MASTER THESIS

TOWARDS A HYBRID CUSTOMER CLASSIFICATION AND CLUSTERING SCHEME FOR EFFECTIVE TARGET-GROUP BASED MARKETING

Amit Das

FACULTY
Faculty of Electrical Engineering, Mathematics & Computer Science

STUDY PROGRAMME
Master's in Business Information Technology

GRADUATION COMMITTEE
Dr. Maya Daneva, Assistant Professor, University of Twente.
Dr. Adina Aldea, Assistant Professor, University of Twente

COMPANY SUPERVISORS
Christian Vonk (La Place)
Dr. Felix Janszen (E-Tail Genius)

September 2019

UNIVERSITY OF TWENTE.

Preface

The last 21 months of my life have been memorable. Right from the time I moved to the Netherlands to pursue my Masters' at the University of Twente, life has taught me a lot. A new country, diverse people from different nationalities and culture, different work ethics, and much more. I started studying at the University of Twente and I am very glad that I chose this university. This university has given me the opportunity to grow as a person, both personally and professionally. Right from completing my courses in the 1st year of Masters', working part time at the University with Career Services and then moving to Utrecht to conduct my Master thesis at La Place, it has been an eventful phase of life. I have probably learned more in this phase of life than I can otherwise recall. With mixed emotions, I am towards the completion of this phase of my life which concludes with the completion of my Master thesis.

I would like to take this opportunity to thank a few people and show my gratitude towards them, without whom this phase would have been incomplete. Firstly, I would like to thank Christian Vonk, my supervisor at La Place, who gave me the opportunity to conduct my Master Thesis amidst a lot of changes within the organization. His direction, guidance and valuable feedback is much appreciated. It was my first experience working at a Dutch company and he made sure I was comfortable and guided me whenever I was lost and looking for ideas. Secondly, I would like to thank my supervisors at my hiring company, Prof Dr. Felix Janszen and Arian Oosthoek for their guidance and support during this process. I am glad that they thought I was capable enough of working at one of their esteemed clients, La Place. Lastly, I would like to thank my University supervisors, Dr. Maya Daneva and Dr. Adina Aldea, for being there with me in this journey of 9 months of completing my Master thesis. They helped me with valuable inputs and feedback throughout the journey of my Master thesis. Meetings with them were always helpful and had a calming effect.

Working at Career Services formed an integral part of my university life. I would especially like to thank Jacqueline for being a great support, not only at work but also personally. Hemo and Selene were great colleagues to work with and I will cherish the memories.

Any journey is incomplete without friends. I made some amazing new friends with whom we did a lot of groupwork in courses. A special mention to Nivedita, Arief, Yaumi, Aimana, Thomas, Dylan, Rob, Dico, Joep, Egle and Teresa, with whom I smoothly carried out and completed my courses. I made a lot of Indian friends here with whom I spent memorable time playing cricket and having some nice food. I am also thankful to Sameeksha, Akshath, Sujith, Sujan, Rakshith, Siddharth, Anish, Dwara, Jathin and Amit for the memories.

On a finishing note and most importantly, I would like to thank my family and friends back home, who are like a lifeline when you live so far away. They had the belief in me that I could succeed and hopefully I did justice to their trust in me. Thank you all!

Management Summary

The HORECA industry in the Netherlands is competitive. La Place is known to be one of the most popular restaurant chains in the Netherlands, offering customers a typically 'Dutch' experience, amongst other cuisines. With a range of restaurants and food joints for customers to visit, it is difficult to keep customers loyal. One way of encouraging customers to visit is to offer them incentives in terms of promotions and deals. However, regular promotions and deals are detrimental to company's revenues. La Place is very sensitive in terms of pricing of its products and has a relatively high cost of operation because of the freshly prepared food, high quality ambience and relatively high cost of labor. So, regularly offering discounts and reducing product prices is not the most sustainable option. There is a need to be more precise with the way marketing and promotions are done. In addition to an 'intended for all' strategy, strategies must be designed to incentivize the loyal customers of La Place. This would not only help maintain their loyalty but spreads goodwill and a good word-of-mouth, possibly attracting more customers. Moreover, customers who have fallen out of the loyalty programme or are likely to fall out, need to be paid special attention and care to encourage them to return. This would be beneficial for both La Place and its customers.

Purpose: The goal of this study is to analyze the potential of using machine learning techniques like classification and clustering for grouping customers based on upcoming promotion and deals. Special attention to loyal customer groups would reduce the likelihood of them falling out and thereby, add to the turnover of La Place.

Methodology/ Approach: A structured literature review was conducted prior to conducting this research. Based on that, the LRFM model and K-means clustering were found to be popular and widely used techniques for customer classification and clustering, respectively. A generic method is proposed and designed to classify and cluster customers. This method is them implemented through a case study within La Place. There were 7 main classes of customers defined based on the recency of their first/last transaction. Results from the case study were analyzed to check for the effectiveness of the method. To validate the prototype and receive suggestions for improvement, an evaluation process was conducted with two experts who have some knowledge about applications of machine learning and data science in business domains. Based on their feedback and overall review, a discussion is made about the usefulness of this research and its limitations. Finally, suggestions for future work are proposed to La Place with a view of implementing this within their organization.

Results: The customer and transactional data analysis was done for the period between June-August 2019. Different customer groups were revealed based on the recency of their transactions. Surprisingly, it was found that more than half the customers of the loyalty programme haven't visited La Place for more than 2 years. These were termed as 'Lost' customers. Almost 30% of the customers hadn't visited La place during this period. This meant that during this period, only about 15% customers of the entire customer base visited La Place. Around 12% were frequent visitors, 1.5% were new customers and the remaining 1.5% were unexpected customers as they were believed to be lost.

Around 15% of the overall customers are termed as 'Monitor' group. They are customers who did not visit during this period but have visited within the last 365 days. Around 15% of the customers are termed as 'Inactive'. They are customers who haven't visited in the last 365 days but had visited a year ago. Some customers turned up unexpectedly during this period. They were earlier either 'Lost' customers or 'Inactive' customers. This group is termed as 'Redormed'. All customer groups identified are mutually exclusive and independent of each other.

Within each of these groups, customers were found to have a preference to visit La Place on a specific part of the day or a specific day of the week. Majority of the customers within these groups did not have specific preferences. For such groups without a preference, clustering was used to find patterns based on their transactional history. Based on knowledge about the clustering variables, inferences were drawn about the identified clusters having specific preferences for part of the day or day of the week.

Eventually, an approach was shown how random forest classifiers could be used to predict future visitors based on their defined labels. This would need training the model with historic data. The model would learn based on the data and then predict who would be the future visitors based on historical data.

Recommendations:

Almost half of the total revenue was generated from the frequent visitors, emphasizing that special attention must be paid to these customers. Frequent visitors should be encouraged to keep visiting La Place and all attempts must be made to ensure that they do not fall out. They were also the most loyal customers in terms of their association with La Place. This should be honored, and appropriate incentives should be offered to them in terms of promotions/deals different from others. New customers should be encouraged to keep visiting La Place regularly and possibly turn them into frequent visitors. The reason why the 'Monitor' group of customers stopped visiting must be tried to be found out based on their past experiences with La Place. This

is a big group and efforts should be made to revive this group and turn them into frequent visitors. On the other hand, they can also become inactive and care must be taken to avoid this from happening. It is alarming to see a substantially large 'Inactive' group. They have high likelihood of turning into 'Lost' customers. It might already be too late to revive them but based on their past opinions, lessons must be learnt to be avoided in the future, and efforts should be made to try and make them visit again. 'Redormed' customers are probably one-off visitors but the fact that they returned after a long time must be appreciated.

In terms of improving the proposed classification method, several steps can be taken. This includes suggestion like inclusion of product level preferences for each customer, inclusion of other attributes like the preferred type of store and the location of store (highway/city), analysis of customer feedback and survey responses about their experience with La Place, sentiment analysis (Big Data) of social media data about the opinion of people about La Place, monitoring actual in-store data for a defined period. The proposed method does have some limitations which can be overcome with future improvements. The validity and reliability of the proposed method can only be verified once it has been put into practice and analyzing its results.

This research is the beginning of something innovative to build stronger relationships with customers. Continuous work needs to be done in order to improve this so that it benefits La Place in the long run.

Contents

Chapter 1. INTRODUCTION 9
1.1 Problem Identification and Motivation 12
1.2 Research context 12
1.3 Research Objective 13
1.4 Research Question 14
1.5 Research Methodology 15
1.6 Structure of the report 17
Chapter 2. LITERATURE REVIEW 18
2.1 Classification Techniques 21
2.2 Classification Techniques 23
Chapter 3. DESIGN 26
3.1 Models, algorithms and techniques 26
3.1.1 The LRFM Model 26
3.1.2 The K-means clustering algorithm 27
3.1.3 Random Forest Classifier 28
3.2 Design of the proposed method 29
3.3 Design choices, assumptions and constraints 30
3.4 Tools 31
Chapter 4. CASE STUDY 33
4.1 About the Company 33
4.2 The model 34
4.3 Tools 37
4.4 The Tableau version of the K-means clustering algorithm 39
4.5 The Tableau Prep flow 42
Chapter 5. RESULTS 44
5.1 Analysis of the groups 46
5.1.1 Frequent Visitors 47
5.1.2 Monitor Group 52
5.1.3 Redormed Group 56
5.1.4 New Customers 59

Chapter 6. EVALUATION 63
6.1 Setup 63
6.2 Respondents 63
6.3 Questions setup for evaluation 64
6.4 Evaluation Results 64
6.4.1 Qualitative results 64
6.4.2 Quantitative results 65
Chapter 7. CONCLUSION AND DISCUSSION 67
7.1 Conclusions 67
7.2 Discussion 71
7.2.1 Classification vs Clustering OR Both combined? 71
7.2.2 What's in it for the Marketing team? 72
7.2.3 Contributions 73
7.2.4 Limitations 74
7.2.5 Validity and Reliability 75
7.3 Recommendations for future work 75
7.4 Advice to La Place 76
References 78
APPENDIX 81
Appendix A. Details about K-means clustering algorithm 81
Appendix B. Tableau Prep Flow 85
Appendix C. Additional Tables 85
Appendix D. Code for Random Forest 85
Appendix E. Questionnaire used for Evaluation 86

LIST OF FIGURES

Figure 1 Classification framework for data mining techniques in CRM (cited from [2]) 10
Figure 2 Design Science Research Methodology (DSRM) Process Model 16
Figure 3 Flowchart of the Grounded Theory Results 21
Figure 4 Proposed method 30
Figure 5 Generic Data flow architecture 32
Figure 6 Data flow architecture for La Place 39
Figure 7 Process diagram of Tableau Prep Flow 44
Figure 8 Customer Classification Overview 45
Figure 9 Revenue split based on Recency 46
Figure 10 Frequent Visitors Overview 47
Figure 11 Frequent visitors - No daypart preference 48
Figure 12 Summary statistics - Clustering Frequent visitors without a daypart preference 49
Figure 13 ANOVA statistics - Clustering Frequent visitors without a daypart preference 49
Figure 14 Frequent visitors - No day preference 51
Figure 15 Summary statistics - Clustering Frequent visitors without a day preference 51
Figure 16 ANOVA statistics - Clustering Frequent visitors without a day preference 52
Figure 17 Monitor Group Overview 53
Figure 18 Monitor Group - No daypart preference 54
Figure 19 Summary statistics - Clustering Monitor group without a daypart preference 54
Figure 20 Monitor Group - No day preference 55
Figure 21 Summary statistics - Clustering Monitor group without a day preference 55
Figure 22 Redormed Group Overview 56
Figure 23 Redormed Group - No daypart preference 58
Figure 24 Summary statistics - Clustering Redormed group without a daypart preference 58
Figure 25 Redormed Group - No day preference 59
Figure 26 Summary statistics - Clustering Redormed group without a day preference 59
Figure 27 New customers Overview 60
Figure 28 New customers - No daypart preference 61
Figure 29 Summary statistics - Clustering New customers without a daypart preference 61
Figure 30 New customers - No day preference 62
Figure 31 Summary statistics - Clustering New customers without a day preference 62
Figure 32 Sample output from Random forest classifier 70

LIST OF TABLES

Table 1 Definitions of LRFM model 27
Table 2 Definitions of the proposed model 36
Table 3 Recency matrix based on Transaction history 37

Chapter 1. INTRODUCTION

Customers are the most valuable asset of any business and are at the heart of customer relationship management (CRM). CRM is viewed as a comprehensive process of acquiring and retaining customers, and with the help of business intelligence, tries to maximize the customer value to the organization [1]. CRM is mainly comprised of a set of processes and enabling systems that support a business strategy to build long term and profitable relationships with specific customers [2]. Any CRM strategy could be viewed as a closed cycle consisting of four different dimensions; customer identification, customer attraction, customer retention, and customer development [1]. An effective classification or segmentation technique can be a powerful way of knowing customers [3]. Data about customer behavior and/or demographics can be used to divide customers into different segments. Such classification techniques can be used in targeting promotions and recommending the right products to the customers. Price promotions influence consumers' experience positively when certain sales bargains are made, since some consumers value unexpected and spontaneous discounts and sales promotions [4]. Rather than targeting all the customers for all types of company promotions, which does not result positive response, there is always an enthusiasm as who are the people who should be targeted for the specific offer. In target marketing, it has been an arduous task to single out customers who are likely to be fascinated for a new product or service. At this juncture, data mining techniques can be applied for filtering out the target customers out of the pool. This would increase the overall effectiveness of the marketing campaign [5].

Data mining techniques are extensively used for classification and pattern extraction from customer data which is very important for business support and decision-making. Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions. Classification and clustering (segmentation) are two of the most important techniques used in marketing and customer-relationship management to understand customer groups [6]. A typical classification framework for data mining techniques in CRM can be seen in Figure 1 [2]. This describes four main steps related to customer relationship management, namely; customer identification, customer attraction, customer retention and customer development. It can be clearly seen that customer identification is the first step. Target customer analysis and customer segmentation are at the core of this step as it is most important to know who your customers are and analyze their behavior. Classification and clustering are two commonly used data mining techniques associated with customer identification.

