest sales in over the last 12 months and S stands for the period when did the customer first bought the product. For the description of the "Viata" RFMHS description refer to appendix 3.

Figure 5.1, gives an overview of the retention email policy driven decision modelling notation for the coupon offered. For the detailed policy driven decision with respect to RFMHS conditions and the coupon offered, refer to appendix 2. The selection is made in the Qlikview business intelligent tool in RFMHS selection and the list of customer's emails are extracted and retention emails are sent through the Copernica software. For an example, the list of customers whose market basket contains Etixx products will fall in sports category and frequency will be one will get a retention email and the promo code based on the purchase amount.

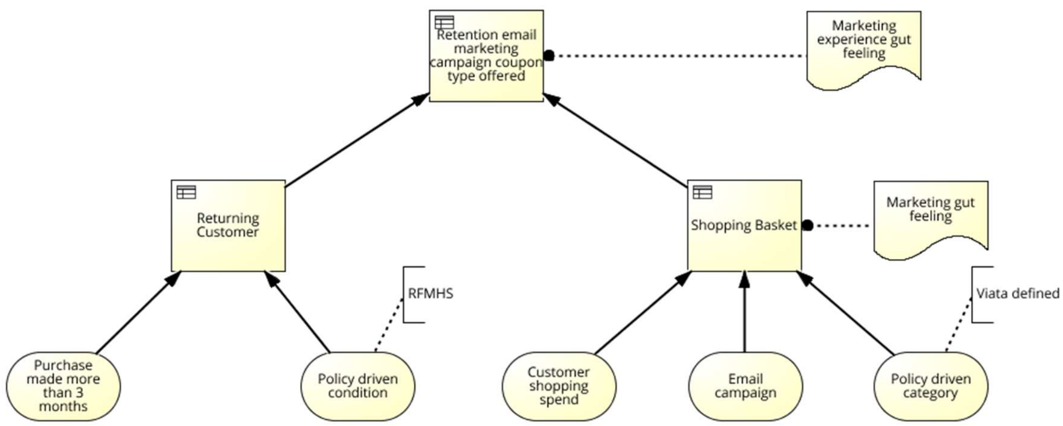


Figure 5.1: Retention email coupon Decision model notation

Figure 5.1 gives a general overview of how coupons type is offered to retention email customers. The complete list of DMN table rules can be found in the appendix 2.

5.2 Email marketing component

The e-commerce company uses Copernica software for email marketing purposes. The data of the customers are extracted from the Qlikview business intelligence tool. Based on RFMHS analysis, they are segmented according to their frequency, spending and purchasing behaviour. In the end of the process, their emails are compiled based on the segmentation and imported to Copernica software for further email marketing.

The customers are classified into two categories as first time customers and returning customers. First time customers are those who made a purchase for the first time through the company's website. The returning customers are identified as the customers who made a purchase more than one time through the website. The customers who had opted for subscription email will get a promotional email every two weeks, regardless of first and returning customers. Promotional emails are also send out to the users who subscribed for the email even though they did not purchase anything from the website. The registered and returning customers who are not active for more than three months, they get a retention email. The retention email is send usually in the middle of the month. The retention email marketing process mainly gives priority to returning customers. The returning customers gets a retention email based on the policy driven selection of RFMHS values.

5.2.1 Promotional email marketing

Customers who opted-in for subscribed email including the returning customer will get a general promotional email. Customers who has hard-opted-out for email will not get any emails at all. Based on the previous marketing experience and customer behaviour, the company have segmented their customers in to three categories as follows:

1. Customers who bought baby products will get promotion related to baby products and the generic email content.
2. Customers who are above 50 years of age will get different promotional and the generic email content.
3. Customers who do not fall in any of the two categories, they will be offered general email promotions.

If the customer falls in more than one category, they will get promotional email from the respective categories. Regardless of any categories, generic promo content will be always there. Figure 5.2, explains the promotional email process given below.

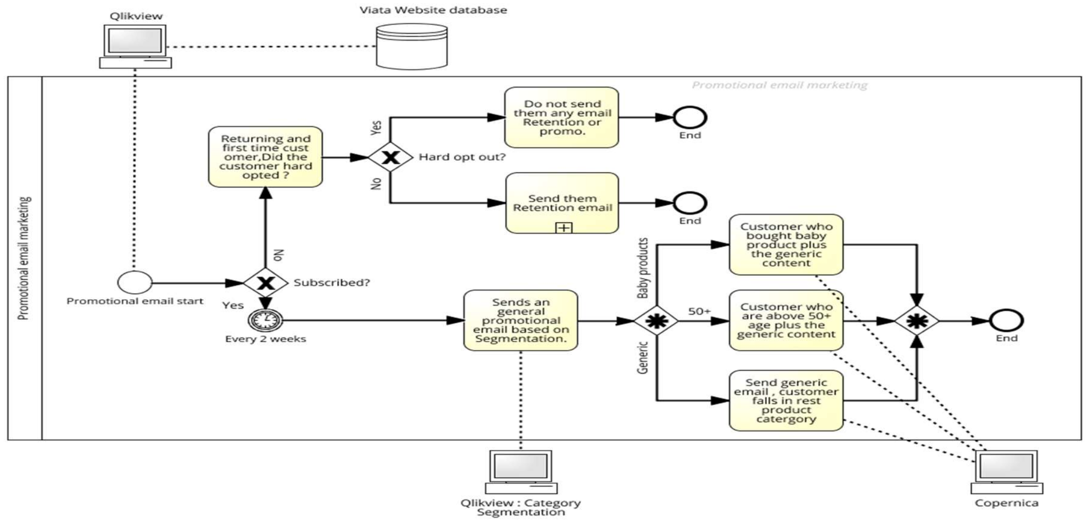


Figure 5.2: Viata online pharmacy promotional email process.

5.2.2 Retention email marketing

Customers who are inactive for more than 3 months, they get a retention email based on their past purchasing behaviour, spending and market basket behaviour. The company, has defined eight product categories, which can be found in the section 5.1. Based on RFMHS analysis, the customers receive a retention email accordingly. Customers who hard-opted-out for email will not get any emails. The retention email is send usually in the middle of the month. Figure 5.3, shows the various the process involved in the retention email. For the complete overview of the Viata email marketing process can be found in the appendix 4.

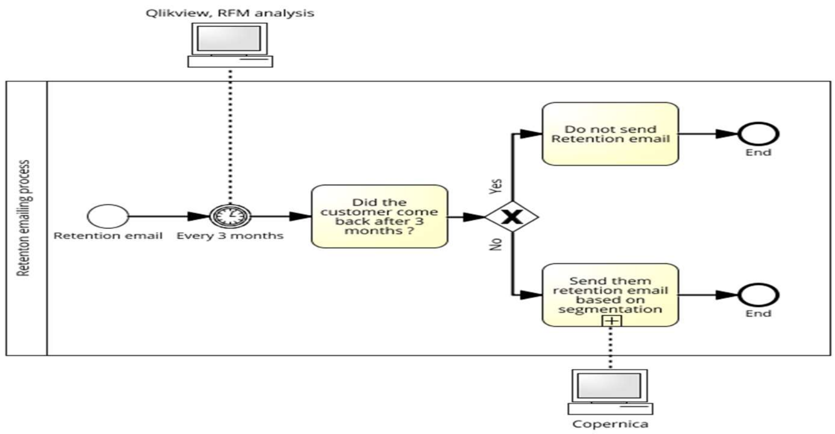


Figure 5.3: Viata online pharmacy retention email process

5.2.3 Retention plan decision process

Customers get the retention email based on the RFMHS analysis. Different retention emails are sent based on the retention plan. The retention plan is developed based on the gut feeling and past marketing experience. There are 9 different retention email content for email marketing. They are as follows:

1. Viata baby products.
2. Viata fitness and sports.
3. Viata general retention.
4. Viata beauty and care products.
5. Viata generiek(rest) products.
6. Viata joint and muscle products.
7. Viata slimming and hair products.
8. Viata vicks products.
9. Viata auto campaigns.

Each of these email content has different promo codes and are further categorized accordingly to retention plan. Each of the email content, shows the process, decisions and the number of emails sent and how many orders came out from the emails sent. The analysis on retention email with respect to specific category is explained more in detail given in the appendix 5. The order analysis was extracted with respect to each category from the Copernica software and then compared with the specific selections made in the Qlikview. This process was time consuming because there was no link between the dataset. More over every category of retention email in Copernica software, statistics were extracted and manually compared to obtain the results for individual retention campaign analysis and orders. Table 5.1 and figure 5.4, which is in the next section gives the result of the analysis.

5.3 Retention email analysis

To analyse the impact of the retention email, we have extracted the email statistics data from the Copernica software. The statistics data presents the number of emails send for each user created content of the email. The statistics report of emails is further divided into 6 parts as follows:

1. **Error:** Number of emails send had encountered error somewhere in delivery process.
2. **No response:** The software was not able to register the response from the subscriber.
3. **Impression:** This shows the number of email where opened but did not clicked on the content of the email.
4. **Click:** Number of users clicked on the hyperlink of the email.
5. **Unsubscribed:** The number of users who clicked to unsubscribe the emails.
6. **Complaint:** Number of users who reported a spam or abuse otherwise.

Starting from 2014 April till 16/12/ 2016, there were 118,530 emails where send under retention email campaign. There were 1,424 errors, 75,002 no response, 33,929 impressions, 6,543 clicks, 1,808 unsubscribed and 75 complains under retention email campaign.