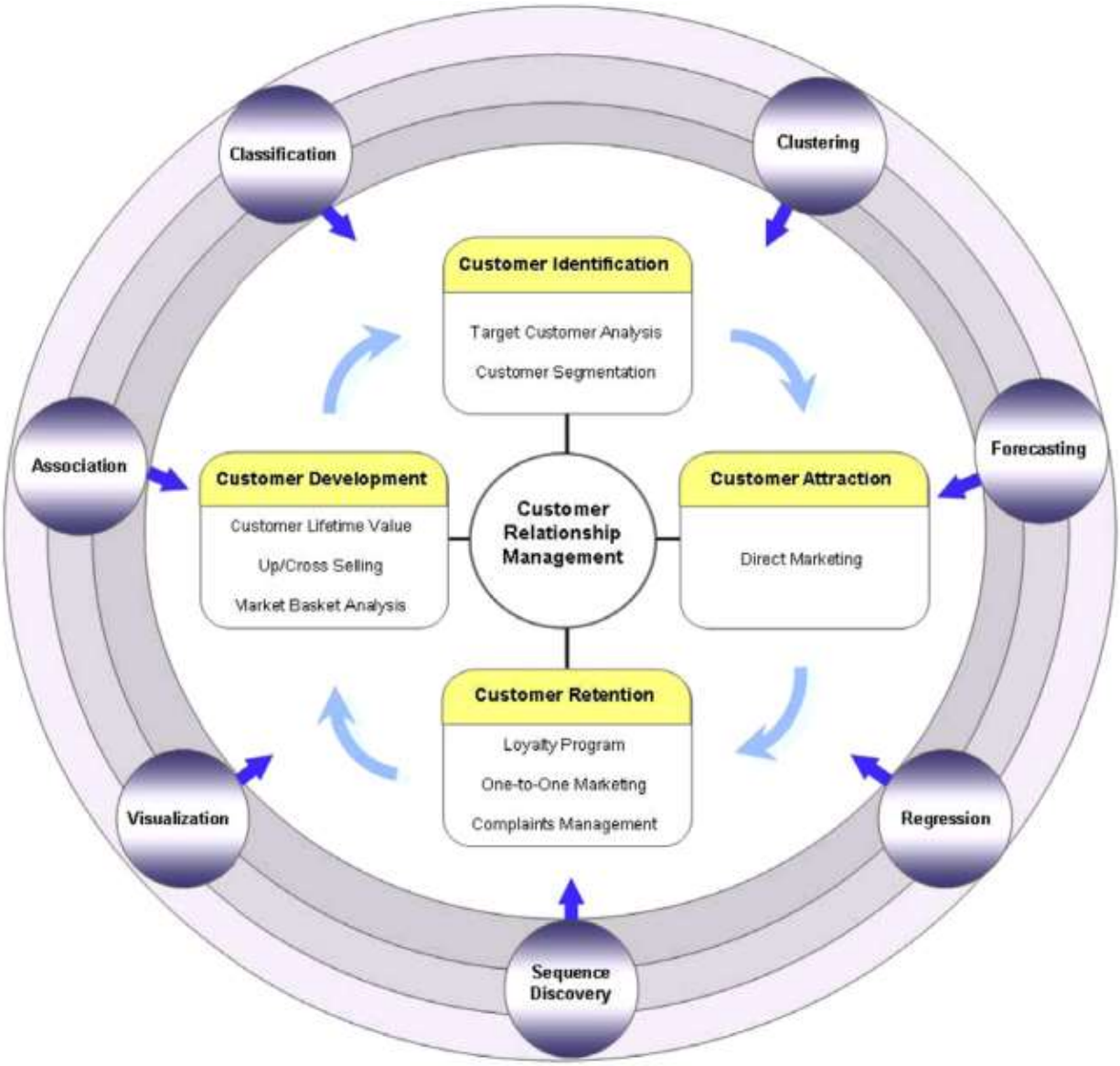


Figure 1 Classification framework for data mining techniques in CRM (cited from [2])

Customer classification and clustering enable the firms to group similar customers together and help managers to better understand the customers' needs; because it is much easier to identify and analyze the characteristics of customer groups rather than studying each customer individually [7]. Dividing the customer base into homogenous groups enables them to deploy different marketing campaigns according to the characteristics of that group [8]. Success of a company depends on its ability to build and maintain loyal and valued customer relationships [9].

By dividing their customers into different clusters, firms can better decide how to effectively allocate their limited resources to different groups of customers based on their value. Also, by

using customer clustering techniques, firms can effectively design their customer retention strategies and maximize their overall profitability [7]. Identifying different groups of customers and their needs can lead to customer satisfaction, which in turn contributes to customer loyalty [7]. When the customers discontinue doing business with a company and move to its competitors, various negative consequences such as losing current revenues due to discontinued business relations or loss of good reputation and credibility could be expected in the long term. This loss of credibility can lead to a loss of current and potential customers' trust in products and services. In general, the probability of successfully selling a product/service to current active customers is roughly 60-70 percent, while this probability is only 5-20 percent for prospective customers [10]. It is more beneficial to identify key customers and retain them rather than acquiring new customers to fill the empty place of those who have decided to discontinue doing [11]business with the organization. This is mainly because the cost of new key customer acquisition is five times more than the cost of current customer retention [7]. Many firms understand the Pareto principle, that 20% of the customer base generates 80% of the profits, but the task of customer segmentation is to find out who belongs to the 20% [3]. If the customer loyalty is enhanced, customer life-cycle will be likely optimized, and eventually the firm will become more profitable. Customers will be more satisfied and a commitment towards relationship with customers would most likely have a positive impact on the firm's brand loyalty and awareness.

Customer segmentation is defined as the process of dividing the entire customer base of a company or the market into smaller groups to make marketing actions more effective and concise. Each segmented group is a smaller segment of the total customer base with similar characteristics. Cluster analysis can be used to create customer segments. Customer are then mapped onto these segments. This makes it possible to increase targeting effectiveness of marketing promotions and to improve response to changing needs [12]. Clustering is an unsupervised learning technique for grouping similar data points. A clustering algorithm assigns many data points to a smaller number of groups such that data points in the same group share the same properties while, in different groups, they are dissimilar. In general, customer segmentation is defined as the process where customers of an enterprise are divided into groups based on their purchasing behavior and characteristics [1]. Segmentation based on a combination of customers' geographic, demographic, and behavioral attributes is considered a better approach compared to the use of a single category of attributes [1].

Classification, on the other hand is a supervised learning technique. It corresponds to the unsupervised procedure of clustering but mainly differs considering that in classification, a training data set of correctly identified observations is already available. Customer classification requires the selection of a set of features or attributes of customer data. The results of customer classification will be dependent on this set of selected attributes. It is one of the widely used

machine learning techniques. Classification is the problem of identifying to which category does an object belong. This assigning of the object to a specific category is based on similarity between the objects in that category, like clustering. Classification is an example of pattern recognition and is used in a variety of applications like finding which emails should be classified 'spam' or as 'non-spam'. Classification is one of the most widely used techniques as it helps to identify customers who are likely to be fascinated by a new product or service. Some of the common customer classification techniques include Naive Bayes Classification, Decision Tree, Artificial Neural Network (ANN), J 48 graft, LAD Tree, Radial Basis Functional Network, Multilayer Perception Neural Network, Ross Quinlan decision tree model (C5.0), C 4.5 Decision tree Algorithm [5] [13]. Out of the several standard techniques, selecting the best classification technique is a major task in data mining. As one size does not suit for all, likely one technique does not produce better yield for all types of data set [5].

1.1 Problem Identification and Motivation

Existing research suggests that companies either do not reveal how to target customers based on their transactional behavior or there isn't enough information available how this can be done. With the help of well-known machine learning techniques like classification and clustering, it is possible to group customers as per their preferences. This helps in enabling different communication means with customers informing them about what is the best for them and thereby maintaining better relationships. Customers will in turn feel more valued and this will imminently increase the company turnover and possibly also the customer base. After all, CRM is mainly about intensifying and solidifying the relationship between the company and its customers, ideally in a one-to-one relationship. However, with a large customer base, it becomes almost impossible to keep such personalized relationships. This is where customer classification can help by grouping similar customers together and then designing ways of reaching out to them specific to their characteristics. It is vital that the number of these groups should be manageable for the company so that special attention can be given to each group and different strategies can be made to get the best out of each group. Overall, this will be a win-win situation for both parties involved, the customers and the company.

1.2 Research context

A generalization may be formal or informal, expressed in words or diagrams, may be known to be true often but false sometimes, and may not all be connected deductively [11] [14]. This research intends to propose a generalizable method which can be adopted by companies to

group customers using classification and clustering techniques. The method is designed to assist companies with the absence of a CRM system or similar advanced customer analysis tools. Moreover, it is also possible that the findings of this research may be add value to those offered by currently existing CRM systems. The purpose of this research is not to replace any existing system, but to work as an aid to managers and analysts in identifying their most valuable customers. The proposed method is meant to be generalizable to companies which store transactional data of customers, mostly from the HORECA industry. However, certain parts of this research would remain company specific such as the available tools and techniques depending upon their availability, expertise and the need of using them.

In order to ensure that the proposed method can be effective, a case study is conducted [14]. The case study was conducted at La Place, a popular Dutch restaurant chain. Details of can be found in Chapter 4. The adoption of the method proposed for La Place remains inspired from the generalizable method proposed in Chapter 3. However, certain aspects like the components of data flow architecture and chosen techniques for future predictions explained in Chapter 3 are robust and will tend to be more company specific.

1.3 Research Objective

Many small to medium-sized companies find it difficult to understand where, why and how much they should invest for their marketing activities [9]. The overall objective of conducting this research was trying to solve this business problem. In order to do so, a case study was conducted at La Place. The current way of designing promotions and deals at La Place is to advocate the variability of the menu by highlighting different products mainly focusing on the seasonality and time of the year. The promotions are sold at a discount, for example a coffee and apple pie deal for Euro 3.00 instead of the normal price of Euro 5.50, ideally suited for breakfast. The deals are targeted at the entire customer base and not to specific groups. This can cause customers to visit only to take benefits of the ongoing promotions and never return. Customers who avail promotions one-time and never visit again do not really benefit the overall revenues. Ideally, it would be nice to tailor these promotional offers to specific customers, who have a higher probability of returning.

This is where this research can add value. The objective of this research is twofold: (1) to propose a method to La Place which would be capable of explaining how to classify customers, and (2) to evaluate the method for its usefulness and usability. The proposed method will be applied to the already identified classes of customers. The result of these steps can be used to find small and specific groups of customers, which can be of specific interest to the marketing teams in order to

reach out with promotions and deals. By addressing the business problem of La Place, it can be safely assumed that the proposed method can indeed be useful to similar companies.

1.4 Research Question

Based on the research objective stated in Chapter 3, the research questions are formulated. In order to answer the research questions, existing literature is used, and a method is proposed to achieve the research objective.

The main research question formulated from the research objective is as follows:

How can classification and clustering techniques be used in effective target group-based marketing schemes?

The goal of the main research question is to understand how data mining techniques like classification and clustering can be used to find similar groups of customers. In order to answer the main research question, firstly it is important to have a good understanding of the different customer classification and clustering techniques which are widely and effectively used by companies. Then , it is important to understand what the additional value is of doing such a customer classification and clustering scheme. Additionally, such a scheme can be used to possibly predict future visitors. Therefore, some research sub questions are defined in order to contribute towards answering the main research question and the research objective. The first research sub-question is formulated as follows:

SQ1: Which are the most widely used customer classification and clustering techniques?

The goal of SQ1 is to know about existing classification and clustering techniques which are popularly used by companies to group customers. This sub-question is answered primarily in Chapter 2 through a systematic literature review. This forms a basis for answering the second research sub-question, which is as follows:

SQ2: How can classification and clustering techniques be combined to group customers?

Based on the findings of SQ1, one or more classification and clustering techniques are selected. The goal of this sub-question is to propose a method to groups customers based on their behavior and preferences, by combining one or more classification and clustering technique. A case study is used in Chapter 4 to show how the proposed method can be used in practice. It is then important to understand the additional business value of the results of these techniques. Therefore, the third research sub-question is formulated as follows:

SQ3: What is the business value of incorporating results of these techniques in marketing strategies?

The goal of SQ3 is to understand if it is indeed useful and worth investing time and effort in implementing the proposed scheme. This sub-question is answered with the help of existing literature and findings from a case study. It is attempted to find if some research papers reveal how some companies benefited from such schemes in the past. Based on company's historical data, it is possible to find which specific groups of customers need to be focused upon in order to reduce marketing/promotion costs and increase company's profits. Chapters 5 and 7 are used to summarize answers for this research sub-question. Finally, it can be useful for companies if they can predict their future visitors based on historical data. The fourth research sub-question is formulated as follows:

SQ4: How can we predict future visitors based on the results of these techniques?

The goal of SQ4 is to explain how the results of such a scheme can be used to predict future visitors based on historical data. This sub-question is answered with the help of implementing random forest classifier on the already identified groups of customers from SQ2 and can be found in Chapter 5.

1.5 Research Methodology

In order to ensure a stable research framework and methodology, the Design Science Research Methodology (DSRM) is used [15]. This research framework is commonly used for Information system research in Design Science. In other words, the DSRM model is used in the designing of a software (artifact/prototype) that is reused in the context of a research field and evaluating that software (artifact/prototype) in the intended context.

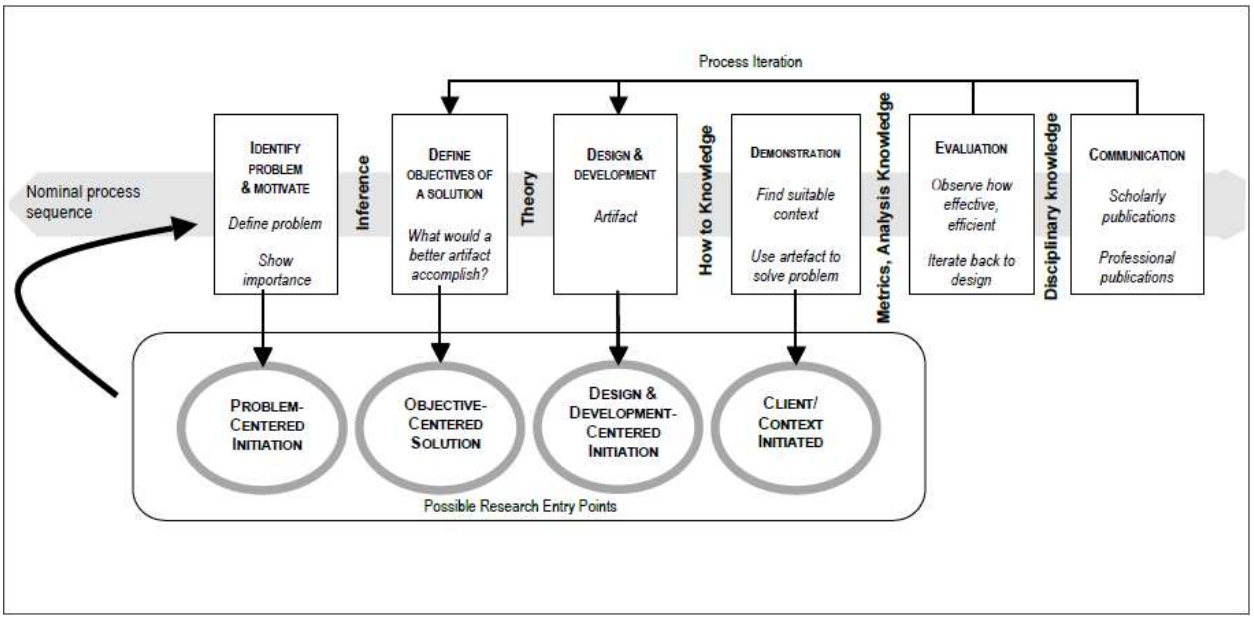


Figure 2 Design Science Research Methodology (DSRM) Process Model

The same steps from the DSRM are followed for conducting this research, as shown in Figure 2.

- The first step includes the "Identification of the problem and motivation". In this step, the research problem is defined, and its importance is shown with the problem at its center. It leads to 'Problem-centered initiation' as a possible research entry point. This is mainly covered in Chapter 1 of this thesis and is primarily used in answering SQ1 and SQ3.
- The second step is "Defining the objectives of a solution". This step uses inferences to determine what would a better artifact accomplish by solving the earlier stated problem. The objective(s) of the solution is the center of attraction and this step leads to 'Objective-centered solution' as a possible research entry point. This is mainly covered in Chapters 1 and 2 of this thesis and is also used to answer SQ1 and SQ3.
- The third step is the "Design and development" of the artifact. This is where the artifact is designed and developed using theory and knowledge about the problem being addressed. It is explained in Chapter 3 of this thesis and is mainly used in answering SQ2. This leads to 'Design and development centered initiation' as a possible research entry point.
- The fourth step is "Demonstration" of the artifact. The artifact from step 3 is demonstrated or implemented in a given problem context and leads to 'client/context initiated' research entry point. This is explained in Chapters 4 and 5 of this thesis with the help of a case study. The designed is artifact is used to solve a problem and mainly answers SQ2 and SQ4.