The next step is to extract data from the Qlikview to analyse, how many sale orders came out from these retention email marketing campaign. There were in total 543 sale orders came out, where 271 sale orders used coupon codes provided along with the emails.

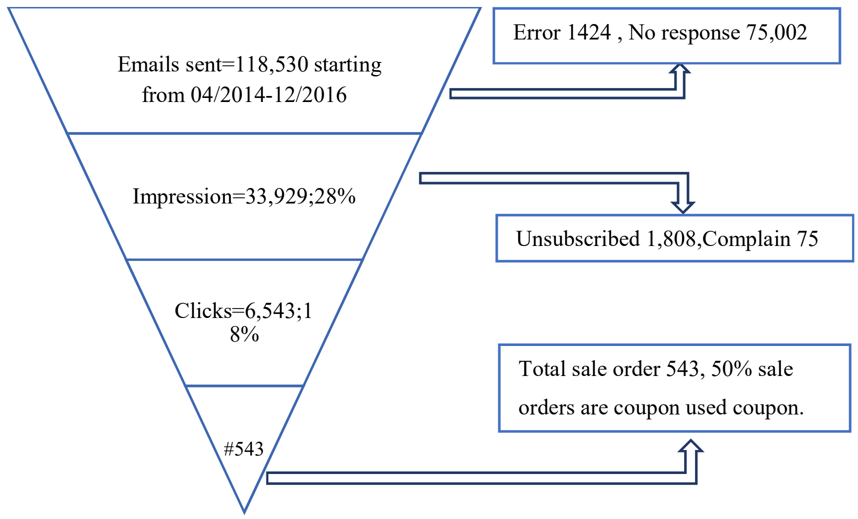


Figure 5.4: Viata online pharmacy retention email sale order analysis

5.4 Individual retention campaign order analysis

Based on Copernica statistics results obtained, we will only be focusing on impression data for each individual retention email campaign. The table 5.1, shows the total number of retention emails sent to the customers based on retention campaign, number of customers who had impression on email and number of order generated from each retention campaign. It also shows the number of orders, where coupon was used from each individual retention email campaign.

The net contribution is defined as:

Net contribution = Gross margin -handling costs-packaging cost-transaction cost-shipping cost,

Where gross margin = sale price of the product-cost price of the product.

Retention campaign	**# Emails sent**	**Users email impression**	**#NC**	**NC coupon orders**	**# orders**	**Coupon orders**	**% orders**
1. Etixx	4398	1490	€629.51	€284.17	42	21	50%
2. Ret_1	2229	632	€11.81	€6.76	17	15	88.2%
3. Baby	1899	402	€28.30	€8.97	10	8	80%
4. Eucerin	2490	836	€269.19	€152.96	26	10	38.5%
5. Generiek	61720	17401	€1,220.07	€463.61	209	117	56%
6. Gewricht	7705	2411	€175.66	€108.42	17	8	47.1%
7. Hair	716	183	€73.49	€29.11	6	4	66.7%
8. Vermageren	7644	1784	€150.07	€36.17	21	9	42.9%
9. Verzorging	10983	3019	€638.98	€166.12	89	47	52.8%
10. Vitamin	2099	622	€115.18	€35.44	9	5	55.6%
11. Reflex spray	1957	675	€12.59	NA	1	NA	0%
12. Vicks	482	131	€73.46	€65.72	5	2	40%
13. Auto campaign	3761	1284	€24.08	€18.17	23	20	87%

Table 5.1: Viata online pharmacy retention email sale order analysis

Figure 5.5, represents the retention email campaign report where the bubble size is the emails sent and on the y axis represents the average net contribution and x axis the conversion rate i.e. orders divided by total emails sent. Combining the data from Copernica and the Qlikview , various conclusion can be drawn by observing the figure 5.5.

Effective campaigns with respect to average net contribution by order and conversion rate:

The retention campaign which falls under high conversion rate and high average net contribution by order are Etixx, hair and Eucerin retention campaign respectively. These are the effective campaign and the rest campaign, the company should improve in future to increase the average net contribution and the conversion rate.

However, the generic (Rest products) has the largest size of emails sent, though with lower conversion rate and lower average net contribution rate. We can draw a conclusion rather than sending emails randomly for rest products, it is recommended to send emails by product wise specifically to the customers to achieve better results. The list of coupon condition which were used in the email campaign is given in appendix 6.

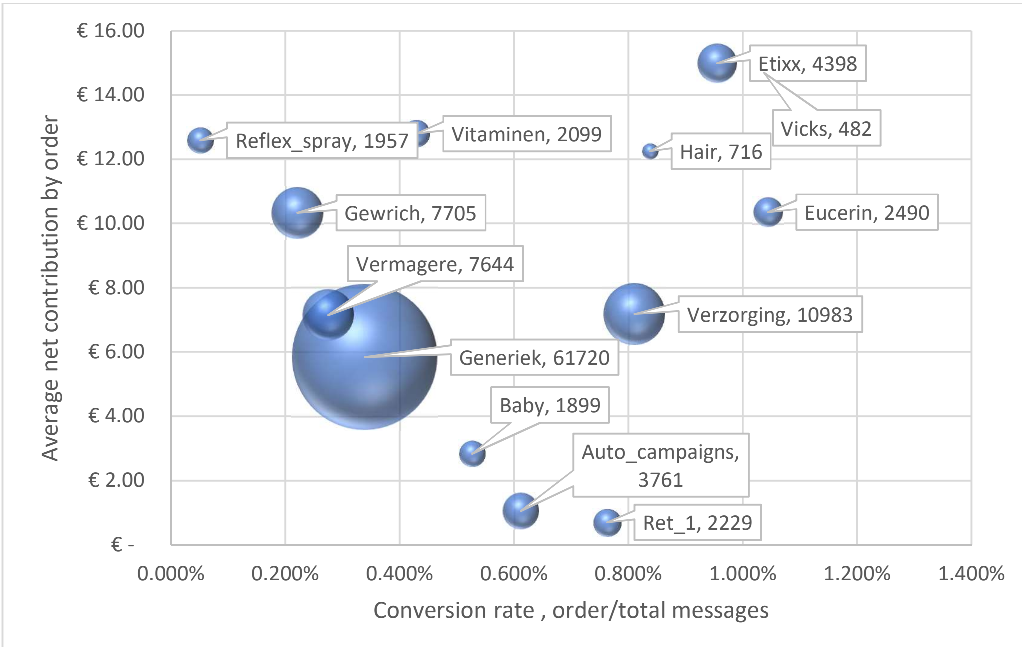


Figure 5.5: Viata online pharmacy retention email campaign report

6 Chapter 6

6.1 Retention email data mining

After having an overview and report on the retention email campaign, an interesting observation can be concluded from the figure 5.5, that is, almost 50% of the sale orders were coupon used. Hence, coupon plays an important role in the retention email marketing. Data mining algorithms can be applied on the retention email dataset to predict coupon usage. For preparing and cleaning the data, Microsoft excel 2016 is used and for the data mining algorithms application, “R” studio software version 1.0.136 is used throughout the project.

To predict whether the customers will use the coupon or not, two algorithms are applied as follows:

1. Decision tree classification algorithm.
2. Random forest classification algorithm.

In this case, the two algorithms are applied on the retention email dataset. Retention email data set will be referred as “Dataset_1”. This section explains the description of the variables, source of the Dataset_1, visual analysis of the Dataset_1, implementation of the algorithms and finally the comparison of the algorithms.

6.2 Descriptive analysis retention email Dataset_1

Dataset_1 is acquired from the Qlik view business intelligent tool. The data had been selected from 1st December 2014 till 31 Dec 2016 under the retention email campaign. The following variable had been obtained from the retention email campaign which is given in the table 3. However, there are many other variables we can choose from the raw data. Due to the scope of business requirement, which is prediction of coupon usage the following variable have been considered to develop the coupon prediction model. The structure of the data set is called as data frame in R studio. This dataset contains both numerical and categorical variables.

Variable name	Variable description	Variable type
1. Land	Country of the customer	Factor, 4 levels, Belgium, Netherlands, France, Germany
2. Age	Age of the customer	Numerical
3. Basket	Total number products bought by the customer	Numerical
4. Rev	Revenue generated by the customer	Numerical
5. NC	Net contribution towards the order by the customer	Numerical
6. Geslacht	Gender of the customer male/female	Factor, 2 levels, f, m
7. Coupongebruikt	Did the customer used the coupon? ja, nee	Factor, 2 levels, ja, nee

Table 6.1: Dataset_1 variable description

The given data had been cleaned and prepared thoroughly in Microsoft excel 2016. The cleaning process includes removing duplicate names, cleaning the cells contain missing values and rearranging the data to import to R studio. There are in total 540 observations and 7 variables in the Dataset_1. The summary of the categorical and numerical type variable is given in the table 6.2 and 6.3 below:

Land		**Gelascht**		**Coupongebruikt**	
Netherlands	313	Female	337	Yes	284
Belgium	195	Male	203	No	256
France	29				
Germany	3				
Total	540		540		540

Table 6.2: Dataset_1 categorical variables statistics

Statistics	Age	Rev	NC	Basket
Min	21	6	-11	1
1st Quadrant	43.75	29	1	2
Median	53	43	5	3
Mean	52.53	49.51	6	4
3rd Quadrant	62	61	11	5
Maximum	101	273	51	62

Table 6.3: Dataset_1 numerical variable statistics

After having a clear overview and explanation of the variables in the table 6.1,6.2,6.3, we can proceed further to analyse visually the Dataset_1. Figure 6.1, shows the histogram plot shown by coupon usage by country and gender.

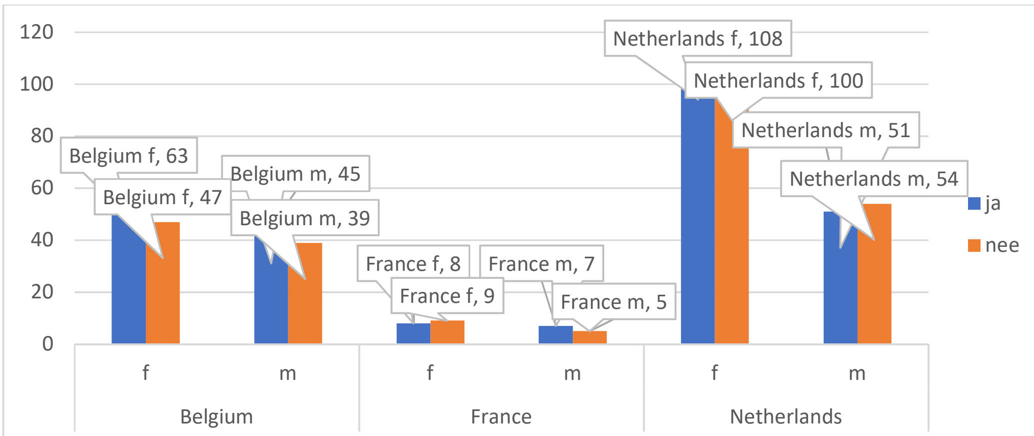


Figure 6.1: Dataset_1, Coupon used by country and gender

Clearly from the figure 6.1, the highest number of coupon users are from Netherlands and female are the highest users of the coupon in the retention email dataset.

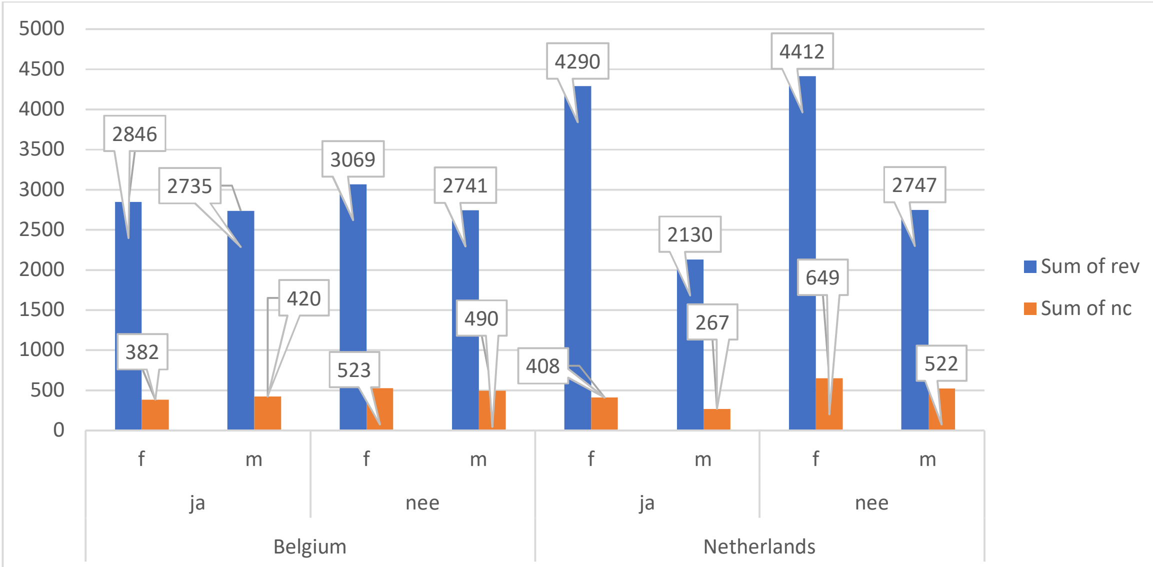


Figure 6.2: Dataset_1: Rev and NC chart by coupon usage

Figure 6.2, shows the histogram of the revenue and net contribution from the retention email dataset. Through observation from the histogram figure 6.2, is that the customers from Netherlands gives the highest revenue of 13,579 Euro's and females users are higher in contributing revenue. Interestingly the net contribution of Belgium (1,185 Euro's) and Netherlands (1,1846 Euro's) is almost the same.

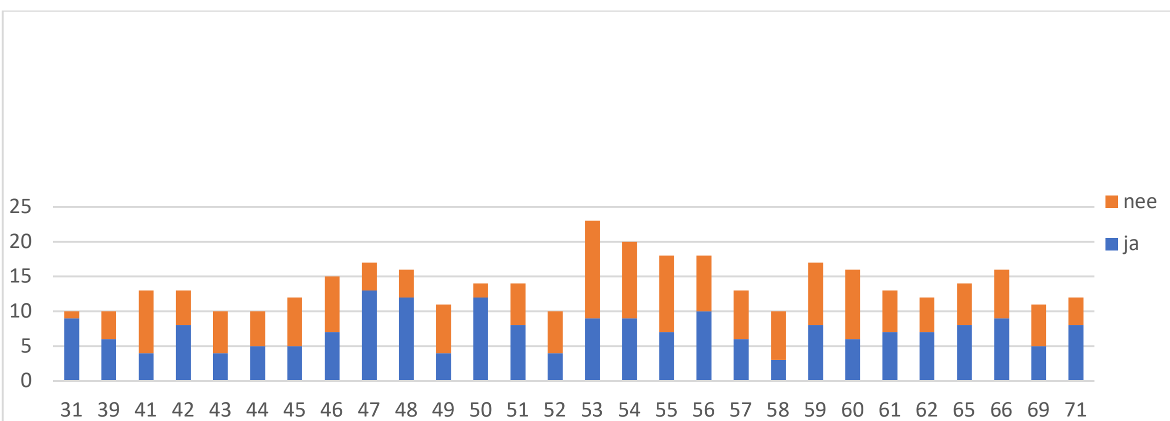


Figure 6.3: Dataset_1: Top 28 coupon usage by age

Figure 6.3, shows the top 28 coupon users with respect to age from the retention email dataset. Generally, coupon users are between 30 and 70 years of age. After having a visual examining of the coupon used by different variables, already we have some insight about the characteristics usage of the coupon from the Dataset_1. The coupon usage statistics is normally distributed that is around 47% have used the coupon and the rest did not used the coupon.

6.3 Decision tree algorithm application: Dataset_1

This section presents the application of the decision tree algorithm and building a prediction model on the retention email dataset to predict the coupon usage. In the retention email data set, the decision tree model needs to be trained first and then to test the prediction model based on random selection of the test data. The target variable here is to predict in the Dataset_1 is coupongebruikt. Since the target variable values is already given in the Dataset_1 , so this is a supervised learning and a classification problem.

There are mainly two type of algorithm in decision tree, they are classification and regression methods. Here our goal is to predict whether the customer will use coupon yes or no based on the given independent variables. Therefore, this is a classification problem, hence decision tree classification is used to predict the coupon usage.

The package to run the decision tree algorithm in R studio is called as “rpart”. To plot the decision tree output in graphical form, we make use of the package in R studio called as “rpart.plot”. The variables which are categorical, needs to be converted to factor variable so that R studio can identify these variables as categorical type. To make the prediction model, first the data needs to be split in to two parts randomly to train the model in one set and to test the model in the other set. The Dataset_1 had been split into the training set which is 80% and the test data which is 20%.The training data contains 464 observations and the test data set contains 115 observations. Seed needs to be set first, before running the algorithm so that the algorithm will use same order of the data set every time. The command to set the seed in R studio is,

- set.seed(1234)

To build and train the decision tree following command is used in R studio on the Dataset_1_train.

- Dt<-rpart(coupongebruikt~.,Dataset_1_train, method="class")

The result of the decision tree is stored in the new variable called as Dt. Here, the target variable coupongebruikt is run against all the variables given in the Dataset_1_train. The method here instructs the decision tree to use the classification approach. The output of the decision tree is given as:

"n= 464

1) root 464 213 1 (0.45, 0.54)

2) rev>=48.5 215 83 0 (0.61, 0.38)

4) age>=31.5 202 74 0 (0.63, 0.36) *

5) age< 31.5 13 4 1 (0.30, 0.69) *

3) rev< 48.5 249 81 1 (0.32, 0.67)

6) rev< 25.5 84 29 0 (0.65, 0.34)

12) nc>=-1.5 65 15 0 (0.76, 0.23) *

13) nc< -1.5 19 5 1 (0.26, 0.73) *

7) rev>=25.5 165 26 1 (0.15, 0.84,) *"

The above results give the variable importance gain information by each variable. Figure 6.4, shows the plot of the above results generated by the R studio. To display the decision tree in the nice and detailed graphical form, "rpart.plot" command is used as follows:

- rpart.plot(Dt, type=2,extra=101)