- The fifth step involves the “Evaluation” of the artifact. The artifact is observed for its efficiency and effectiveness. If required, iterations to the third step are proposed. This is covered in Chapter 6 of this thesis.
- The final step is “Communication” about the usefulness of this study and its contributions. This can be done with the help of scholarly publications and showing contributions to science and technology. This is briefly explained in Chapter 7 of this thesis.

1.6 Structure of the report

For the purpose of making it easier for the reader, this thesis is structure into different chapters. Chapter 1 provides a brief introduction about the role of customer relationship management, followed by information about classification and clustering. Chapter 2 uses literature review to know about the most widely used classification and clustering techniques within companies. Chapter 3 describes the design of the method (artifact) and the different techniques and algorithms used. Chapter 4 is a demonstration of the artifact by means of a case study at a company. Chapter 5 discusses the results of implementing the designed artifact in the given problem context of the company. Chapter 6 evaluates the artifact and is used for feedback about the artifact. Chapter 7 is used for a general discussions, limitations and conclusions about the artifact. Chapter 8 makes recommendations for future work to make the artifact better.

Chapter 2. LITERATURE REVIEW

The main purpose of conducting a literature review was to find which classification and clustering techniques were widely used in practice by companies and how are they combined. The findings of the literature review are mainly used in answering SQ1 and SQ2. In order to conduct the literature review, a systematic literature search was carried out using the guidelines based on the *Grounded theory* [16]. This also formed the basis for conducting the 'Research Topics' which preluded conducting this research to get an idea about the theory available in literature. For conducting the literature review, five iterative steps were defined namely,

a. Define criteria for inclusion and exclusion and determine search terms, sources and field of research.
b. Search for literature.
c. Select and refine the sample of selection.
d. Analyze the selection of literature.
e. Present and structure the content.

The most important steps are described below:

A. **INCLUSION AND EXCLUSION CRITERIA**

All research and studies were included that:

i. presented theoretical and practical aspects of implementation of classification and/or clustering techniques in organizations, in order to understand the history and background of these techniques and the way they are implemented.
ii. were published after the year 2000, as data mining mainly blossomed in the past 2 decades.
iii. were case studies of the application(s) of these techniques in both the retail and wholesale markets.
iv. were published in well-known and reliable journals/publications. H-index[1] was used a quality aspect for ranking purpose and only those with H-index higher than 10 were included.
v. were only in English text.
vi. were limited to fields of Computer Science, Business and decision making and Decision Sciences. Other irrelevant fields like Social sciences, Neurosciences or

[1] https://en.wikipedia.org/wiki/H-index

Mathematics were not considered important because the emphasis was more on the application in retail and wholesale markets.

All research and studies were excluded that:

i. were not freely accessible and needed a paid download or access request from the author(s).
ii. applied these techniques to non-customer data.
iii. did not have enough detail about the implementation of the technique and probably provided just an abstract or high-level overview.
iv. were deemed irrelevant upon reading the titles and abstracts.
v. were found to be duplicates.

B. **SEARCH**

In this step, the focus was mainly on the search queries and the data sources to be queried. Reliable and trusted web search engines like Scopus[2], Science Direct[3] and Google Scholar [4] were used. In some cases, when it was not possible to download articles from these sources, other sources like the LISA (University of Twente Library)[5] were used.

The search strategy applied consisted of keywords which would enable retrieval of relevant results and were mapped towards answering the research questions listed earlier. A pre-exploratory "general" literature search was conducted to construct the list of concepts and keywords and to understand the topics better before building a search query. This was an iterative approach but helped in building towards a concrete list of concepts and keywords we were looking for. Using this as an input, one of the main advanced queries used in Scopus was as follows:

(("Customer" AND "classification") OR ("Customer" AND "clustering") AND (PUBYEAR > 2000) AND (LANGUAGE (english))) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "BUSI") OR LIMIT-TO (SUBJAREA , "DECI")) AND (LIMIT-TO (ACCESSTYPE(OA)))

The same query was used for answering both the research questions as the focus was on finding more (bigger set) results first, and then refining them carefully to ensure not missing out on important content. Not all subject areas were used, and the focus was mainly on subject areas like business, decision making and computer science, where we expect most extensive and relevant use of customer classification techniques. In addition to the results obtained from this

[2] https://www.scopus.com/home.uri
[3] https://www.sciencedirect.com/
[4] https://scholar.google.nl/
[5] https://www.utwente.nl/en/lisa/library/

query, additional references were obtained from other sources like Google Scholar and Science Direct which could not be accessed freely from Scopus.

C. SELECTION OF THE SAMPLE AND ANALYSIS OF THE SELECTION

In this step, the results of the above search query are reviewed and filter the articles relevant to our research. This was the most time-consuming task, considering that the search query (even after being very specific) resulted in about 2377 results. This long list was exported, and the titles of all articles were reviewed to check if they remained relevant to what we were exactly looking for. Additionally, articles from other sources like Google Scholar and Science Direct were included resulting in a total of 2467 results. The titles of all these articles were screened to filter the articles which would be useful. This was an important step as a majority of 2215 (about 90%) articles were discarded. The reason for discarding these articles was because based on the titles, they did not really deal with 'customer' classification/clustering and did not have enough implementation details. At times, they appeared to be more theoretical or literature reviews and were from different subject areas and industries. This left a smaller set of 252 articles for abstract screening. Based on careful reading of abstracts, about 180 articles were excluded and out of the remaining 72, nine duplicates were removed. Duplicate removal is done at this stage as recommended by the *Grounded Theory*. After duplicate removal, 63 articles were reviewed completely based on their full-text, out of which 45 articles did not provide enough information or were incomplete.

The final list consisted of 18 articles, however forward and backward citations resulted in 5 more articles being included in the literature review, making it 23 articles. Forward and backward citations were carried out using the second-level or second-generation reference search. It was only possible to do this after the full text of the sample have been read. Figure 1 shows a flowchart diagram of the results of the Grounded theory.

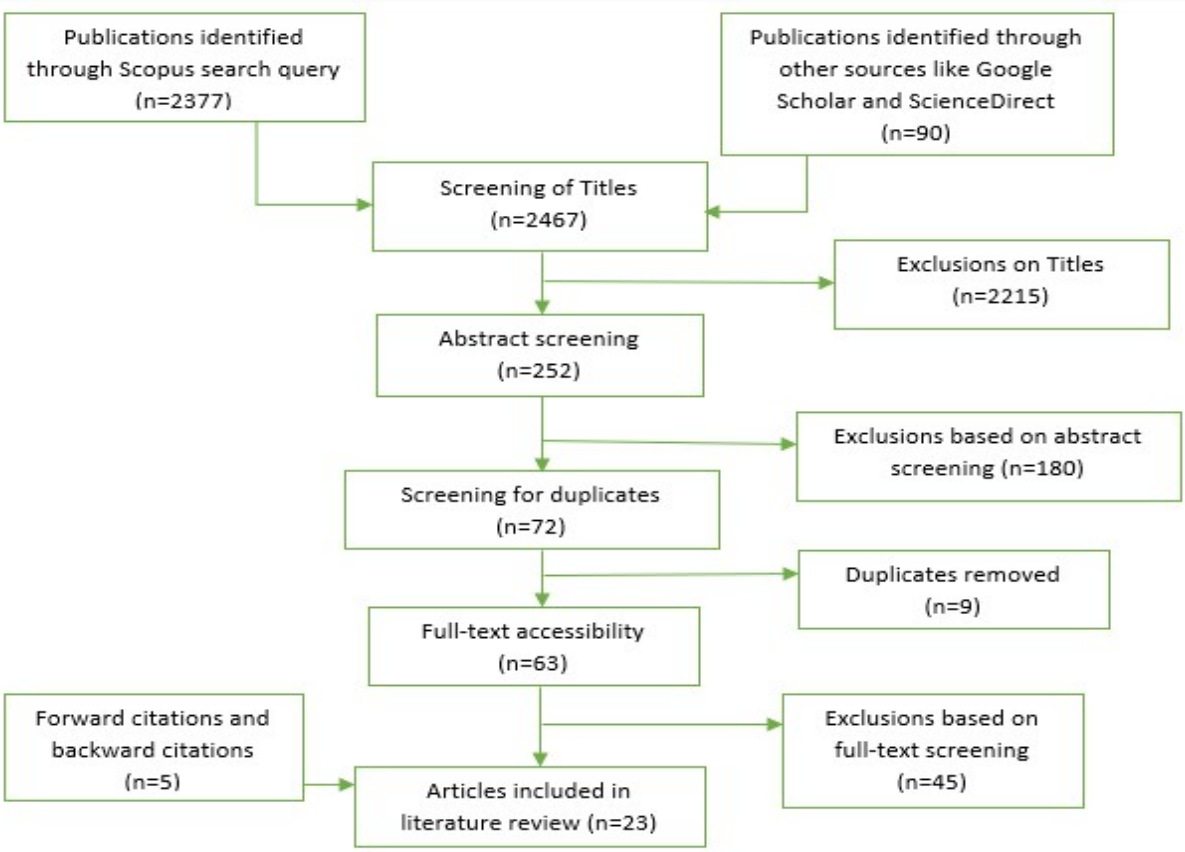


Figure 3 Flowchart of the Grounded Theory Results

2.1 Classification Techniques

It is important to understand different classification techniques used in practice to classify customers. This would help in answering SQ1. A data mining architecture based on clustering techniques which assists experts to segment customers based on their purchase behaviors was presented [17]. In the proposed architecture, diverse segmentation models are automatically generated and evaluated with multiple quality measures. The architecture is based on the RFM (Recency, Frequency and Monetary) and SAX (Symbolic Aggregate Approximation) models, which are commonly used models in customer segmentation. There are four main components, namely, the input component which is responsible for the data transformation, the generation component which is responsible of the creation of the clustering results, the selection component which scores the clustering results to only keep the more relevant segmentations, and the visualization component to graphically represent the clustering results. The results are validated by applying the proposed methodology to a dataset consisting of the purchase log of

10 000 customers over 62 weeks and used the F-measure and R^2 (Recall) measure scores and explained the results with the help of test and training data.

One of the most renowned models for customer value analysis is the RFM (Recency, Frequency and Monetary) model and has been used by many researchers to perform customer segmentation [18] [19] [20]. The RFM model analyzes the behavior of the customers and therefore it is considered as a behavior-based model [21]. RFM model has a limitation as it is not able to divide customers into segments based on the length of their relationship with the company [22]. An approach combining the fuzzy c-means clustering and genetic algorithms to cluster the customers of steel industry is described [7]. This approach used the LRFM model which is an extension of the RFM model, *L* standing for Lifetime or Length of a customer's relationship with the company. This combination of algorithms was used because the performance of fuzzy c-means algorithm is strongly affected by the selection of the initial centroid clusters and the performance of the combined algorithms had a lower mean squared error (MSE) compared to fuzzy c-means clustering. With the help of GA -Fuzzy clustering software which combines fuzzy c-means clustering and genetic algorithm, a dataset of 120 customers was distinctly clustered.

The LRFM model is combined with the k-means++ algorithm to identify clusters in a dataset of customers from a well-known, multi-national, Fast-Moving Consumer Goods (FMCG) company in Egypt [1]. The number of clusters *k* is found by integrating a bootstrapping phase in addition to employing both the Calinski-Harabasz index and Rand cluster validity index. For validation purposes, the Welch ANOVA test is deployed to ensure distinguishable characteristics among the clusters.

A two-stage clustering approach was proposed to cluster 597 specialty food retail customers into four clusters namely "standalone rationals", "foodies", "cherry pickers" and "indulgencers" [23]. This approach used the Ward's method, followed by the k-means clustering algorithm. Results were validated using the MANOVA and ANOVA analysis and there was considerable inter-cluster difference among members, indicating good results.

The IBM I-Miner tool is used to cluster customers and identify the high-profit, high-value and low-risk customers [6]. There were two phases. The first phase involved the data cleaning, followed by development of patterns via demographic clustering algorithm using IBM I-Miner tool. The second phase consisted of data profiling, developing clusters and identifying high-value, low-risk customers. This is a relatively simple approach to quickly cluster customers without much complex calculations.

A customer segmentation framework based on data mining was proposed and a new customer segmentation method was constructed based on survival character [3]. The framework presents the whole process of segmentation, which includes data acquiring, issue mapping, segmentation model establishing, data mining arithmetic selecting and function analysis. This framework used k-means clustering by which customers having similar survival characters (churn trend) are clustered together. This is followed by predicting each cluster's survival/hazard function using survival analyses, then the validity of clustering is tested, and customer churn trend is identified. Results are validated against a dataset of mobile customers from a Chinese telecom company.

It is not only important to segment customers but also to create marketing strategies specific to those customer segments [9]. This research proposed a new Life Time Value (LTV) model and customer segmentation considering customer defection and cross-selling opportunity. The model was applied to 2000 customers from a telecommunication company, divided them into clusters based on their LTV and then proposed marketing strategies based on the observations of each cluster.

Customer value has become an important parameter in customer segmentation [24]. It is termed as Customer Lifetime value (CLV) and often acts as a differentiator between different customer segments, especially in CRM and decision-making for marketing. A framework to predict CLV was proposed based on adapted weighted RFM analysis [24]. K-means algorithm was used for clustering and number of clusters was decided based on Dunn index. ARIMA (Auto Regressive Integrated Moving Average) time series model was used on the identified clusters of customers to estimate their future CLV. Successful experimental results have been shown from the retail banking sector from this research.

Apart from all customer classification and segmentation techniques discussed, which use unsupervised and supervised techniques, the role of capturing social media data, lifestyle and other characteristics is very important and an emerging trend [12]. A complete customer clustering scheme would be a combination of transaction behavior and demographic behavior [12]. Social media complements, rather than overlaps, existing transactional data in databases. Additionally, it can help customers reach even a larger base of customer through common communities, networks, etc.

2.2 Classification Techniques

Like classification, it is important to understand different clustering techniques used in practice to cluster customers. This would also help in answering SQ1. A prediction model to

identify customers who would most likely respond to the prospective offerings of the company based on their past purchasing trends was built by [5]. The author used a combination of different classification techniques like Naïve Bayes, KNN (K nearest neighbor) and SVM (Support vector machines), and then compared the performance of these techniques. It was found that the Naive Bayesian Classification achieves the highest accuracy and specificity in this specific context. However, the worst classification was performed by Support Vector Machine technique as SVM technique achieves lowest sensitivity, accuracy and specificity. Additionally, SVM achieves the highest error rate with many false positive and false negative cases when compared to their peers.

The application of Information Theory in selecting customer features was analyzed [25]. Feature selection is an important data mining technique which helps in selecting only the relevant customer attributes for customer classification and discards the irrelevant ones. This work presents a variable selection method based on entropy and mutual information. *Information Gain* or *Max Relevance* is the concept of selecting variables that provide most information. The Information Theory concept is applied to a dataset about a bank telemarketing campaign. Based on Entropy and Information Gain calculations, the most relevant variables for feature selection are selected as a maximum product of Chi-square test and Reliability percentage. To evaluate the model, LIFT metrics was used, which reflected that 80% of the positive responses can be achieved by 50% of the sample scored with this method.

Data accuracy is an important dimension of data quality and plays an important role in the results of customer classification. Therefore, it is important to analyze the impact of data accuracy on different customer classification techniques [26]. This work analyzed this impact using a dataset of two empirical direct marketing data sets provided by the Direct Marketing Educational Foundation (DMEF)[6] and found that decision trees perform better than RFM analysis, CHAID and logistic regression under optimal data accuracy.

RFM, CHAID and logistic regression are popular customer classification techniques. A comparison between them was done [27]. In order to compare these techniques, they were applied on two different datasets. It was found that the RFM could be more successful overall when the response rate to marketing e-mails or campaigns is high. CHAID performed better than RFM when the response rate was low and when the target customer base to be reached out via emails and campaigns was small. Overall, CHAID and logistic regression performed better than RFM.