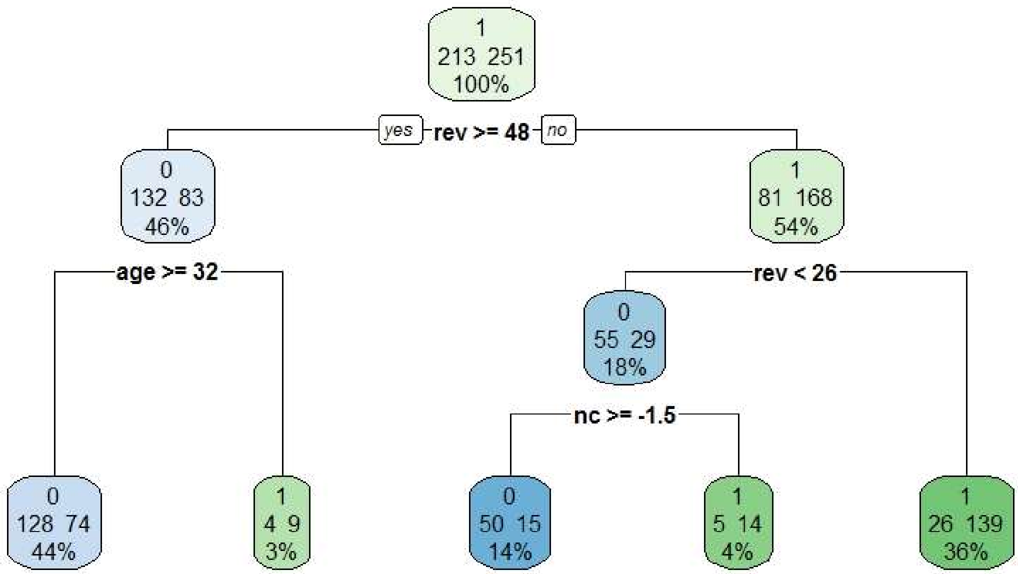


Figure 6.4: Decision tree on coupon usage

In figure 6.4, the algorithms start with the variable revenue because the variable has the highest information gain comparing with all other variables. The 0 indicates coupon not used and 1 indicates coupon has been used in the Dataset_1_train. The total number of observation in the dataset is 213+251=464. Where 213 used the coupon and 251 did not used the coupon. Next step is to validate the model in the test data to determine the accuracy of the model. The R command to predict the test data is:

- Dtpredict<-predict(Dataset_1_train, Dataset_1_test,type="class")

The command above stores the prediction in to the new variable called as Dtpredict. This gives the list of prediction in the Dataset_test_1. To cross validate the model:

- T<-table(predictions=Dtpredict, actual=Dataset_1_test$coupongrebruikt)

The result of the table is stored in to the new variable. The results of the T variable are given in the table 6.4 as follows:

	Actuals	
Predictions	0	1
0	49	11
1	12	43

Table 6.4:Decision tree test data confusion matrix

Here, in the above matrix, 49+43=92 times the prediction is made accurately with prediction model created and 12+11=23 times the model predicted wrongly. This means the model has accuracy of 80% on the test data set.

6.4 Random forest algorithm application: Dataset_1

The package to run the random forest algorithm in R studio is called as "randomForest". Load the random forest package in the R studio. Seed needs to be set before the execution of the random forest algorithm.

- Install.packages("randomForest") //To install the package.//
- library(randomForest) //To load the package//
- set.seed(1234) //To set the seed//

We need to split the Dataset_1 in to training and testing datasets. The same split dataset used in decision tree algorithm will be used here in random forest algorithm. To train and build a random forest model:

- modelrandomt←randomForest(coupongebruikt~.,data=Dataset_1_train,mtry=3,ntree=1000)
- modelrandomt

In the R command above, we have used all the variables presented in table 6.1, against the target variable. We have selected the number of variable at each split randomly which in this case it is 3. This means the algorithm will choose three variables randomly every time to plot the decision tree in the Dataset_1_train. The higher the number of trees plotted the higher the accuracy of the model will be given. In this case, we have selected 1000 trees to plot. The results are given below in the table 6.5:

"OOB estimate of error rate: 29.53%"

	0	1	Class error
0	150	63	0.295
1	74	177	0.294

Table 6.5: Random forest confusion matrix

Here the OOB means out of bag estimate of error rate or misclassification rate is 29.53%. This means the model accuracy is 70.47%.

- importance(modelrandomt) //Shows the result of importance of variable. //
- varImplot(modelrandomt) // Shows the figure of importance of variables//

By executing the above command, Figure 6.5, shows the variable importance as follows

	MeanDecreaseGini
land	11.27
age	45.97
rev	88.78
nc	49.75
basket	28.22
geslacht	5.06

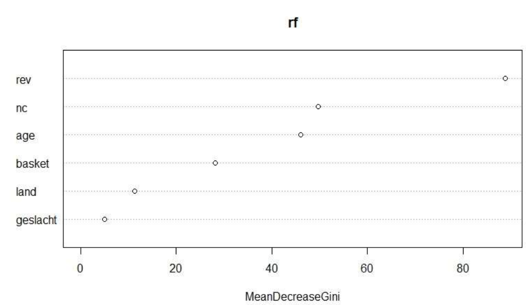


Figure 6.5: Dataset_1_train variable importance

By observation of the figure 6.5, clearly revenue variable has the highest importance followed by nc, age, basket, land and geslacht variables. To predict and test the model in the testing data set the following R command is used:

- Modelrandomtp<-predictt(modelrandomt,Dataset_1_test,type="class")
- t<-table(predictions=Modelrandomtp,actuals=Dataset_1_test$coupongebruikt)
- sum(diag(t))/sum(t)

The command above tests the model created in the training data against the testing data. The results are displayed in the table 6.6 as follows:

	Actuals	
Predictions	0	1
0	44	13
1	17	41

Table 6.6: Random forest prediction

Here, in the above matrix, 44+41=85 times the prediction is made accurately with prediction model created and 13+17=30 times the model predicted wrongly. This means the model has accuracy of 73.91% on the test data set. The table 6.7, gives the summary of the different results model built on the selection of mtry numbers.

Number of variable split tries	**Number of trees plotted**	**OOB**
Mtry not used.	1000	28.02%
1	1000	30.6%
2	1000	29.53%
3	1000	29.53%
4	1000	31.47%
5	1000	32.11%

Table 6.7: Random forest accuracy with different variable split tries

By executing the algorithm number of times, it is best to run the algorithm without selecting the number of variables to split at each try because it gives the higher OOB error rate, that is accuracy of 71.98%. To test the random forest model created on the train dataset the result is given as follows:

	Actuals	
Predictions	0	1
0	44	10
1	17	44

Table 6.8: Random forest confusion matrix on test dataset

The above results show that the algorithm predicted 88 times correctly and 27 times wrongly predicted. The accuracy of the random forest on the retention email dataset is 76.52%

6.5 Algorithm comparison: Decision tree vs Random forest

After applying both the algorithms on the retention email Data_set 1, we can compare the results as shown in the table 6.9, as follows:

Classification Algorithm	**OOB**	**Test data set accuracy**
Random forest	28.02%	76.52%
Decision tree	NA	80%

Table 6.9: Decision tree vs random forest results

Hence, we can conclude that to predict the coupon usage through the decision tree classification algorithm has the highest accuracy on the retention email test data set of 80 % whereas random forest algorithm has 76.52% accuracy.

6.6 Association mining: Market basket analysis on Retention email

To find the interesting relationship of the products associated from the retention email data set, here we will apply the association rule mining algorithm. The retention email data set contains 1363 observations and 2 variables. The first variable is transaction id and the second variable is productnaam. Each row gives the single transactions of the product. The package in R studio to install the association rule algorithm is "arules" and the package to run the graphical representation of the result is "arulesViz".

- Installed.packages("arules")
- Installed.packages("arulesViz")
- library(arules)
- library(arulesViz)

The dataset needs to be converted in to list of transactions, so that R studio can recognize the first column as transactions and second column as product names. In the Dataset_1, there are 941 items and 585 transactions present.

- txns <- as(Dataset_1, "transactions")
- itemFrequencyPlot(txns, topN = 25)

To run the apriori algorithm, we need to define the minimum confidence and support percentages to generate the rules. In the large data set to select the appropriate value of support and confidence values can be very time consuming. In general, it is best to have a visual examination of the dataset for the Top 20 items frequencies plot to have an idea of selecting the support and confidence values to execute the algorithm. In the figure 6.6, we can observe that the most frequent items appeared less than 5% in the total transactions in the retention email dataset.

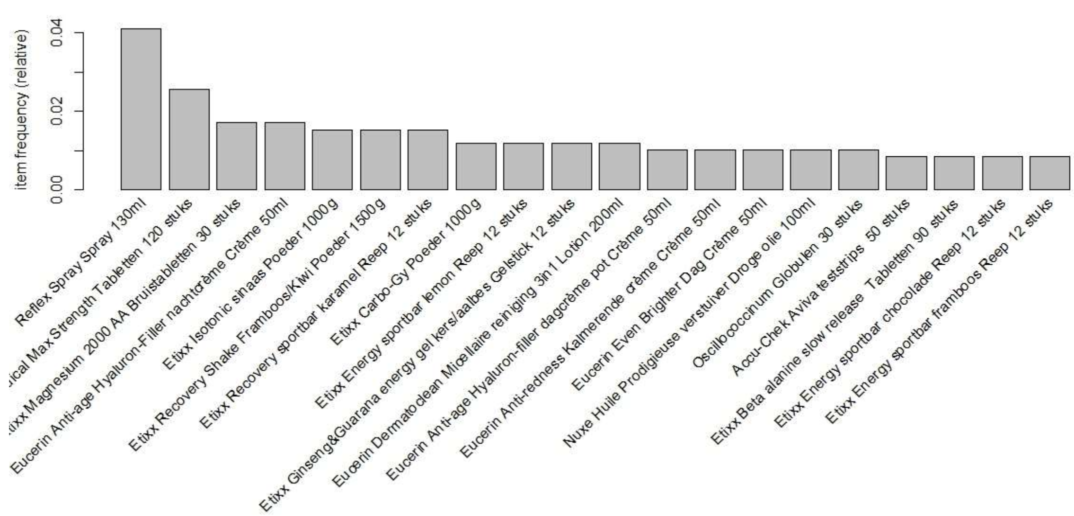


Figure 6.6: Top 20 item frequency plot

By default, R selects the minimum length of product 1 and maximum length of products as 10. To see how many rules are generated from the retention email data set, we select the minimum values of confidence and support as 0.001 %. By executing the R command as follows:

- rules <- apriori(txns, parameter = list(sup = 0.001, conf = 0.001, target="rules"))

The algorithm generated 468,880 set of rules from the data set. The rule length distribution is *given as follows:*

"rule length distribution (lhs + rhs):sizes

1	2	3	4	5	6	7	8	9	10
941	3330	6576	12808	26430	50856	81382	103408	103059	80090"

The above result means that there were 941 rules generated which had only 1 item and 3330 rules generated which had 2 items and so on. To simplify the rules, we set the minimum length of items to 2, support at 0.002% and confidence level at 30%.

- rules <- apriori(txns, parameter = list(sup = 0.002, conf = 0.30, minlen=2 ,target="rules"))

The algorithm generated 90 set of rules from the Dataset_1. The rule length distribution is given as follows:

"rule length distribution (lhs + rhs): sizes

2 3 4

65 21 4 "

The above result means that there were 65 rules generated which had only 2 item and 21 rules generated which had 3 items and 4 rules generated which had 4 items. In this given set of rules, we can sort the rules by lift, confidence or support in decreasing and increasing order by the sort command in R. However, to narrow the rules down, there might be repeated or redundant rules present. To get rid of redundant rules we use following R commands to prune the rules as,

- subset.matrix <- is.subset(rules, rules)
- subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
- redundant <- colSums(subset.matrix, na.rm=T) >= 1
- rules.pruned <- rules[!redundant]
- rules<-rules.pruned
- rules
- plot(rules,col="blue")
- inspect(rules)
- write(rules,file = "rules.csv",sep=",",quote=TRUE,row.names=FALSE)

After pruning the rules, we are left with 41 rules to inspect. The summary is given as follows:

"

rule length distribution (lhs + rhs):sizes

2

41

mining info:

data transactions support confidence

txn 585 0.002 0.3 "

	Support	Confidence	Lift
Min	0.0034	0.4	23.4
1st Quadrant	0.0034	0.5	33.4
Median	0.0034	0.66	55.7
Mean	0.0039	0.71	79.7
3rd Quadrant	0.0034	1	117
Max	0.0084	1	292.5

Table 6.10: Association rules, Summary of quality measures

In figure 6.7, shows the plot of 41 rules which were generated after pruning the rules. The table 6.11, shows the set of top 5 rules which were generated and sorted by support. For the complete 41 association rules refer to appendix 7.

Rules	support	confidence	lift
{Eucerin Anti-age Hyaluron-filler dagcrème pot Crème 50ml} => {Eucerin Anti-age Hyaluron-Filler nachtcrème Crème 50ml}	0.0085	0.833	48.8
{Eucerin Even Brighter Nacht Crème 50ml} => {Eucerin Even Brighter Dag Crème 50ml}	0.0068	1.000	97.5
{Etixx Recovery sportbar karamel Reep 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0068	0.444	28.9
{Etixx Nutritional Energy gel Gelstick 12 stuks} => {Etixx Energy sportbar lemon Reep 12 stuks}	0.0051	0.750	62.7
{Etixx Nutritional Energy gel Gelstick 12 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0051	0.750	48.8

Table 6.11: Sorted by support top 5 rules

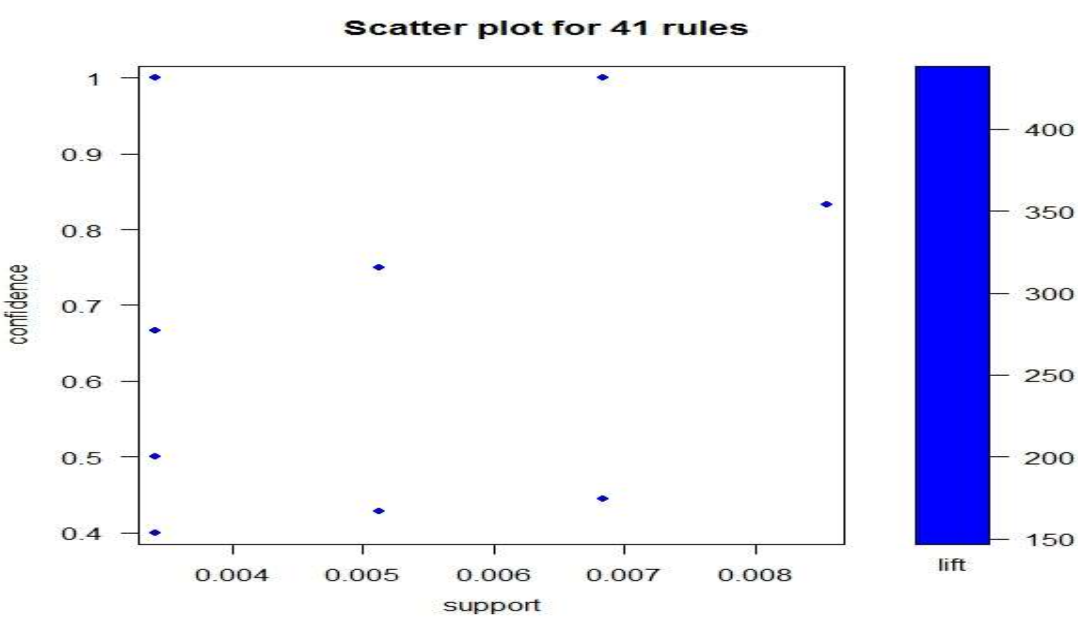


Figure 6.7: Scatter plot association rules for retention email data set

The interpretation of the rules is as follows, in the table 6.11, the rule no 2 is that the customer who bought product Eucerin even bright nacht crème 50ml is most likely to buy Eucerin even brighter dag crème 50ml with the confidence of 70 %. The support indicates, by what proportion of a data the product to be in transaction. For the rule 2 the support is 0.003% and the lift is 131.

Support= number of transaction involving both the products which in this case Eucerin even bright nacht crème 50ml and Eucerin even brighter dag crème 50ml / Total number of transactions.

Confidence= Number of transactions where Eucerin even bright nacht crème 50ml was also bought when Eucerin even brighter dag crème 50ml was also bought /No of transactions where Eucerin even bright nacht crème 50ml was also bought.

Lift= Ratio of observed support by the expected support.

Now that from the retention email data set, we have generated the important association rules , which can be used to promote the products and give coupons accordingly through the rules recommended. If we want to target specific items to generate the rules, let's say that from the figure 6.6, we want to target the highest frequency of products. In this case, we have selected the item Etixx Isotonic sinaas Poeder 1000g s for generating target rules. In total, there were 215 rules generated.

- rules1<-apriori(data=txn, parameter=list(supp=0.001,conf = 0.001,minlen=2,maxlen=4), appearance = list(default="lhs",rhs="Etixx Isotonic sinaas Poeder 1000g"),control = list(verbose=F))
- rules<-sort (rules, decreasing=TRUE,by="confidence")
- inspect(rules[1:5])

"rule length distribution (lhs + rhs):sizes

2 3 4

25 78 112"

In the below Table 6.12, shows the list of transactions before buying the Etixx sinaas powder 1000 g and the table 6.13, shows the list of transactions after buying the Etixx sinaas powder 1000g.

Rules	support	confidence	lift
{Etixx Citrax Poeder 400g} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Recovery Shake chocolade Zakjes 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Starter Pack XS} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Ice Power cold gel Gel 150ml} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Recovery Shake framboos/kiwi Zakjes 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Cycling preparation pack 2} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Isotonic Energy gel lime 40g Gelstick 12 stuks,Etixx Recovery Shake chocolade Zakjes 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Carbo-Gy Poeder 1000g,Etixx Recovery Shake chocolade Zakjes 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56
{Etixx Ginseng&Guarana energy gel kers/aalbes Gelstick 12 stuks,Etixx Recovery Shake chocolade Zakjes 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0017	1	65.56

Table 6.12: List of rules before buying Etixx sinaas 1000g

Rules	support	confidence	lift
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Citrax Poeder 400g}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Recovery Shake chocolade Zakjes 12 stuks}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Starter Pack XS}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Ice Power cold gel Gel 150ml}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Recovery Shake framboos/kiwi Zakjes 12 stuks}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Cycling preparation pack 2}	0.00169	0.111111111	65.555556
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx Recovery Shake chocolade Poeder 1500g}	0.00339	0.222222222	43.703704
{Etixx Isotonic sinaas Poeder 1000g} => {Etixx HMB 1000 Tabletten 60 stuks}	0.00169	0.111111111	32.777778
{Etixx Isotonic sinaas Poeder 1000g} => {Gratis - Etixx Energy sportbar chocolade Reep 1 stuks}	0.00169	0.111111111	32.777778

Table 6.13: List of rules after buying Etixx Sinaas 1000g

We can conclude this section as, in the Table 6.11, the list of rules can be used to offer along with the coupon recommendation based on the list of rules presented. Depending upon the product targeting marketing, as an example given the list of rules in Table 6.12 and 6.13 can be used to offer coupon to the given list of products to customers who bought Etixx sinaas powder 1000g.