[6] http://www.directworks.org/

Customer churning is a major issue for most companies to deal with [28]. Customer churn is when an existing customer, user, player, subscriber or any kind of return client stops doing business or ends the relationship with a company. Churn prediction is a supervised problem and techniques like classification can help solve this problem with techniques and methods like logistic regression, Naive Bayes, decision trees, artificial neural networks and SVMs to predicts churning customers from the loyal ones. This issue is addressed by using multiple classifiers ensemble (MCE), which has become a powerful tool for improving classification performance in numerous fields [28]. MCE combines the results of base classifiers in reaching a final decision, however the challenge is how to select an optimal subset of base classifiers. This paper considers MCE which has a built-in model-selection function for predicting churning customers using group method of data handling (MCES-GMDH). The MCES-GMDH model's training contains two phases: base classifiers training and classifiers ensemble selection. The higher the classification accuracy and the diversity among results, the better is the ensemble model's performance. To validate results, they applied the MCES-GMDH to two different datasets and discussed about some other classifier ensemble methods bagging and boosting, GAMensPlus, and GAMens**Error! Bookmark not defined.**. Eventually, it was concluded that for these datasets, the MCES-GMDH performs better than any of the other classifier ensemble methods. It was also shown that the MCES-GMDH model proposed here can better deal with the classification issue with imbalanced distribution.

Most of the customer classification techniques use transactional data of customers. This is highly volatile, sparse and skewed data because most customers have very few transactions. This is referred to as imbalanced distribution and can cause bias in the accuracy predicting classifiers. As buying behavior of customers change over time, their transactional data changes rapidly and so does their classification profiles. In order to tackle this issue, a comparison of four different customer classification approaches was done [29]. The goal of this research was to decide whether to change class labels of dynamic customer profiles or adapt the classifiers. It was concluded that it is more logical and effective to adapting the classifiers in shorter time windows of analysis whereas changing class labels or re-labelling them is the correct approach to go ahead with for longer time windows.

Chapter 3. DESIGN

This chapter describes the design of the artifact. The artifact is a generalizable method which can be adopted by companies storing transactional data of customers. The main focus of this chapter is about the design choices in terms of the selected model, algorithm and techniques. Based on the literature review from Chapter 2, the LRFM model, the K-means algorithm and the Random forest classifier were selected. Details about these can be found in the next sections of this chapter.

3.1 Models, algorithms and techniques

The findings from the Literature Review indicated that the RFM model is one of the mostly popular and widely used methods for customer classification [30]. However, RFM has the limitation that it does not consider the Length of a customer's relationship with the company. Therefore, the LRFM model was proposed as an improvement to the RFM model [21]. Using this method, it is possible to classify, or label customers based on their length (L), recency R), frequency (F) and monetary (M) values. The K-means clustering algorithm is used for clustering. A detailed description of the LRFM model and K-means clustering algorithm is provided below and then we discuss how they are used in the design of the prototype.

3.1.1 The LRFM Model

The traditional approach to use the LRFM model is to sort the customer data and then divide the data into five equal segments for each dimension of LRFM. The top 20% segment is assigned as a value of 5, the next 20% segment is assigned as a value of 4, and so on. Thus, each customer based on LRFM model can be represented by one of 625 (5*5*5*5) LRFM cells, namely, 5555, 5554, 5553, . . . , 1111. The definitions are as follows: length (L) is the time length between the first purchase date until the last purchase date, recency(R) is the time length since the most recent purchase; frequency (F) computes the number of purchases during the same period; and monetary (M) refers to the total amount of money spent on all purchases given the same period. The definitions are summarized in Table 1. Based on the values of these dimensions for each customer, they can be grouped into different groups as proposed in [31]:

1. best customers with always a high monetary value and high frequency (L ↑ R ↑ F ↑ M ↑ , L ↓ R ↑ F ↑ M ↑ , L ↑ R ↓ F ↑ M ↑ , and L ↓ R ↓ F ↑ M ↑),

2. frequent customers with always a high frequency of visiting (L ↑ R ↑ F ↑ M ↓ , L ↓ R ↑ F ↑ M ↓ , L ↑ R ↓ F ↑ M ↓ , and L ↓ R ↓ F ↑ M ↓),
3. spender customers with always a high monetary value and low frequency (L ↑ R ↑ F ↓ M ↑ , L ↓ R ↑ F ↓ M ↑ , L ↑ R ↓ F ↓ M ↑ , and L ↓ R ↓ F ↓ M ↑), and
4. uncertain customers with always low frequency of visit and monetary value (L ↑ R ↑ F ↓ M ↓ , L ↓ R ↑ F ↓ M ↓ , L ↑ R ↓ F ↓ M ↓ , and L ↓ R ↓ F ↓ M ↓).

Variables	**Definitions**
Length (L)	*Number of days from the first purchase date to the last purchase date*
Recency (R)	*Number of visits from the first day of the study period to the last purchase date.*
Frequency (F)	*Number of visits in a specified time period*
Monetary (M)	*Total amount spent by the customer in the study period*

Table 1 Definitions of LRFM model

3.1.2 The K-means clustering algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed apriori (known beforehand).

K-means requires an initial specification of cluster centers. Starting with one cluster, the method chooses a variable whose mean is used as a threshold for splitting the data in two. The centroids of these two parts are then used to initialize k-means to optimize the membership of the two clusters. Next, one of the two clusters are chosen for splitting and a variable within that cluster is chosen whose mean is used as a threshold for splitting that cluster in two. K-means is then used to partition the data into three clusters, initialized with the centroids of the two parts of the split cluster and the centroid of the remaining cluster. This process is repeated until a set number of clusters is reached. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

3.1.3 Random Forest Classifier

Random Forests is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job.

Random Forests provides an improvement over bagged trees (decision trees)) by a small change that *decorrelates* the trees. As in bagging, you build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a *random sample of m predictors* is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those m predictors. This is the main difference between random forests and bagging; because as in bagging, the choice of predictor *m=p*.

In order to grow a random forest, you should:

- First assume that the number of cases in the training set is *K*. Then, take a random sample of these *K* cases, and then use this sample as the training set for growing the tree.
- If there are *p* input variables, specify a number m<p such that at each node, you can select *m* random variables out of the *p*. The best split on these *m* is used to split the node.
- Each tree is subsequently grown to the largest extent possible and no pruning is needed.
- Finally, aggregate the predictions of the target trees to predict new data.

Random Forest is very effective at estimating missing data and maintaining accuracy when a large proportion of the data is missing. It can also balance errors in datasets where the classes are imbalanced. Most importantly, it can handle massive datasets with large dimensionality.

However, one disadvantage of using Random Forests is that you might easily overfit noisy datasets, especially in the case of doing regression.

3.2 Design of the proposed method

In this study, a method is designed to carry out the classification and clustering of customers, right from initiation to identifying relevant groups for marketing strategies. The method can be seen in Figure 4. There are mainly three phases:

- Phase 1 deals mainly with the data collection, integration, cleaning and transformation activities. This can be thought of as a preparatory phase. This phase also includes integration with correct data sources, removing corrupt or incorrect data, handling missing data and calculation of new attributes relevant to the classification scheme specific to the context.
- In phase 2, the actual customer classification and segmentation takes place and their results are combined. One or more classification/clustering techniques are chosen and then applied on the customers from phase 1. The output of phase 2 is groups of customers with common characteristics.
- Finally, in phase 3, the groups identified in phase 2 are analyzed and marketing strategies are proposed for these groups. If the identified groups happen to be too large or complex to understand, they can be passed through phase 1 again to narrow down the results further.

The three phases within the method are connected in a cyclic fashion and one-directional.

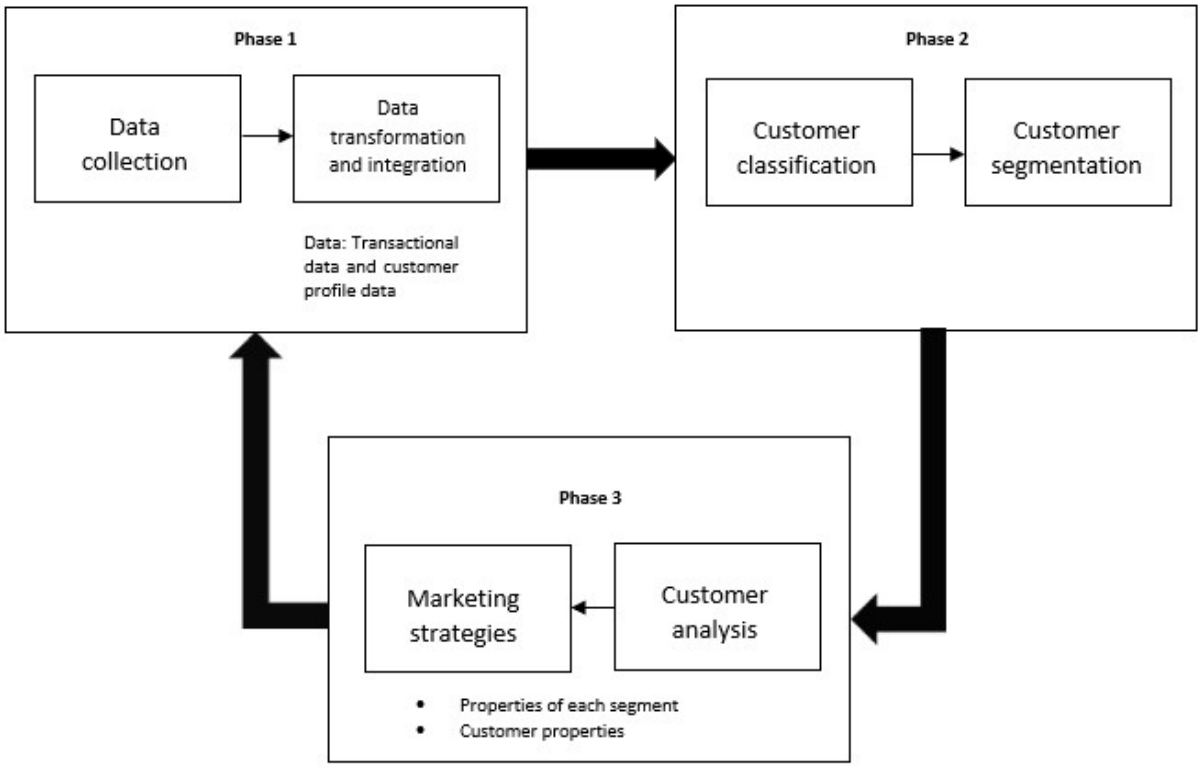


Figure 4 Proposed method

3.3 Design choices, assumptions and constraints

This section describes the design choices, assumptions and constraints. These are based on multiple factors like the utility of the method being developed, fore sighting upcoming issues, data quality issues and so on.

Following design choices, assumptions and constraints were considered:

- The focus was on a classification scheme based on how recently customers have visited. So, it is more concentrated on Recency (R) rather than Frequency (F)
- Only transactional data of customers with membership/loyalty cards are considered. This does not guarantee all transactions of membership/loyalty card holders were recorded as some customers tend not to use their cards on every visit due to different reasons.
- Only the high level transactional data is used such as the revenue, number of products, weekday of the transaction and daypart of the transaction. Other in-depth details like the products within each transaction, whether on promotion or not are considered to be outside the scope of this research.

- Transactions with negative transactional values such as revenue, number of products should be cleaned as they are incorrect entries and can lead to misleading results.
- All computations and calculations should be done strictly only within the selected BI tool.
- The quality of the remaining data should be trusted.
- Transactions only within normal working hours should be considered for consistency and standardization purposes.

3.4 Tools

The main source of customer and transactional data would normally be from widely used data sources which includes database servers like Microsoft SQL server, Tableau Server and other ERP systems like SAP. These systems should be capable of making live connections to external data sources and have high reliability in terms of data replication and backups. Sometimes, the data in these source systems cannot be changed and needs pre-processing before it can be used further. These source systems should be able to make connections with popular Business Intelligence Tools such as Tableau, Power BI, QlikView and SAP HANA to name a few. All simple and complex calculations/ transformations need to be made within these BI tools. These BI tools should finally be connected to popular and powerful tools for data analysis like Python, R and Microsoft Excel. These tools can perform simple tasks like presenting charts and views, and even more complex tasks like regression, clustering and data modelling. A general data flow architecture for the proposed method would look like Figure 5.

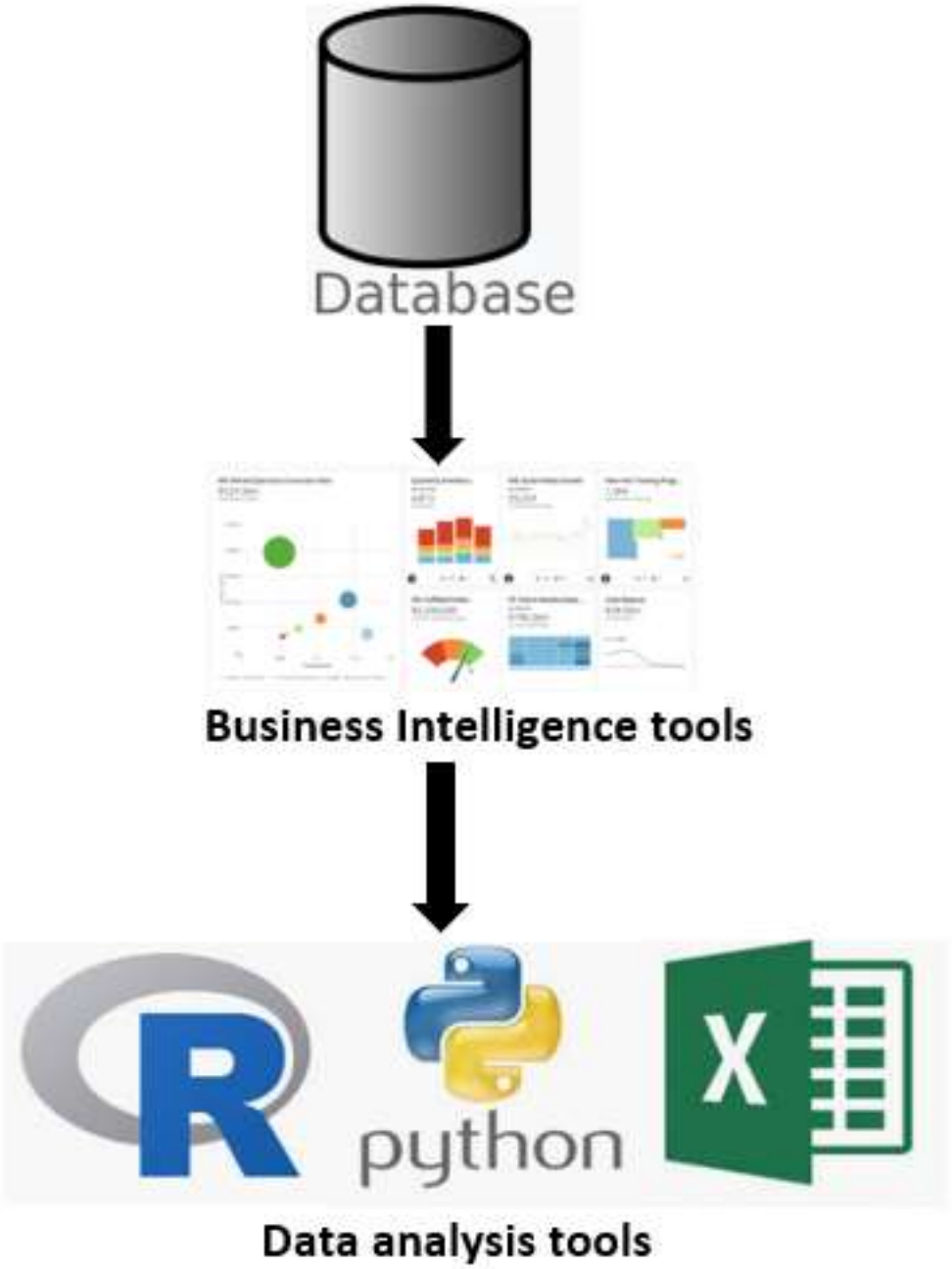


Figure 5 Generic Data flow architecture

Chapter 4. CASE STUDY

This chapter deals with the demonstration of the proposed method (artifact) in a real-world scenario with the help of a case study at a company.