7 Chapter 7

7.1 Retention email process through data mining

Finally, to close the loop for the retention email marketing, the general process for data mining for the retention email campaign and to make the data driven decision through data mining techniques for the retention email campaigns in near future for the company the process is shown in the figure 7.1. To measure the effective success of the retention email campaign, where the data mining techniques is applied for the defined respective retention email data can be measure through the count of sale orders and highest profit margin generated through the specific retention email campaigns. This in turn can be compared with the company's policy driven decisions.

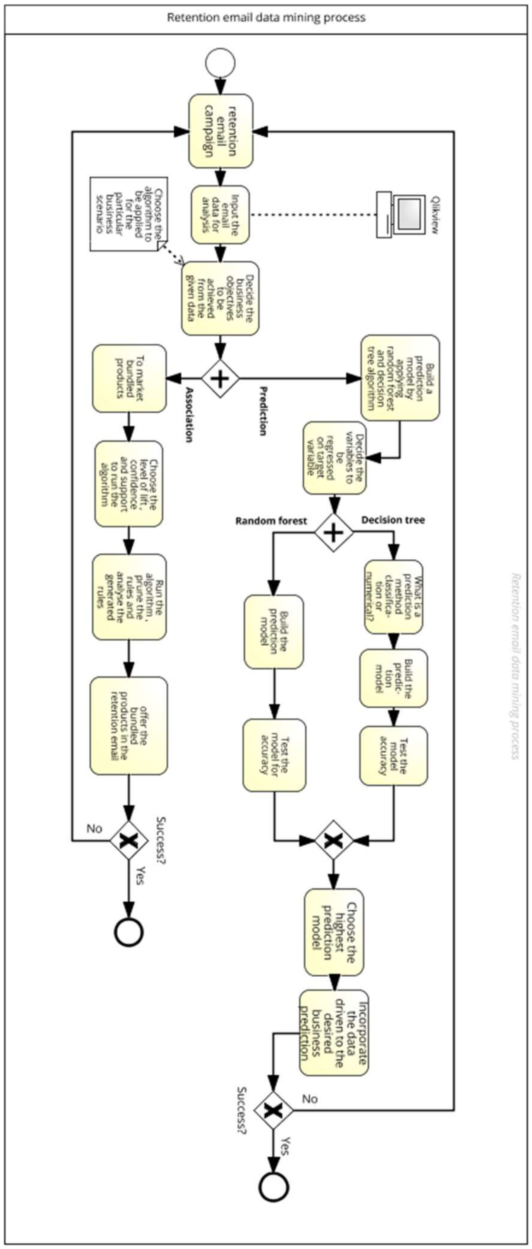


Figure 7.1 : Retention email data mining process

7.2 Recommendation: process model for retention campaign test idea process

Viata online pharmacy had been applying marketing strategies based on their experience and gut feeling. However partial decisions taken in the marketing area were partly driven on the descriptive data. To be more data driven and use the data mining techniques in the future online marketing campaigns , the company can use the test process given in the figure 7.2, for testing a new idea for the retention email campaign.

The process is general and can be used with any campaign idea. For an example, if the company wants to test the email campaign idea for the specific segment of the customers. Let's, assume the customers are only from the Netherlands and next step is to analyse their market basket through the association rules and then send emails with the respective bundles created by the marketers, based on the data. Then execute the campaign and analyse the results. Based on the results the decision taken by the marketers can be non-conclusive or positive or negative. If the decision s non-conclusive then retest the campaign with large set of data or drop the idea. If it is positive then proceed by using the new idea in the company standard email strategy. For negative decision, change the idea and run the test again.

7.2.1 Guidelines to set up a good test

Generally, A/B tests are used widely by marketing professionals to take marketing decisions based on the scenario. This test is a way to determine which variation of the testing ideas have improve conversation rates. Conversion rates in online marketing are generally defined as the total number of users divided by the total number of reacted user on the specific campaign or an idea. The best practice is to run the test with the selection of a large sample size to give conclusive results(Miller, 2010).Time duration must be defined appropriately to run the tests. To evaluate the differences of the test idea and to conclude the results, should be statistically significant at 10% or 5% or 1% level i.e. confidence of 90% or 95% or 99%. Generally, it is best to decide the sample size before executing the tests to have conclusive results.

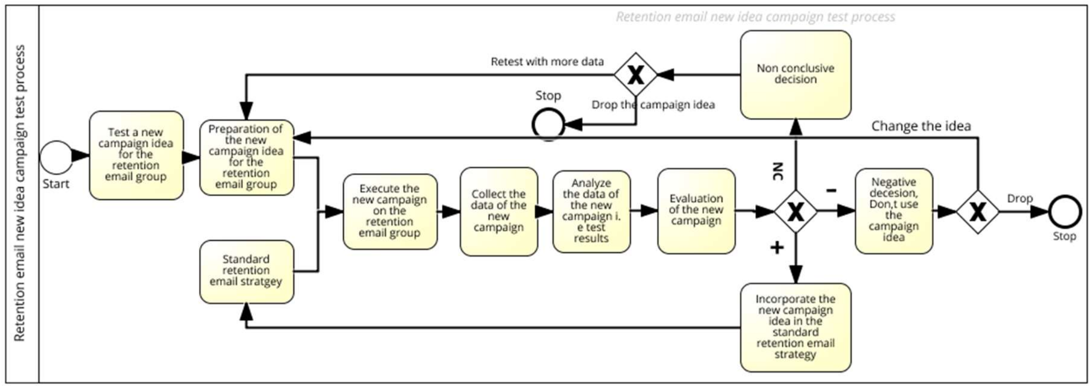


Figure 7.2 : Retention campaign test idea design process

7.3 Conclusion

This master thesis project discussed and presented the Viata online pharmacy business model in detail. Through the understanding of the business model the goal was decided to target the marketing component of the company for further analysis to apply data mining technique. Through presenting the detail process of the marketing component of the company, key areas were identified to apply data mining techniques. Sale order analysis were presented from the retention email process of the company. The result of the sale order there by shows that coupon plays an important variable to apply prediction algorithm. Two prediction algorithms were applied namely decision tree and random forest on the retention email data set. The comparison of the two algorithms were presented and the results were, decision tree algorithm was better in predicting the coupon usage in retention email data set. Finally, association rule mining algorithm was applied to find interesting product association in the retention email data set. As these rules generated and present in this thesis, will be one of the factors to decide on bundles for the marketers in future campaigns. Additionally, to test the data driven campaign ideas a test model was designed for the company to test the impact of new retention data driven ideas. Finally, concluding this thesis, the retention email data mining process has been presented.

References

Chen, Y.-L., et al. (2005). "Market basket analysis in a multiple store environment." Decision support systems **40**(2): 339-354.

Coussement, K., et al. (2014). "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees." Journal of Business Research **67**(1): 2751-2758.

Cuadros, A. J. (2014). "Customer segmentation model based on value generation for marketing strategies formulation/Modelo de segmentación de clientes basado en la generación de valor para la formulación de estrategias de mercadeo/Modelo de segmentação do cliente baseado na criação de valor para a formulação de estratégias de marketing." Estudios Gerenciales **30**(130): 25-30.

Cuadros, A. J. and V. E. Domínguez (2014). "Customer segmentation model based on value generation for marketing strategies formulation." Estudios Gerenciales **30**(130): 25-30.

Dursun, A. and M. Caber (2016). "Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis." Tourism Management Perspectives **18**: 153-160.

Hosseini, S. Y. and A. Ziaei Bideh (2014). "A data mining approach for segmentation-based importance-performance analysis (SOM–BPNN–IPA): a new framework for developing customer retention strategies." Service Business **8**(2): 295-312.

Kashwan, K. R. and C. Velu (2013). "Customer segmentation using clustering and data mining techniques." International Journal of Computer Theory and Engineering **5**(6): 856.

Kashwan, K. R. and C. M. Velu (2013). "Customer Segmentation Using Clustering and Data Mining Techniques." International Journal of Computer Theory and Engineering **5**(6): 856.

Kaur, M. and S. Kang (2016). "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining." Procedia Computer Science **85**: 78-85.

Khajvand, M., et al. (2011). "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study." Procedia Computer Science **3**: 57-63.

Koh, H. C. and G. Tan (2011). "Data mining applications in healthcare." Journal of healthcare information management **19**(2): 65.

Loraine Charlet, A. and A. Kumar (2012). "Market Basket Analysis for a Supermarket based on Frequent Itemset Mining."

Radulescu, A., et al. (2012). "RFM based segmentation: An analysis of a telecom company's customers." Marketing From Information to Decision(5): 52-62.

Rajagopal, D. S. (2011). "Customer Data Clustering using Data Mining Technique."

Torgo, L. and L. Torgo (2011). Data mining with R: learning with case studies, Chapman & Hall/CRC Boca Raton, FL:.

Varma, M. K. S. and N. K. Rao "Comparative Study of Different Data Mining Prediction Algorithms."

Weber, B. G. and M. Mateas (2009). A data mining approach to strategy prediction. Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on, IEEE.

Wirth, R. and J. Hipp (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Citeseer.