4.1 About the Company

This study was conducted at La Place, a popular Dutch restaurant chain. La Place is currently owned by Jumbo, a popular supermarket chain in the Netherlands. Until 2015, La Place was owned by the Dutch store giant V&D (Vroom & Dreesmann), which later went bankrupt and had to stop operations. As of 2015, V&D operated 67 branches throughout the Netherlands. La Place, was its former subsidiary restaurant chain which had both in-house and standalone restaurants throughout the country. The department stores sold clothing and shoes, jewelry, cosmetics, books, home-entertainment products, electric goods, stationery, cards and posters, furniture and home wares. Most branches also had a La Place in-house restaurant, a travel agent and an ATM. Larger branches also had a bakery. The variety of services and products they offered attracted a huge customer base. However, after bankruptcy, a large portion of this customer base was lost due to the shutting down of its stores and this obviously, had major impact on the customer base of La Place. So, it has become increasingly important for La Place to better understand its customers and maintain stronger relationships with them.

With a huge focus on customer loyalty, La Place started a 'Loyalty programme'. The members of this programme had unique card numbers and whenever they made a purchase using this card, their transactions were recorded. This programme was started when the operations were run by the V&D group and remains active till today, with about 800k customers. The notion behind starting this programme was that customer loyalty was a major growth lever in the market. The market is very competitive in terms of pricing and what made La Place unique is the experience it offered to its customers. Fresh and healthy food, large variety in assortment, colorful display of assortment, clean and vibrant ambience, polite and friendly service from employees are some of the features that make La Place unique and one of the very few restaurant chains of its type. However, it is known that La Pace is comparatively at the higher end on pricing because of the features it provides when compared with other restaurants and cafes in the Netherlands. In such a competitive market with so many other options available for customers to eat, it was realized that the population or the customer base will be reasonably stable. This obviously means that people will not eat much more at normal prices unless there are substantial promotions or deals. Thus, in order to increase the revenues, customers who are already familiar

with La Place need to visit more often and need to be retained. This is where customer relationship management can help.

As stated earlier in , the context of the problem does not remain confined to a specific company like La Place. Primarily, this research proposes a method which is generalizable and can be adopted by companies storing transactional data in the HORECA (Hotel/restaurant/café) industry. Then this method is demonstrated for a company like La Place, but the overall goal is to make it generalizable to other companies without major changes. Other companies from the HORECA industry who store transactional information of identifiable customers would also benefit from understanding the behavior of its customers and proactively responding to ensure not losing out on valuable customers in a competitive world. It is worthwhile mentioning that not many HORECA companies traditionally store transactional data of customers and La Place is one of the few doing so.

4.2 The model

The LRFM model forms the basis of designing our artifact. Since the LRFM model can lead to many groups, it was decided to limit the number of groups. The proposed model is inspired from the LRFM model. The Recency (R) and Frequency (F) were combined and only Recency was used. Recency is defined as the number of days since the last transaction. Different labels were defined on the basis of how recent the last transaction was. Lifetime (L) was defined as the number of days from the first transaction of the customer (until the day the flow was run). The Monetary value (M) was defined as the average ticket value of each customer for their whole Lifetime. Age based classification was also done as the marketing team strongly believed in that. The names of the proposed classes/labels and their values were motivated by prior business knowledge of the supervisor from La Place. For example, last 13 weeks was used as a reference period because of a preference for performing a running quarterly analysis in comparison to previous years. Another example is the division of dayparts into 4 classes depending upon the time of the day and the promotions for specific parts of the day. Some of them were inspired from terms used by the Marketing team such as ‘Baby boomers’ for age. Definitions of these attributes used in our research and the different labels can be found in Table 2.

Attributes	Definitions	Classes/Labels
Recency	*Number of days since the first or the last transaction depending upon the label (13 weeks is selected as a reference period mainly because of a running quarter analysis)*	***New**: customers who made their first transaction in the last 13 weeks and had no previous transactional history* ***Frequent:** customers who made at least one transaction in the last 13 weeks and at least one more transaction within the past 365 days.* ***Monitor**: customers who did not make any transaction in the last 13 weeks but had visited at least once in the last 365 days.* ***Inactive**: customers who haven't visited in the last 365 days but had visited in the year before (between 365 days-730 days)..* ***Lost**: customers who did not make any transactions for over 2 years (730 days).* ***Redormed1**: customer who had last visited at least once one year ago (between 365-730 days) and turned up in the last 13 weeks.* ***Redormed2**: customers who had last visited beyond 2 years ago (730 days) and turned up in the last 13 weeks.*
Lifetime	*Number of days since the first transaction*	***New**: customers who made their first transaction in the last 6 months.*

		***Middle**: customers who made their first transaction in the last 2 years, excluding the 'New' customers* ***Old**: customers who made their first transaction prior to 2 years (more than 730 days ago).*
Monetary	*The aggregated average ticket value of each customer*	*No Labels*
Age	*The actual age of the customers (in years). (The labels are motivated by the marketing team)*	***Generation Z**: customers aged between 9 to 19 years.* ***Millenials**: customers aged between 20 to 39 years.* ***Generation X**: customer aged between 40 to 53 years.* ***Baby Boomers**: customers aged between 54 to 73 years.* ***Silent Generation**: customers aged between 74 to 90 years.* ***Unknown Age:** remaining customers for whom the age was not known or missing.*
Daypart	*The part of the day when the transaction was recorded*	***Breakfast** is when the transaction was recorded between 0800-1200 hours* ***Lunch** is when the transaction was recorded between 1200-1500 hours* ***Midday** is when the transaction was recorded between 1500-1700 hours* ***Dinner** is when the transaction was recorded between 1700-2100 hours*

Table 2 Definitions of the proposed model

'Recency' is the most important attribute because in a competitive business, the frequent customers are the ones who contribute the most to the company's revenues. So, every company would like to know who these customers and try their best to retain them. The most interesting part of the 'Recency' based labels are that customers will move within these groups every 13 weeks. A 'New' customer can become a 'Frequent' customer if they keep visiting regularly. On the other hand, 'Frequent' customers can turn into 'Monitor' if they stop visiting. Similarly, 'Monitor' customers can turn into 'Inactive' and 'Inactive' into 'Lost' or 'Redormed1'. Lost customers can become 'Redormed2'. In order to increase company revenue, all attempts should be made to keep 'Frequent' customers loyal, turn 'New' and 'Monitor' into 'Frequent', 'Inactive' into 'Redormed1' and 'Lost' into 'Redormed2'.

A 'Recency' matrix is shown in Table 3 for better understanding of this attribute.

	First/Last transaction date based on definition			
Labels	Last 13 weeks	Up to 1 year	Between 1-2 years	Beyond 2 years
New	*Yes*	*No*	*No*	*No*
Frequent	*Yes*	*Yes*	*Not relevant*	*Not relevant*
Monitor	*No*	*Yes*	*Not relevant*	*Not relevant*
Inactive	*No*	*No*	*Yes*	*Not relevant*
Lost	*No*	*No*	*No*	*Yes*
Redormed1	*Yes*	*No*	*Yes*	*Not relevant*
Redormed2	*Yes*	*No*	*No*	*Yes*

Table 3 Recency matrix based on Transaction history

4.3 Tools

As shown in Figure 5, a source system is needed which stores the customer and transactional data. At La Place, the source system was a Microsoft SQL server[7]. The Business Intelligence tools available were Qlik View[8], Tableau Prep and Tableau Desktop. The combination of Tableau Prep and Tableau desktop was used mainly because of their capability to connect independently to Microsoft SQL Server and their ability to communicate between each other. QlikView was discarded as it was not able to perform most of the functions offered by Tableau. Apart from these, no other BI tools were available within the organization and there were minimal chances

[7] https://www.microsoft.com/en-in/sql-server/sql-server-2019
[8] https://www.qlik.com/us/products/qlikview

of acquiring a new one, something like the well-known Microsoft Power BI[9]. This was mainly due to organizational policies and an upcoming business transition which made the future uncertain. Tableau Prep is a powerful data shaping, combining and cleansing tool. It has the capability of combining with different data sources like SQL databases, Oracle databases, Enterprise resource planning (ERP) systems like SAP, Excel files and even on-premise servers. To combine data from different sources, Tableau Prep has powerful in-built functions like Join, Pivot, Union and normal cleaning/filtering steps. It is possible to design a relatively complex data flow, right from reading the input data by connecting to a data source, to giving a structured data output, with intermediate complex transformation steps. It is easy to understand even complex data flows because of the transparency, ease of use, immediate results and smart calculations. Other similar tools in the market are Pentaho, Rapid Miner, KNIME and so on (more information in Appendix).

Tableau Desktop is one of the most powerful Business Intelligence (BI) tools currently available in the market. Like Tableau Prep, it can connect to many data sources. It is mainly used for visualization purposes with the help of designing simple charts, views, dashboards and stories. It is also capable of performing statistical operations like trend analysis, regression and clustering. Tableau Desktop is also capable of integrating with external connections like Python and R libraries for performing advanced statistical analysis.

The data flow architecture suitable for La Place within the available technology landscape can be seen in Figure 5.

A data flow is built in Tableau Prep and exported as a .hyper file or a .xlsx file. The structure of this data flow can be seen in Appendix B. This file is imported into Tableau desktop to visualize the results. Finally, Tableau desktop is connected with an external service for R/Python and Microsoft Excel to perform some statistical operations like clustering and implementing the random forest classifier.

[9] https://powerbi.microsoft.com/en-us/

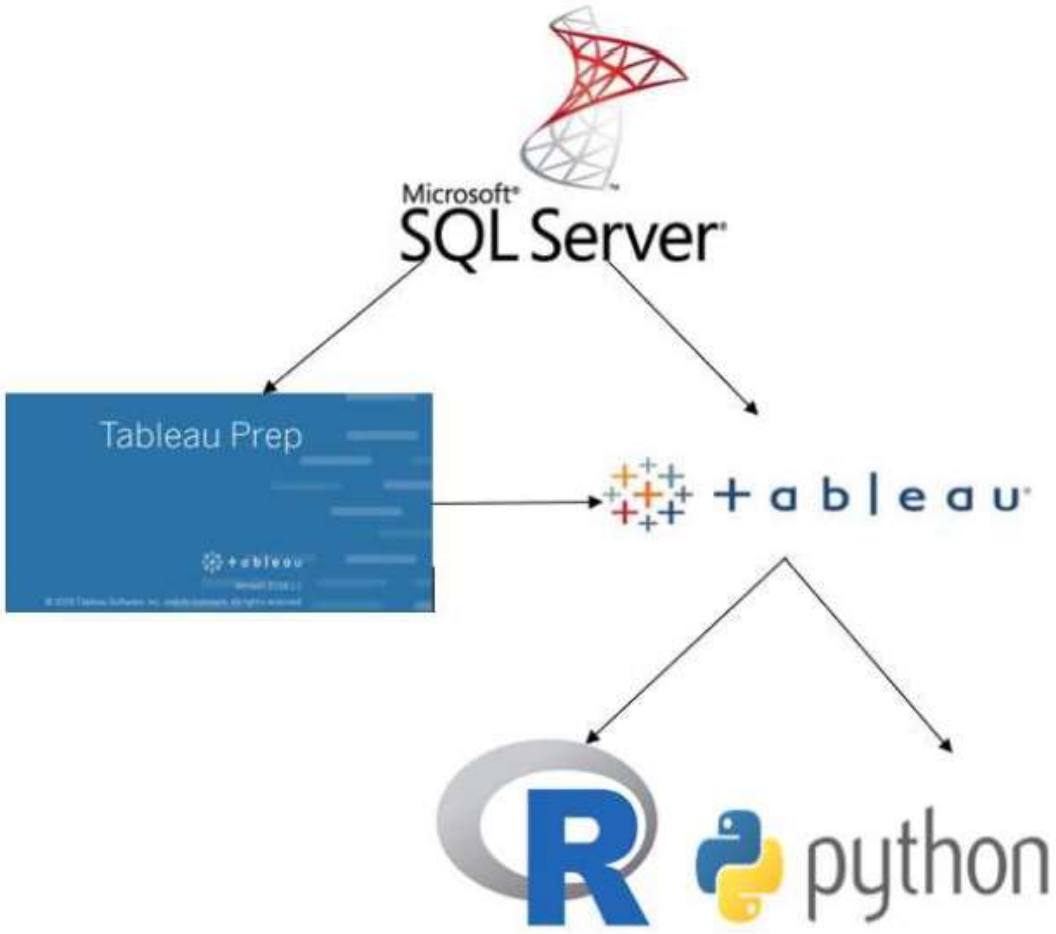


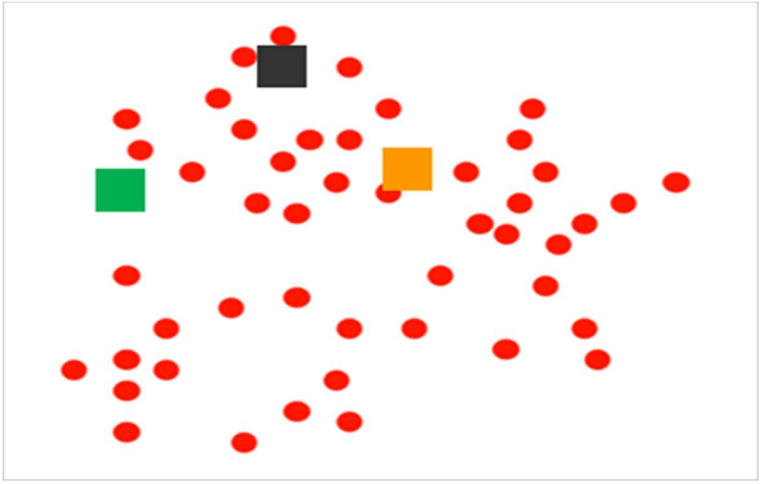
Figure 6 Data flow architecture for La Place

4.4 The Tableau version of the K-means clustering algorithm

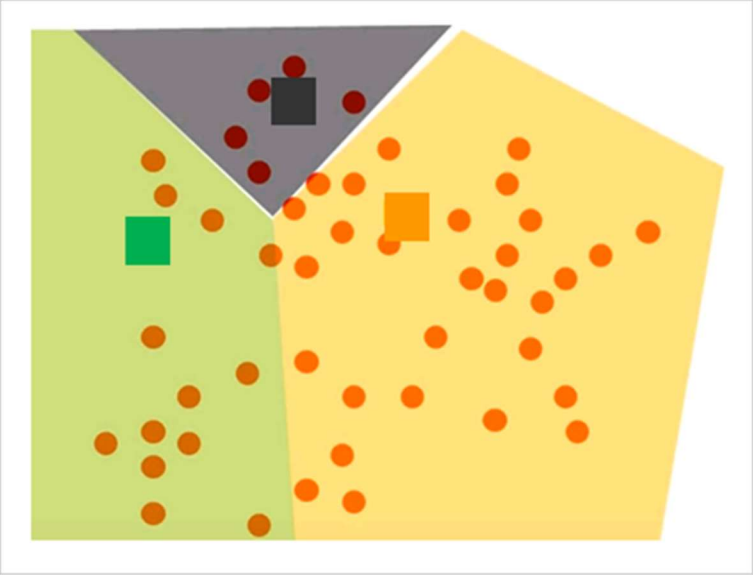
Since Tableau was used for implementing the K-means clustering algorithm, the focus is on the Tableau version of this algorithm. For a given number of clusters k, the algorithm partitions the data into k clusters. Each cluster has a center (centroid) that is the mean value of all the points in that cluster. K-means locates centers through an iterative procedure that minimizes distances between individual points in a cluster and the cluster center. In Tableau, you can specify a desired number of clusters, or have Tableau test different values of k and suggest an optimal number of clusters.

Tableau uses Lloyd's algorithm with squared Euclidean distances to compute the k-means clustering for each k. Combined with the splitting procedure to determine the initial centers for each k > 1, the resulting clustering is deterministic, with the result dependent only on the number of clusters.

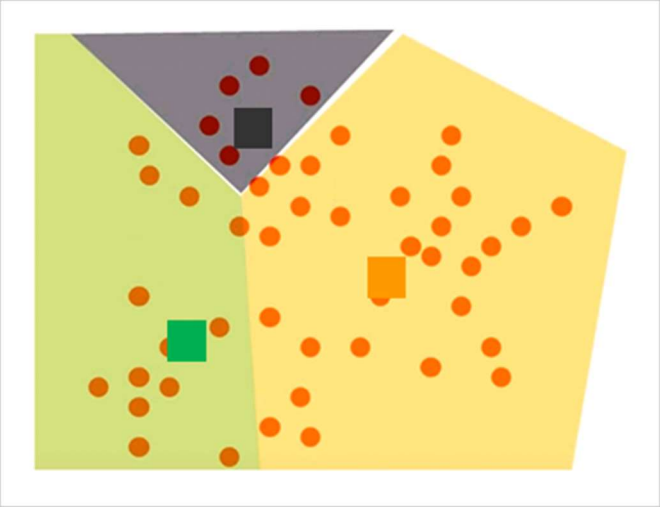
The algorithm starts by picking initial cluster centers:



It then partitions the marks by assigning each to its nearest center:



Then it refines the results by computing new centers for each partition by averaging all the points assigned to the same cluster:



It then reviews the assignment of marks to clusters and reassigns any marks that are now closer to a different center than before. The clusters are redefined, and marks are reassigned iteratively until no more changes are occurring.

Criteria used to determine the optimal number of clusters

Tableau uses the Calinski-Harabasz criterion to assess cluster quality. The Calinski-Harabasz criterion is defined as:

$$\frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

where SS_B is the overall between-cluster variance, SS_W the overall within-cluster variance, k the number of clusters, and N the number of observations.

The greater the value of this ratio, the more cohesive the clusters (low within-cluster variance) and the more distinct/separate the individual clusters (high between-cluster variance). Since the Calinski-Harabasz index is not defined for k=1, it cannot be used to detect one-cluster cases.

If a user does not specify the number of clusters, Tableau picks the number of clusters corresponding to the first local maximum of the Calinski-Harabasz index. By default, k-means will be run for up to 25 clusters if the first local maximum of the index is not reached for a smaller value of k. The maximum value of clusters can be set to 50.

4.5 The Tableau Prep flow

The main part of the proposed method would be the connection between the BI tool and the data source. This is where most of the calculations, transformations and computations would take place. Since Tableau is the main BI tool selected for this research, the heart of our classification scheme is the flow created in Tableau Prep. This flow can connect to the SQL server anytime and read transactional data of customers up to the previous day.

The idea behind developing the flow was that it should be able to connect to the Live SQL server and extract transactional data for each customer. The process diagram of the Tableau Prep flow can be seen in Figure 7. In this figure, the boxes indicate activities and the arrows show the sequence in which the activities are executed. The complete flow can be seen in APPENDIX B. Below is a description of the steps in the flow:

1. Read 'Customer' and 'Transaction' tables. The Customer table had data about 1.2 million customers. This included information about their age, year of birth gender, city of residence, Customer Card number, registration date and so on. It was seen that data for all data fields was not available for all customers. For example, the age and year of birth was missing for about 470k customers and for some others, there were incorrect entries like age greater than 100 years or less than 5 years. Therefore, data in this table had to be cleaned and the cleaning operations mainly involved removing the incorrect birth years of customer, along with most of the other unnecessary fields which were not required in the design process. Customers only with birth years between 1930-2010 have been considered for further analysis and all other years of birth have been treated as NULL. Age of the customers has also been recalculated based on the correct birth year and only customers aged between 9-89 years have been considered. Eventually, we ended up with around 700k customers for whom we had the correct age data. The reason there was so much emphasis on age was because the Marketing team strongly believed in a classification based on age/age groups of customers. Customers were labelled into different age groups based on the definitions in Table 2.

   The 'Transactions' table was read. This table had information about all 60 million transactions since 2015. This table has transactional information like the transaction number, customer card number, paid amount per transaction, number of products, transaction date/ time and so on. However, the main focus here is on the customers who made the transactions. This table only had information about 820k customers, indicating that around 400k customers had never made a transaction even though we had data

about them in the 'Customers' table. So, these 820k customers formed our customer base because this whole research aimed to classify only those customers for whom there was some transactional history.

2. The 'Transactions' table had very limited information about the customers who made those transactions. Therefore, the 'Customers' and 'Transactions' tables were joined to get all information about those 820k customers. So, this gave complete information about each transaction and the customer who made it.
3. The 'Transactions' table into two parts. This is mainly for the 'Recency' calculations. First part has those transactions which were recorded within the last 13 weeks and the second part had all transactions older than 13 weeks. We found around 400k transactions by approximately 120k customers in the last 13 weeks and around 8 million transactions by around 800k customers. It must be specified here that these transaction numbers are aggregated by customer number.
4. The Lifetime, Recency and Monetary values were calculated for each customer in this step. These are based on the definitions in Table 2.
5. The next step was finding the 'Preferred Day' of visiting La Place for these 820k customers. A day of the week is preferred by a customer if the number of transactions made on that specific weekday is a clear majority (more than 50%) compared to the total transactions made by the customer. For example, Friday is a preferred day for a customer with a total of 10 transactions if the customer has made minimum 6 transactions on Friday. In order to make this split per weekday, total transactions of each customer were divided based on the weekday on which they were recorded. Eventually, 7 splits were recorded for each weekday and there were obviously overlapping customers with transactions on different weekdays. With these splits, it was possible not only to find how many transactions were made by each customer on each day of the week, but also a combination of their visits on different weekdays. These splits were brought together to find the 'Preferred weekday' for each customer.
6. Like step 7, the 'Preferred Daypart' of visiting La Place for each customer was computed. A daypart was preferred by a customer if the number of transactions made on a specific daypart is a clear majority (more than 50%) compared to the total transactions made by the customer. For example, Dinner is a preferred daypart for a customer with a total of 5 transactions if the customer has made minimum 3 dinner transactions. Eventually, there were 4 splits for dayparts, namely 'Breakfast, 'Lunch', 'Midday' and 'Dinner'. Again, there were obviously overlapping customers between dayparts. With these splits, it is possible not only find how many transactions were made by each customer on each daypart, but also a combination of their visits on different dayparts. These splits were brought together to find the 'Preferred daypart' for each customer.

7. Results of steps 7 and 8 were combined to get the preferred weekday and daypart for each of the 820k customers.
8. Steps 1, 4 and 7 are combined to get all the defined labels and attributes for all of the 820k customers, and the output is saved as a .hyper extract file.

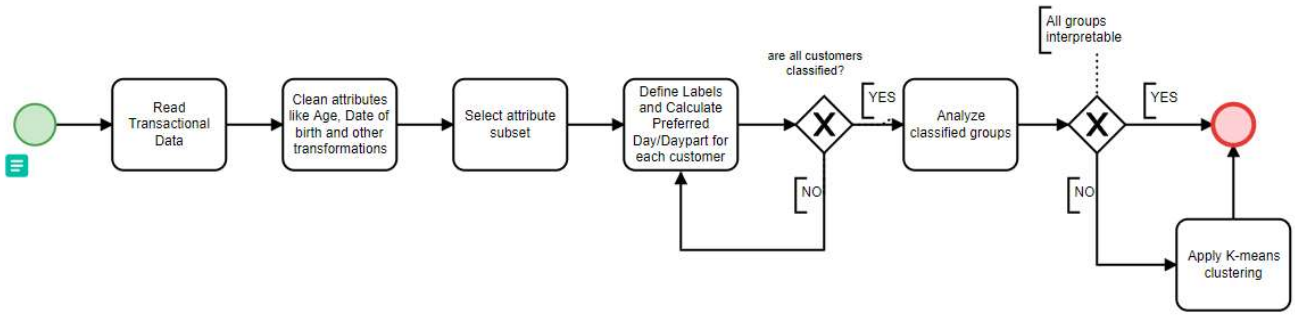


Figure 7 Process diagram of Tableau Prep Flow

Chapter 5. RESULTS

An overview of the customer base can be seen in Figure 8. The overview provides information about the different attributes like the labels defined based on Recency, Age Groups, Lifetime and Monetary values. The design of the overview was a free choice and was designed keeping in mind that it should provide information about Lifetime (L), Recency (R) and Age Group labels in addition to overall information about the number of customers , their age and how much they spend on an average. This overview would help the business in getting a good insight into the numbers and can be further drilled down if required.

Based on 'Recency', the 'Lost' group of customers is the largest and the 'Redormed' groups are the smallest. This is majorly because of the business takeover that happened in early 2016 which caused the closing down of most of the stores across the Netherlands. Also, a lot of other stores operated by the V&D group were shut down. This could be a major reason for a considerably large proportion of users being lost, at almost 59%. Around 14% of the customers are classified in the 'Monitor' group, 13% as 'Inactive' group and 12% as 'Frequent' visitors. The addition of new customers during last 13 weeks was about 1.5% of the total customer base and almost the same proportion of customers were 'Redormed'.

Based on the 'Age groups', it is seen that most customers haven't revealed their date of birth or have mentioned incorrect values. Therefore, almost 34% of the customers fall in the 'Unknown Age' category. Out of all the customers whose age are known, the 'Baby Boomers' are the

majority, justifying the average age of the whole customer base to be on the higher side, at around 49 years.

The 'Monetary' value was computed based on the average spending of each customer on a transaction at La Place. These are aggregated values. So, if a customer has made 10 transactions in total spending 100 Euros, the average ticket value for the customer would be 100/10=10 Euros. The Average ticket value of the entire customer base turns out to be Euros 9.47.

Based on 'Lifetime', the biggest group is formed by the 'OLD' group, which is in sync with the 'Lost' group based on 'Recency'.

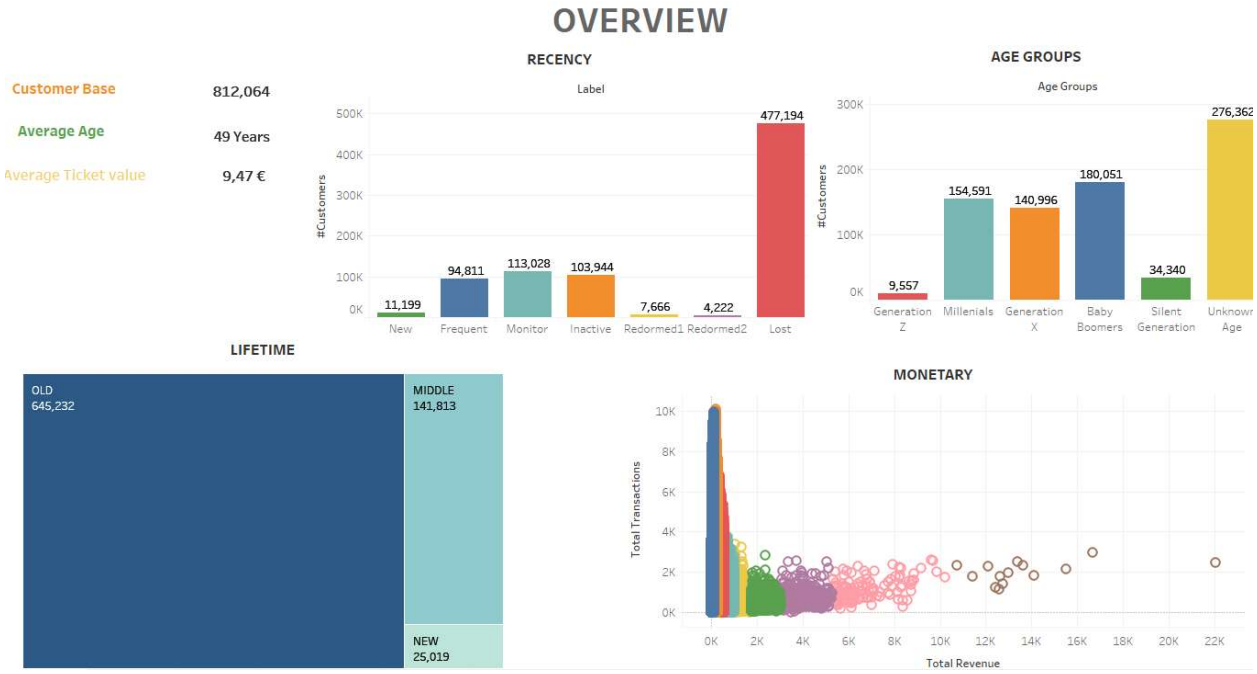


Figure 8 Customer Classification Overview

As per the Pareto principle, it is generally observed that around 20% of customers contribute to 80% of the companies revenue and it is important to know who these customers are [3]. In order to focus on the most important group of customers first, the total revenue in the monitored period was split amount the groups based on 'Recency' and it was expectedly found that the 'Frequent' customers are the biggest contributors to the company's revenue. They contributed about 44% to the total revenue even though they only formed just about 12% of the customer base, as seen in Figure 9. So, our main focus group is the group of 'Frequent' customers, followed by the other 3 groups (Monitor, New and Redormed).

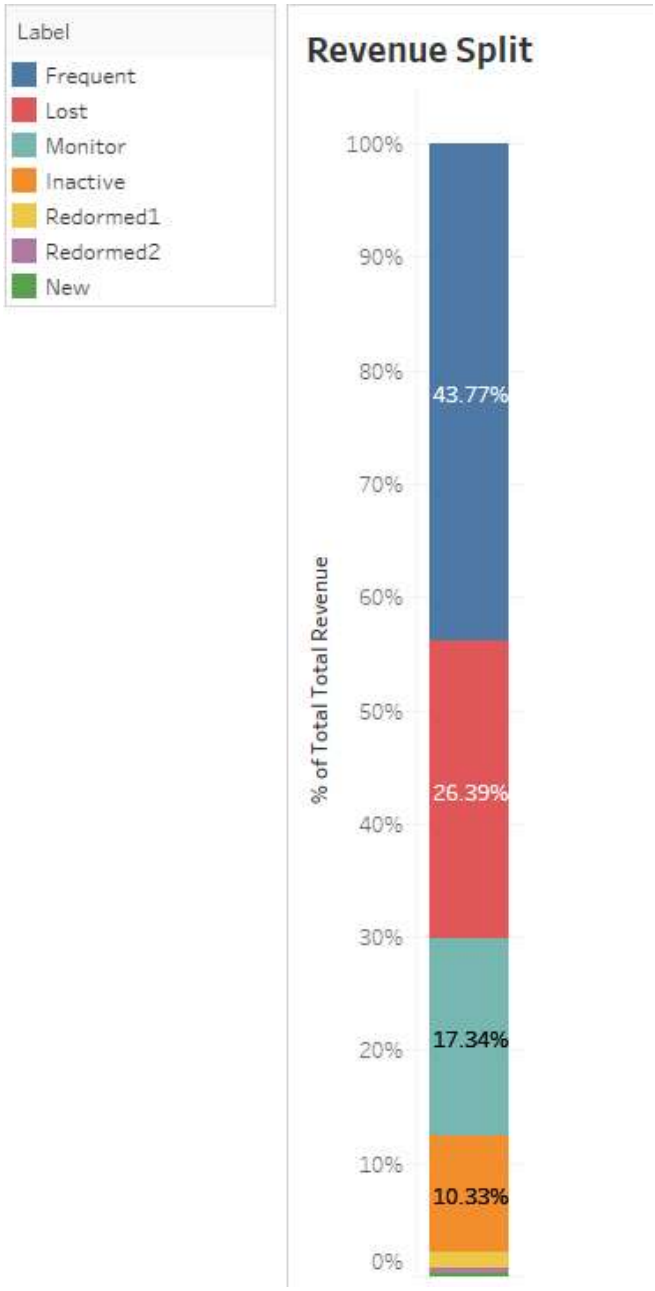


Figure 9 Revenue split based on Recency

5.1 Analysis of the groups

This section analyses 4 different groups, namely the Frequent visitors, Monitor group, Redormed customers and the New customers. This is increasing order of importance of these groups based on their contribution to total revenue and their business value. An overview of each

group is provided to begin with, followed by results from the classification scheme. This is followed by cluster analysis and identified sub-groups are summarized.