Witten, I. H., et al. (2016). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

Zhao, Y. (2012). R and data mining: Examples and case studies, Academic Press.

Ziafat, H. and M. Shakeri (2014). "Using Data Mining Techniques in Customer Segmentation." Journal of Engineering Research and Applications **4**(9): 70-79.

Ziafat, H. and M. Shakeri (2014). "Using Data Mining Techniques in Customer Segmentation." International Journal of Engineering Research and Applications **4**(9): 70-79.

E.Miller (2010)."How not to run an A/B test".

List of Appendices

Appendix 1

The list of “Viata” available data field variable in their database.

Variable name	Variable description
1. Betaald	Payment method used by the customer
2. Campaign	Associated campaign method used
3. Coupon code	Coupon code used by the customer
4. Convprice	Sale price of the product
5. Coupongebruikt	Coupon used flag 0 or 1
6. Coupon naam	Name of the coupon used by the customer.
7. Coupon type	Type of the coupon used by the customer.
8. Customerid	Unique customer id of the customer.
9. Date_birth	Birth date of the customer
10. Email	Email address of the customer
11. Firstorder_order_data	First order date of the customer.
12. Gelascht	Gender of the customer. Male , female.
13. is_firstorder	Flag first order of the customer 0 or 1.
14. Land	Country of the customer
15. Nbr_products	Count of the product bought by the customer.
16. Productnaam	Name of the product.
17. Rfm_frequency, rfm_monitize, rfm_highestsale,rfm_receny, rfm_startdate	RFMHS values of the customers ranked from 0 to 5 scale.
18. S_saleorderid	Saleorder id of the customer
19. Camgaign	Source of campaign used by the customer.
20. Saleorder_netcontr	Customer sale order net contribution generated.
21. Marge	Customers margin
22. Klantid	Unique customer number
23. Stad	Stad of the customer
24. Pakket	Flag the product offered in packet or not
25. Productcat	Categories of the products

Appendix 2

Retention email policy driven decision

Inputs				Outputs
Policy driven category	**Campaign**	**Policy driven RFMHS**	**Spend in €**	**Coupon Type**
Sports and nutrition	Etixx	F1	➢ 30	Etixx Gel
			➢ 40	
			➢ 60	Etixx Bar
			➢ 70	
			➢ 150	Etixx pack
General	Ret_1	F1	➢ 30	Free Shipping
	Generiek	F1	➢ 30	
		H3&F1	➢ 100	
		H4&F1		7 Euro Absolute
		H5		
	Auto campaign	F1	➢ 20	Free shipping
Vicks	Vicks	F1	-	Content
Baby	Baby	F1	➢ 30	Free Shipping
		H3/F1	➢ 30	
		H4/F1	➢ 40	
		H5	➢ 100	7 Euro absolute
Beauty and care	Eucerin	F1	➢ 75	5 Euro absolute
	Verzorging	F1	➢ 30	Free shipping
		H3&F1	➢ 30	
		H4&F1	➢ 100	10 Euro absolute
		H5		
Joint and muscles	Gewrich	F1	➢ 30	Free Shipping
	Reflex spray	F1	-	Content
Hair loss	Hair and nagels	F1	➢ 100	5 Euro absolute
		H4&F1		
		H5		
Slimming	Vermagere	F1	➢ 30	Free shipping
		H3&F1		
		H4&F1	➢ 100	7 Euro absolute
		H5&F2&R1		
Vitamin	Vitamin	F2	➢ 30	Free Shipping
		R1&F1		

Appendix 3

Viata online pharmacy, RFMHS description,

1. Recency- When did customer recently bought the product. From scale 1-5.
   5 - less than 3 weeks.

   4 -more than 3 weeks and less than equal to 2 months.

   3 -more than 2 months and less than equal to 4 months.

   2 -more than 4 month and less than equal to 9 months.

   1 -more than 9 months.
2. Frequency- Total orders over the last 12 months.
   5 -more than 8 orders.

   4 - between 6-8 orders.

   3 - between 4-5 orders.

   2 - between 2-3 orders.

   1 - 1 order only.
3. Monetize- Amount spend over the last 12 months.
   5 - More than 65 Euro.

   4 - More than 55 Euro and less than and equal to 65 Euro.

   3 - More than 45 Euro and less than and equal to 55 Euro.

   2 - More than 30 Euro and less than and equal to 45 Euro.

   1 - less than and equal to 30 Euro.
4. Highest sales – Highest amount of sales in the last 12 months.
   5 - More than 150 Euro.

   4 - More than 75 Euro and less than and equal to 150 Euro.

   3 - More than 45 Euro and less than and equal to 75 Euro.

   2 - More than 10 Euro and less than and equal to 45 Euro.

   1 - less than and equal to 10 Euro.
5. Start- When did the customer order for the first time.
   5 - More than 24 months.

   4 - More than 12 months and less than and equal to 24 months.

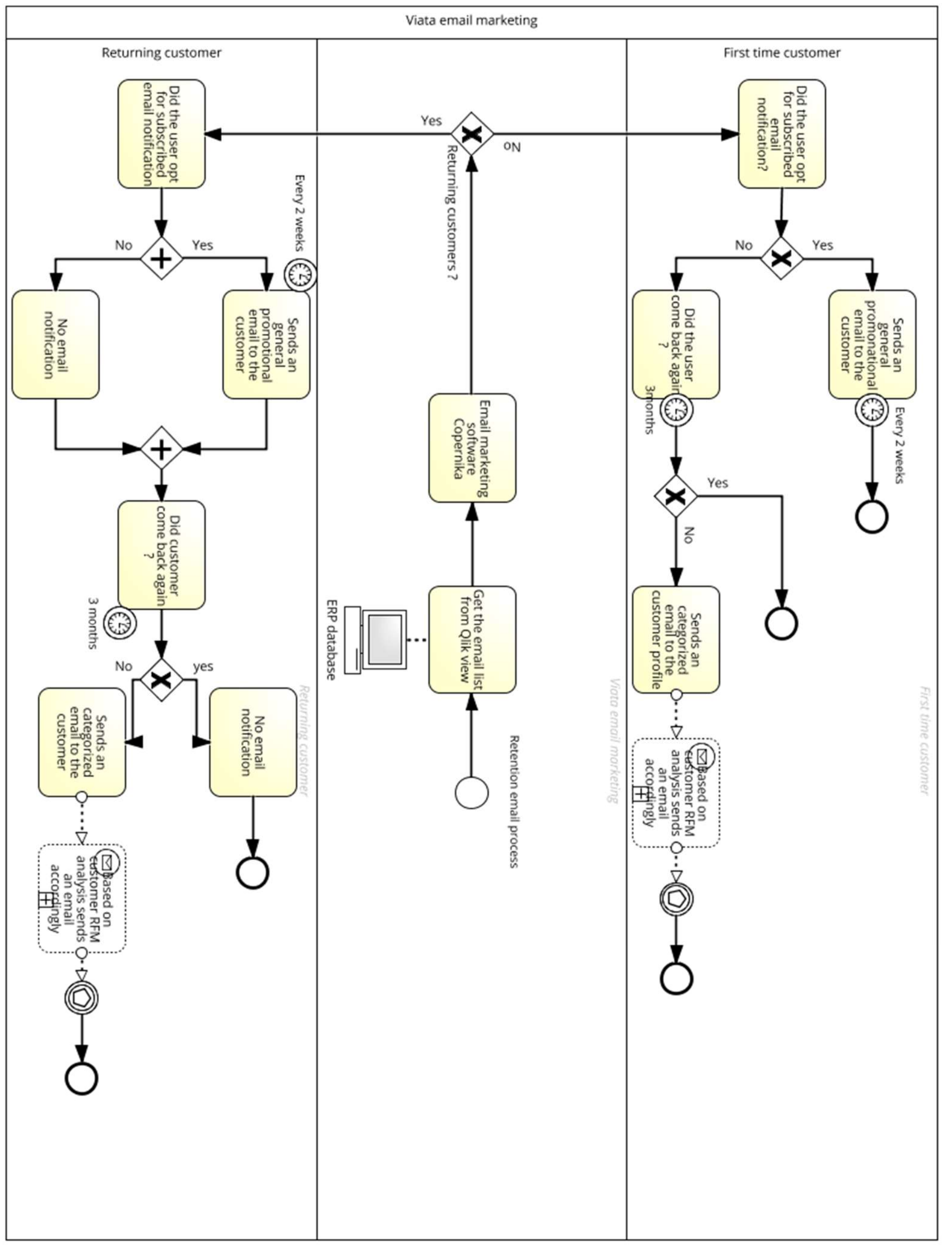
   3 - More than 6 months and less than and equal to 12 months.

   2 - More than 3 months and less than and equal to 6 months.

   1 - less than and equal to 3 months.