5.1.1 Frequent Visitors

An overview of 'Frequent' visitors can be seen in Figure 10. On analyzing the 'Frequent' visitors, some interesting facts were revealed. The average age of this groups is 3 years higher than the average age of the entire customer base. Also, it is observed that almost 35% of the Frequent visitors belong to the 'Baby Boomers' age group. This is considerably higher than the entire customer base, where this age group comprises about 23%. The average number of transactions by Frequent visitors in the last 13 weeks was 3.8.

It was observed that the preferences of visiting La Place for a specific part of the day was the similar between Breakfast, Lunch and Midday. Dinner was the least preferred part of the day. The biggest group had no preference, and this made sense as frequent customers are expected to visit throughout the day because of their liking for the brand and possible proximity of the store.

Similar to daypart preferences, an analysis was done for the day of the week. Saturday was clearly the most preferred day of the week, but the largest group was again formed by the ones which had no preference.

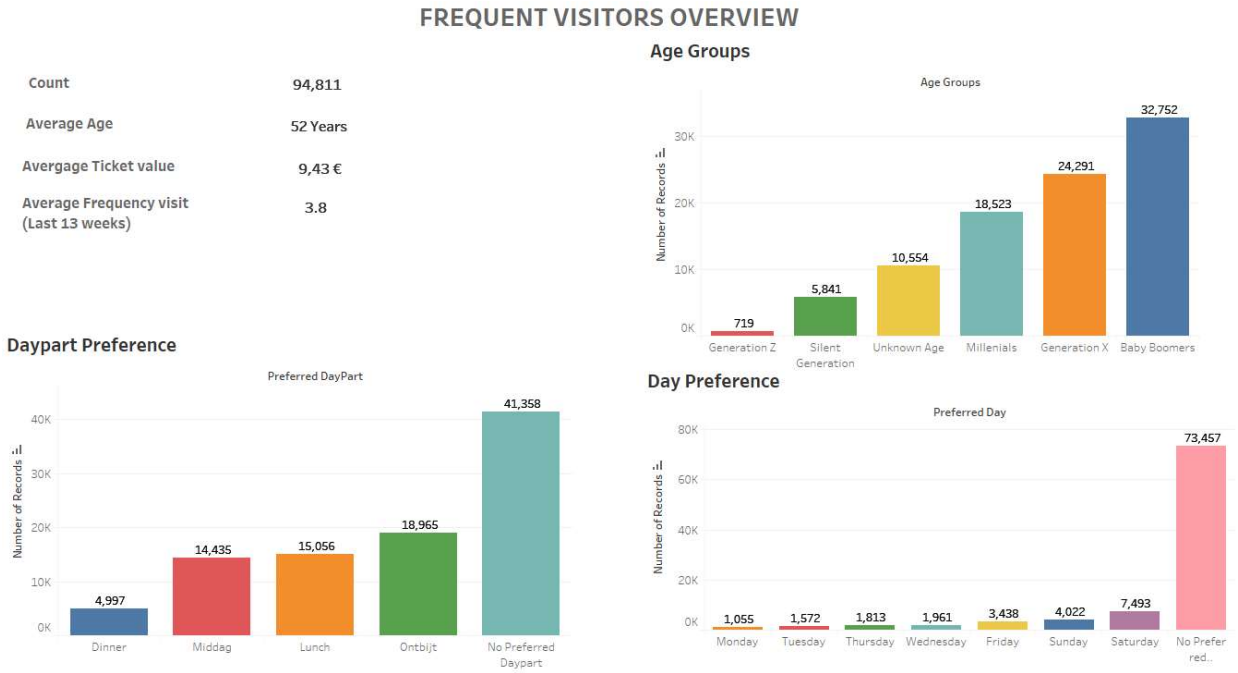


Figure 10 Frequent Visitors Overview

In this case, if we would like to recommend a group of customers to the marketing team to reach out, for an upcoming 'Lunch' deal, it is evident that the group that prefers 'Lunch' would be the main focus group. But we shouldn't forget the biggest group of 'Frequent' visitors which has 'No preference'. Based on their behavior, it could be possible to map some or most of them to possibly have a preference for Lunch. This is where clustering can help.

The group which had 'No preference' for a specific part of the day was clustered using K-means clustering algorithm. This group had around 41k customers and the goal of clustering was to make an inference about their preferred part of the day to visit La Place based on already known information. The variables that were considered for clustering were the Average ticket size, the average ticket value and total transactions. The number of clusters were set to 5 as this gave distinct results unlike 2 clusters which was selected by Tableau automatically. Ticket size is defined as the number of products in each ticket, used as an indicator regarding how many people ate in that transaction. The belief is that customers generally eat more on lunch and dinner than on breakfast and thereby also spend more money. This belief was justified in Figure 11. These variables proved to be distinctive while interpreting the cluster results and can be seen by the ANOVA statistics in Figure 13. The summary diagnostics of clustering results for 'Frequent visitors' without a day preference are shown in Figure 12. Details regarding the ANOVA statistics and summary statistics of clustering can be found in the APPENDIX.

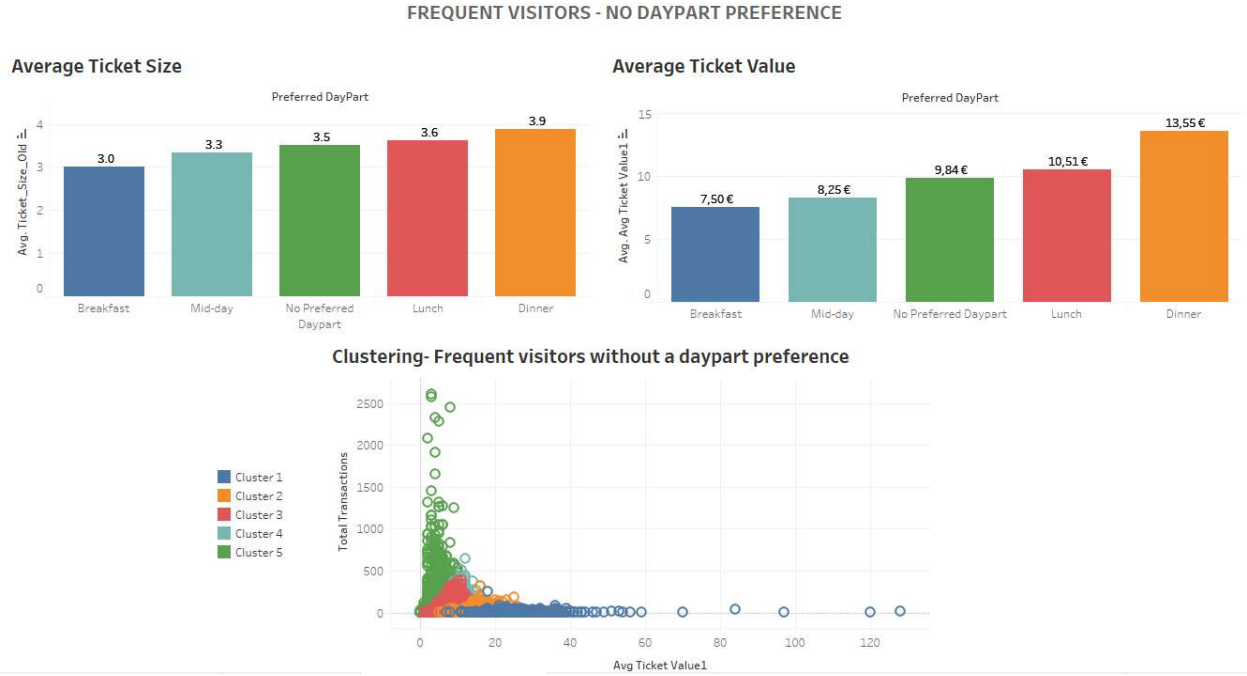


Figure 11 Frequent visitors - No daypart preference

Summary Diagnostics

Number of Clusters:	5
Number of Points:	41358
Between-group Sum of Squares:	100.1
Within-group Sum of Squares:	51.351
Total Sum of Squares:	151.45

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_Old
Cluster 1	15157	8.253	30.101	3.1512
Cluster 2	11384	11.962	25.837	4.0442
Cluster 3	9359	5.0925	62.583	2.2251
Cluster 4	4760	16.843	18.189	5.2631
Cluster 5	698	25.672	10.361	8.072
Not Clustered	0			

Figure 12 Summary statistics - Clustering Frequent visitors without a daypart preference

Analysis of Variance:

			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum of Ticket_Size_Old	8223.0	0.0	55.34	4	69.57	41353
Sum of Avg Ticket Value	7926.0	0.0	43.26	4	56.43	41353
Sum of Total Transactions	608.4	0.0	1.498	4	25.45	41353

Figure 13 ANOVA statistics - Clustering Frequent visitors without a daypart preference

Clearly, clusters 4 and 5 have the highest Ticket size and ticket value, as seen in Figure 12. This is an indicator that these clusters have a strong preference for 'Dinner' as their preferred daypart. Cluster 2 has the ticket size higher marginally higher than dinner making it difficult to infer a preference, but the average ticket value is relatively closer to lunch than dinner. Also, the size of the group is sufficiently large, which probably indicates that this cluster prefers lunch over dinner as more people visit during lunch rather than dinner. Similarly, cluster 3 has the lowest ticket size and ticket value, but the highest number of transactions. This is clearly an indicator of this cluster preferring 'Breakfast'. Cluster 1 has a ticket size which is close to both breakfast and midday, however, the ticket value is closer to midday than breakfast. So, this is possibly a group with a preference for midday rather than breakfast and the lesser number of transactions compared to breakfast support this claim. Therefore, it can be seen that in this case, it is possible to map clusters to their preferred daypart. This is mainly achieved because of knowledge about the variables selected and inductive reasoning. The biggest positive is that inferences can be made about almost all clusters using the cluster analysis.

So, as a recommendation to marketing team for an upcoming breakfast deal, they can consider the following groups: approximately 19k customers who have a clear preference for breakfast and approximately 9k customers revealed using clustering results (cluster 3).

For an upcoming lunch deal, the marketing team should consider the following groups: approximately 15k customers who preferred lunch and about 11k customers revealed using clustering results (cluster 2).

Similarly, for a dinner deal, the marketing team should consider the following groups: approximately 5k customers who clearly prefer dinner and about 5.5k customers revealed using clustering results (clusters 4 and 5).

It is worth mentioning that, the results from clustering are still substantially big groups and can be further drilled down. This depends on other criteria like who does the marketing team want to target, how many of them and so on. In order to do this, it is possible that additional data may be required such as availability of means of reaching those customers.

Also, for the 'Frequent' visitors, it is for known that they will keep visiting. But what needs to be given them as incentive is a reason to visit more and that is how they will keep returning and may be even with new accompanies. A simple offer like a free apple pie on the 5^{th} coffee in a week or a free sandwich after 3 lunch deals in a week, would increase the curiosity among them and make them look forward to reaping benefits. It is certain that the promotional strategies used for this group of customers would be very different from that of other groups.

A similar cluster analysis was done for the 'Frequent visitors' who did not prefer visiting La Place on a specific day of the week, which can be seen in Figure 14. This was a substantially big group with around 73k customers. The same variables were used for clustering. However, the results with only 5 clusters were not distinguishing enough. When the number of clusters were set between 6 to 9, it did not really help in distinguishing the clusters based on the chosen variables. Therefore, the number of clusters were set to 10, which resulted in better distinguished clusters. This helped in achieving better results and make better inductions to the available knowledge about these variables. The choice of the number of clusters really depends on specific needs for example, how big a cluster should be or which are the highest spenders and so on.

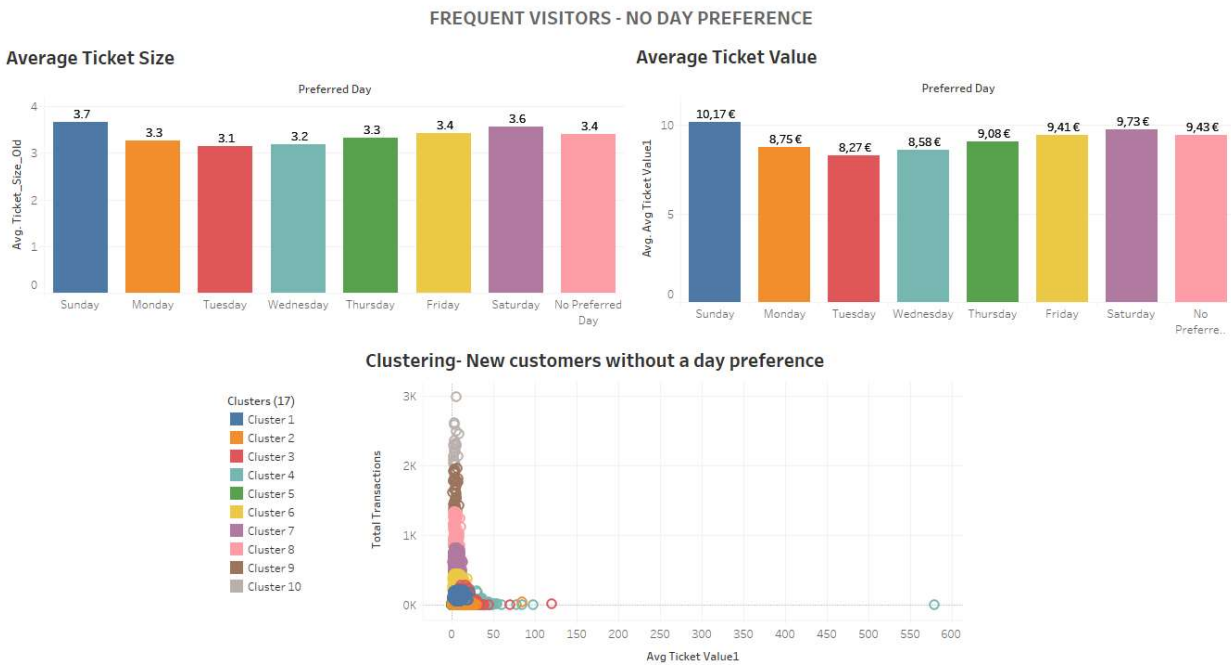


Figure 14 Frequent visitors - No day preference

Summary Diagnostics

Number of Clusters:	10
Number of Points:	73457
Between-group Sum of Squares:	169.96
Within-group Sum of Squares:	30.811
Total Sum of Squares:	200.77

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_Old
Cluster 1	27504	9.1644	26.195	3.3599
Cluster 2	18688	5.9224	25.31	2.237
Cluster 3	5333	6.3795	116.83	2.6899
Cluster 4	16873	13.315	22.019	4.5475
Cluster 5	2938	20.475	13.592	6.9178
Cluster 6	1456	5.4389	283.77	2.5114
Cluster 7	466	4.6567	585.71	2.3862
Cluster 8	142	4.0986	1048.4	2.2729
Cluster 9	37	4.027	1631.8	2.1909
Cluster 10	20	3.8	2291.9	2.1817
Not Clustered	0			

Figure 15 Summary statistics - Clustering Frequent visitors without a day preference

Analysis of Variance:

Variable	F-statistic	p-value	Model Sum of Squares	Model DF	Error Sum of Squares	Error DF
Sum of Total Transactions	7520.0	0.0	68.87	9	74.74	73447
Sum of Ticket_Size_Old	6675.0	0.0	98.29	9	120.2	73447
Sum of Avg Ticket Value	3899.0	0.0	2.796	9	5.853	73447

Figure 16 ANOVA statistics - Clustering Frequent visitors without a day preference

Clearly, clusters 4 and 5 showed weekend preferences because of high ticket size and ticket value (Figure 15). This is probably because they did not visit alone, but with their families, friends and/or colleagues, as expected on weekends. Cluster 1 had a relatively high ticket size within the weekdays, and this is an indicator that this cluster prefers to visit mostly on Thursday and Friday, when the average spend is higher than other weekdays. For the remaining clusters, the inferences were difficult to make as there was very less to distinguish between them. However, the positive part is that Clusters 1,4 and 5 which could be mapped to specific preferred days of the week, constituted about 65% of this group and additional knowledge about them was revealed with the help of this cluster analysis. In this way, clustering helped knowing more about a substantially big group of customers. The variables used for clustering helped in creating distinct clusters and ANOVA statistics in Figure 16 support this claim.