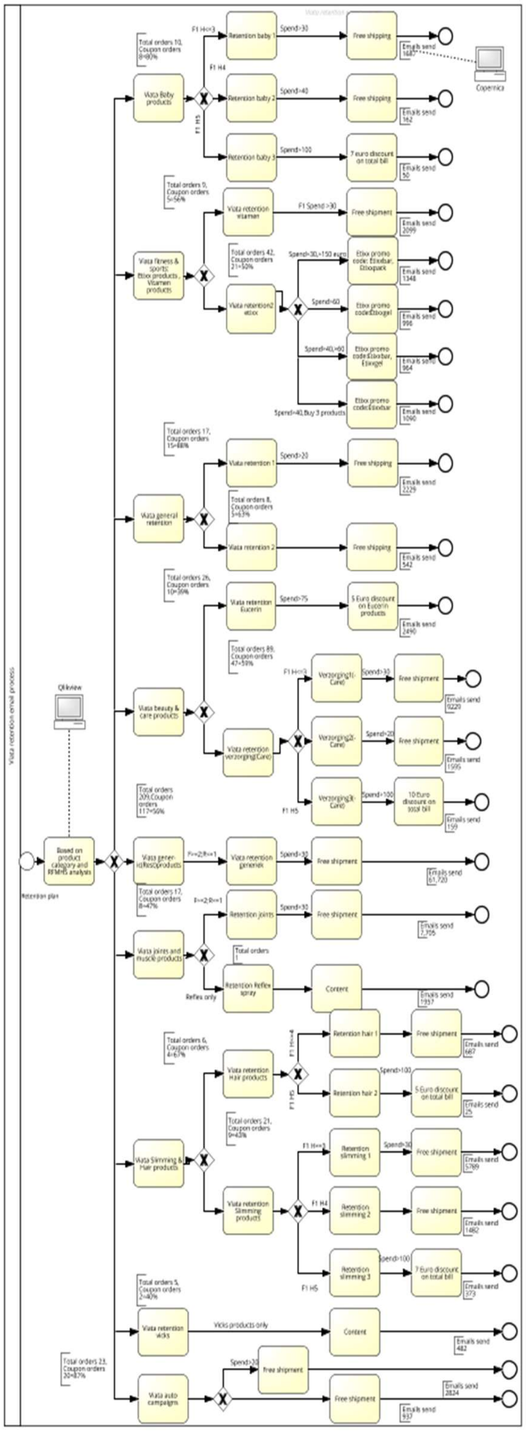
Appendix 4

Viata complete email marketing process



Appendix 5

Viata retention plan decision logic



Appendix 6

Below table gives the list of coupon condition which were used in various retention email campaigns.

Retention email	Coupon conditions
1.Etixx	Spend>30,>150,>60,>40,>70 Etixx gel/bar/pack,buy 3 Etixx product
2.Ret_1	Spend>20 free shipping
3.Baby	Spend>30 Free shipping, Spend>40 Free shipping, Spend>100 7euroDiscount
4.Eucerin	Spend>75 5euro discount Eucerin product
5.Generiek	Spend>30 Free shipping
6.Gewrich	Spend>30 Free shipping
7.Hair	Spend>100 5 euro discount
8.Vermagere	Spend>30 Free shipping, Spend>100 7 euro korting
9.Verzorging	Spend>30 Free shipping, Spend>100 10 euro korting
10.Vitamin	Spend>30 Free shipping
11.Reflex spray	Content
12.Vicks	Content
13.Auto campaign	Spend>20 free shipping

Appendix 7

The below table shows the list of association rules generated from the apriori algorithm. These rules are from the retention email Dataset_1.

rules	support	confidence	lift
{Eucerin Anti-age Hyaluron-filler dagcrème pot Crème 50ml} => {Eucerin Anti-age Hyaluron-Filler nachtcrème Crème 50ml}	0.0085	0.833	48.8
{Eucerin Even Brighter Nacht Crème 50ml} => {Eucerin Even Brighter Dag Crème 50ml}	0.0068	1.000	97.5
{Etixx Recovery sportbar karamel Reep 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0068	0.444	28.9
{Etixx Nutritional Energy gel Gelstick 12 stuks} => {Etixx Energy sportbar lemon Reep 12 stuks}	0.0051	0.750	62.7
{Etixx Nutritional Energy gel Gelstick 12 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0051	0.750	48.8
{Etixx Ginseng&Guarana energy gel kers/aalbes Gelstick 12 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0051	0.429	27.9
{Etixx Ginseng&Guarana energy gel kers/aalbes Gelstick 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0051	0.429	27.9
{Etixx Energy sportbar lemon Reep 12 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0051	0.429	27.9
{Etixx Energy sportbar lemon Reep 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0051	0.429	27.9
{Eucerin Herstellende zalf 10% urea Zalf 100ml} => {Eucerin Hyal-Urea nachtcrème Crème 50ml}	0.0034	1.000	117.0
{Modifast Protiplus Repen melkchocolade/karamel Reep 6x27g} => {Dettol medical ontsmettingsmiddel Chloroxylenol 4,9% Oplossing 5l}	0.0034	1.000	146.3
{Eucerin UltraSensitive reinigingslotion Lotion 100ml} => {Eucerin Anti-redness Kalmerende crème Crème 50ml}	0.0034	1.000	97.5
{Etixx High protein sportbar cocos-vanille Reep 12x50g} => {Etixx Energy sportbar lemon Reep 12 stuks}	0.0034	1.000	83.6
{Sofraline Neusspray 15ml} => {Sofrasolone Neusspray 10ml}	0.0034	1.000	195.0
{Eucerin Dermopurifyer Hydraterende matterende crème Crème 50ml} => {Eucerin Dermopurifyer reinigingsgel Gel 200ml}	0.0034	0.667	97.5
{Infacol baby Oplossing 50ml} => {Nutrilon Omneo 1 Poeder 800g}	0.0034	1.000	292.5
{Etixx Recovery sportbar pinda Reep 12 stuks} => {Etixx Energy sportbar framboos Reep 12 stuks}	0.0034	1.000	117.0
{Bioderma Sensibio H2O Micellaire oplossing 100ml} => {Bioderma Sensibio H2O micellair water Micellaire oplossing 500ml}	0.0034	1.000	146.3
{Eucerin Dermopurifyer scrub Scrubcrème 100ml} => {Eucerin Dermopurifyer reinigingsgel Gel 200ml}	0.0034	1.000	146.3
{Etixx Isotonic Energy gel lime 40g Gelstick 12 stuks} => {Etixx Ginseng&Guarana energy gel kers/aalbes Gelstick 12 stuks}	0.0034	0.667	55.7
{Etixx Isotonic Energy gel lime 40g Gelstick 12 stuks} => {Etixx Carbo-Gy Poeder 1000g}	0.0034	0.667	55.7
{Gratis - Etixx Energy sportbar chocolade Reep 1 stuks} => {Etixx Magnesium 2000 AA Bruistabletten 30 stuks}	0.0034	1.000	58.5
{Eucerin Verzachtende gezichtscrème urea 5% Crème 50ml} => {Eucerin Lip Repair Balsem 10ml}	0.0034	0.667	78.0
{Sterimar Neushygiëne Verstuiver 100ml} => {Sinutab 500/30mg Tabletten 15 stuks}	0.0034	0.667	130.0
{Eucerin Dermopurifyer reinigingsgel Gel 200ml} => {Eucerin Anti-age Hyaluron-Filler nachtcrème Crème 50ml}	0.0034	0.500	29.3
{Eucerin pH5 douche olie pomp Doucheolie 1l} => {Eucerin Anti-age Volume-filler dagcrème normale huid Crème 50ml}	0.0034	1.000	195.0
{Eucerin DermoCapillaire Anti-roos crème droge roos Shampoo 250ml} => {Eucerin Lip Repair Balsem 10ml}	0.0034	1.000	117.0
{Etixx Isotonic lemon Poeder 1000g} => {Etixx Magnesium 2000 AA Bruistabletten 30 stuks}	0.0034	0.667	39.0
{Etixx Isotonic lemon Poeder 1000g} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0034	0.667	43.3
{Etixx Beta alanine slow release Tabletten 90 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0034	0.400	26.0
{Etixx Recovery Shake chocolade Poeder 1500g} => {Etixx Magnesium 2000 AA Bruistabletten 30 stuks}	0.0034	0.667	39.0
{Etixx Recovery Shake chocolade Poeder 1500g} => {Etixx Isotonic sinaas Poeder 1000g}	0.0034	0.667	43.3
{Etixx Energy sportbar sinaasappel Reep 12 stuks} => {Etixx Magnesium 2000 AA Bruistabletten 30 stuks}	0.0034	0.667	39.0
{Etixx Energy sportbar sinaasappel Reep 12 stuks} => {Etixx Recovery sportbar karamel Reep 12 stuks}	0.0034	0.667	43.3

{Eucerin Anti-age Volume-filler nachtcrème Crème 50ml} => {Eucerin Anti-age Volume-filler dagcrème normale huid Crème 50ml}	0.0034	0.500	97.5
{Strepsils + lidocaïne Zuigtabletten 36 stuks} => {Vicks Vaporub Zalf 100g}	0.0034	1.000	195.0
{Etixx Nutritional Energy gel Gelstick 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0034	0.500	32.5
{Etixx Energy sportbar framboos Reep 12 stuks} => {Etixx Carbo-Gy Poeder 1000g}	0.0034	0.400	33.4
{Etixx Energy sportbar framboos Reep 12 stuks} => {Etixx Energy sportbar lemon Reep 12 stuks}	0.0034	0.400	33.4
{Etixx Energy sportbar framboos Reep 12 stuks} => {Etixx Magnesium 2000 AA Bruistabletten 30 stuks}	0.0034	0.400	23.4
{Etixx Energy sportbar framboos Reep 12 stuks} => {Etixx Isotonic sinaas Poeder 1000g}	0.0034	0.400	26.0

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
"Viata online Pharmacy" Retention email Process modelling and application of data mining techniques using R

Richting: **Master of Management-Business Process Management**
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,

Khan, Saud

Datum: **22/05/2017**