5.1.2 Monitor Group

The 'Monitor' Group happens to be the second most important group. This is majorly because this group has the potential to turn into 'Frequent visitors' if they are focused upon. On the other hand, they could also become 'Inactive', which is something La Place would like to avoid. About 15% of the customers form this group and they contribute to about 17% of the total revenue. If they become 'Frequent visitors', their contribution to the total revenue would substantially increase. Therefore, it's vital to do an analysis about them. The average age of this group is 2 years lesser than that of the entire customer base. Unlike the 'Frequent' visitors which had the 'Baby boomers' as the biggest age group, the 'Monitor' group has the 'Millenial' as the biggest age group, followed closely by 'Baby Boomers' and 'Generatie X'. This indicates that mostly young customers constitute this group. Also, this strengthens the fact that the die-hard fans of La Place are the relatively older people i.e. Baby boomers, as seen in 'Frequent' visitors. Younger people are more prone to try different stores and be more dynamic in their choices. Therefore, the means of targeting these customers can be imagined to be different than those of

the ‘Frequent’ group. An overview of the ‘Monitor’ group can be seen in Figure 17. The average ticket value of this group is 60 cents higher than the ‘Frequent’ visitors.

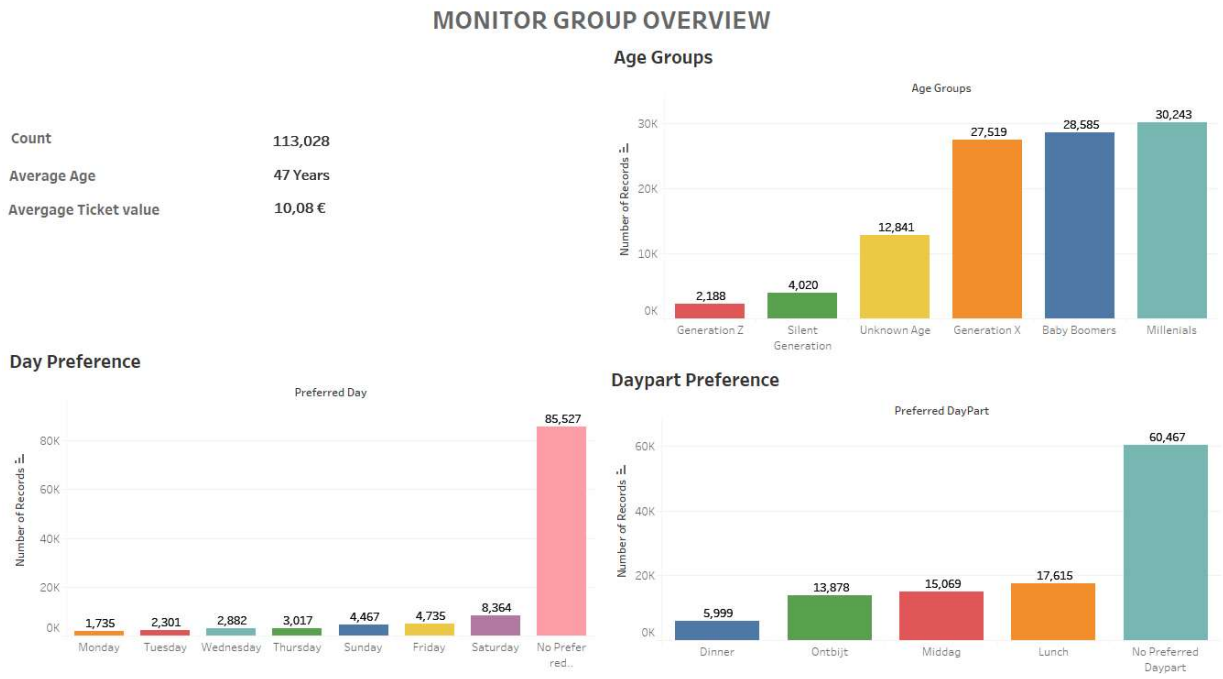


Figure 17 Monitor Group Overview

Again, it was observed that a majority of this group does not have a specific preference for a day of the week or a part of the day. So, cluster analysis was used to infer what kind of preferences customers in these groups would have. The variables that were considered for clustering were the Average ticket size, the average ticket value and total transactions. The number of clusters were set to 5 as this gave distinct results unlike 2 clusters which was selected by Tableau automatically. For approximately 60k customers without a daypart preference, clusters 3,4 and 5 indicated a preference for dinner with a high ticket size and ticket price (Figure 19). Cluster 1 had the lowest ticket size and ticket price, indicating a preference for breakfast. It was difficult to indicate the preference for cluster 2. Using this analysis, inferences about almost 70% of the customers of this group could be made.

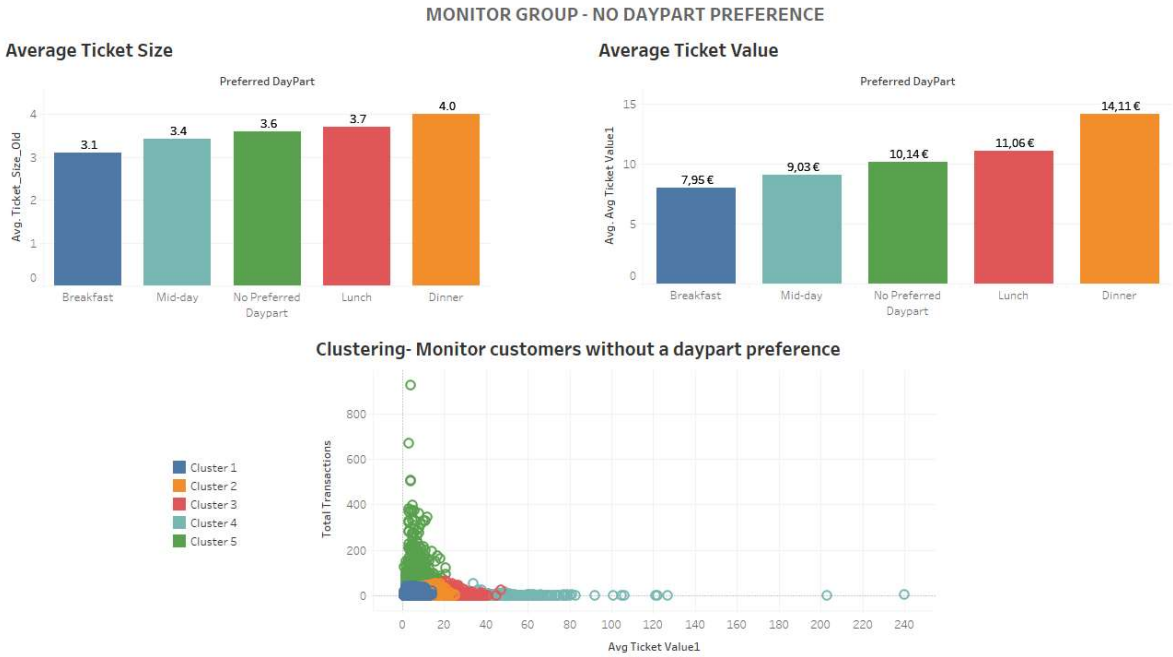


Figure 18 Monitor Group - No daypart preference

Summary Diagnostics

Number of Clusters:	5
Number of Points:	60467
Between-group Sum of Squares:	77.488
Within-group Sum of Squares:	37.075
Total Sum of Squares:	114.56

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_Old
Cluster 1	16280	3.3017	2.9007	1.9269
Cluster 2	19856	8.3138	18.523	3.1299
Cluster 3	18179	13.266	6.5221	4.4087
Cluster 4	5420	22.455	4.3042	6.537
Cluster 5	732	42.892	1.6462	11.469
Not Clustered	0			

Figure 19 Summary statistics - Clustering Monitor group without a daypart preference

Out of the approximately 85k customers without a day preference as seen in Figure 20, cluster analysis revealed that clusters 2,3,5 and 6 had high values for ticket size and ticket value (Figure 21). In this case, same variables were used for clustering however, the number of clusters were set to 6 to get distinctness between them. This indicated a preference for visiting on weekends and constituted about 30% of this group. Cluster 4 with approximately 25k customers

had the lowest ticket value and ticket size, indicating a preference for visiting early in the week. Cluster 1 with approximately around 30k customers had higher than average ticket values and ticket size compared to Monday until Wednesday, indicating a preference for visiting mostly on Thursday and Friday.

Figure 20 Monitor Group - No day preference

Summary Diagnostics

Number of Clusters:	6
Number of Points:	81163
Between-group Sum of Squares:	89.72
Within-group Sum of Squares:	29.522
Total Sum of Squares:	119.24

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_Old
Cluster 1	30675	9.4528	12.96	3.4835
Cluster 2	18198	14.675	8.7731	4.66
Cluster 3	5795	22.569	5.2925	6.5241
Cluster 4	25410	5.1677	16.456	2.2654
Cluster 5	1059	39.044	2.2162	10.273
Cluster 6	26	98.654	1.2692	27.865
Not Clustered	0			

Figure 21 Summary statistics - Clustering Monitor group without a day preference

Based on the preferences, similar marketing strategies can be used as listed for the 'Frequent' visitors. However, the strategies for this group needs to be more appealing and compelling as they have clearly not been interested in visiting over the last 13 weeks.

The 'Monitor' group is a high potential group and additional efforts need to be taken to ensure that they do not fall out into the 'Inactive' group of customers. Based on their transactional history, they should receive personalized offers in addition to the normal ones. After all, it will very difficult to turn them again into 'Frequent' visitors and in order to do so, they should be given a reason to return which is strong enough and they can't resist.

5.1.3 Redormed Group

This forms a very interesting group because these are customers who were not expected to visit based on their last transaction date. However, they turned up in the last 13 weeks. The Inactive customers turned into 'Redormed1' and the Lost customers into 'Redormed2'. Both these groups are combined for analysis into one 'Redormed' group. An overview can be seen in Figure 22.

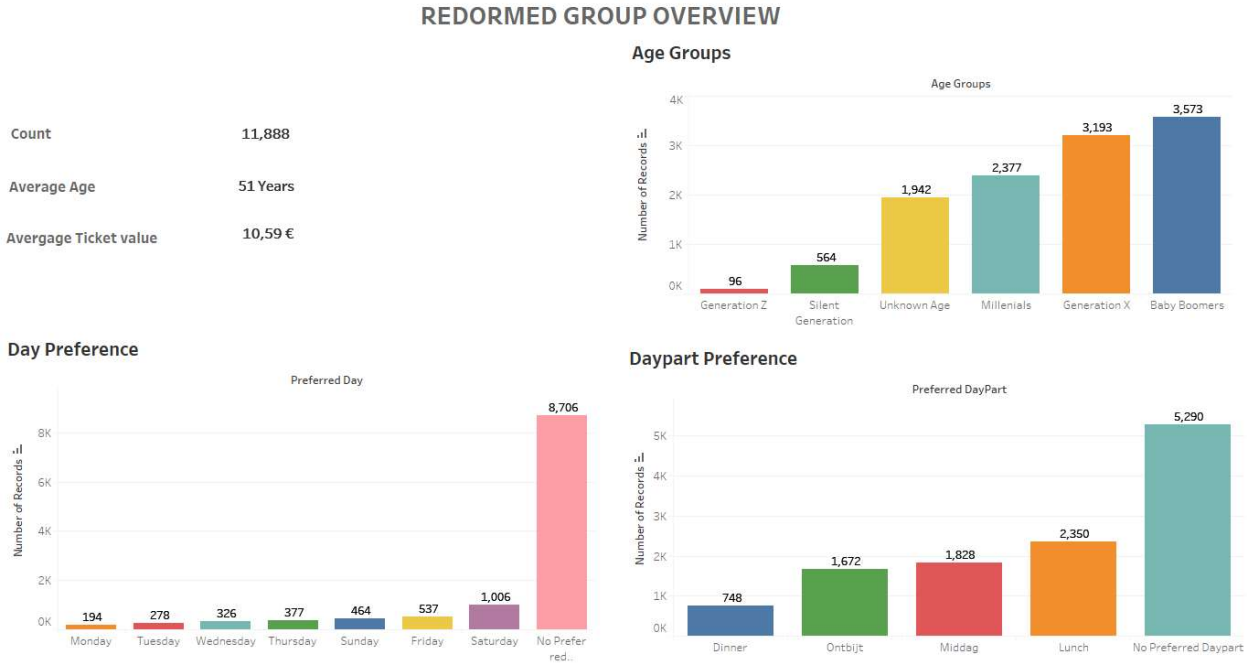


Figure 22 Redormed Group Overview

The most important thing while trying to understand this group is the reason why they turned up in the last 13 weeks. This analysis was done considering the 13-week period between June-August 2019, which also happens to fall between the vacation period in the Netherlands. So, this is probably just one or two visits while they were away on vacation. This was clearly seen in the data when it was revealed that almost 75% of the 'Redormed' customers just made one transaction in the last 13 weeks and had an average ticket value of Euros 13.45, which is almost 50% higher than the average ticket value of the entire customer base. The average number of products for this group is 4, which is also higher than average. The two major age groups are 'Generatie X' and 'Baby Boomer', which are the groups highly likely to visit with families.

Cluster analysis was done on the groups without a preference for a daypart and day of the week. Almost 5.3k customers did not have a preference for a part of the day to Visit La Place. The variables that were considered for clustering were the Average ticket size, the average ticket value and total transactions. The number of clusters were set to 5 as this gave distinct results unlike 2 clusters which was selected by Tableau automatically. Upon using K-means clustering, it was found that clusters 2 and 5 seemed to have a preference for 'Dinner' based on the high ticket size and ticket value, as seen in Figure 24. Cluster 3 did not have a ticket size to match that of Dinner on average, but the ticket value indicated otherwise. Cluster 4 was clearly preferring 'Breakfast'. It was difficult to interpret the preference for cluster 1. Nonetheless, clustering helped in inducing conclusions about almost 60% of the customers without a preference for a part of the day. Results can be seen in Figure 23.

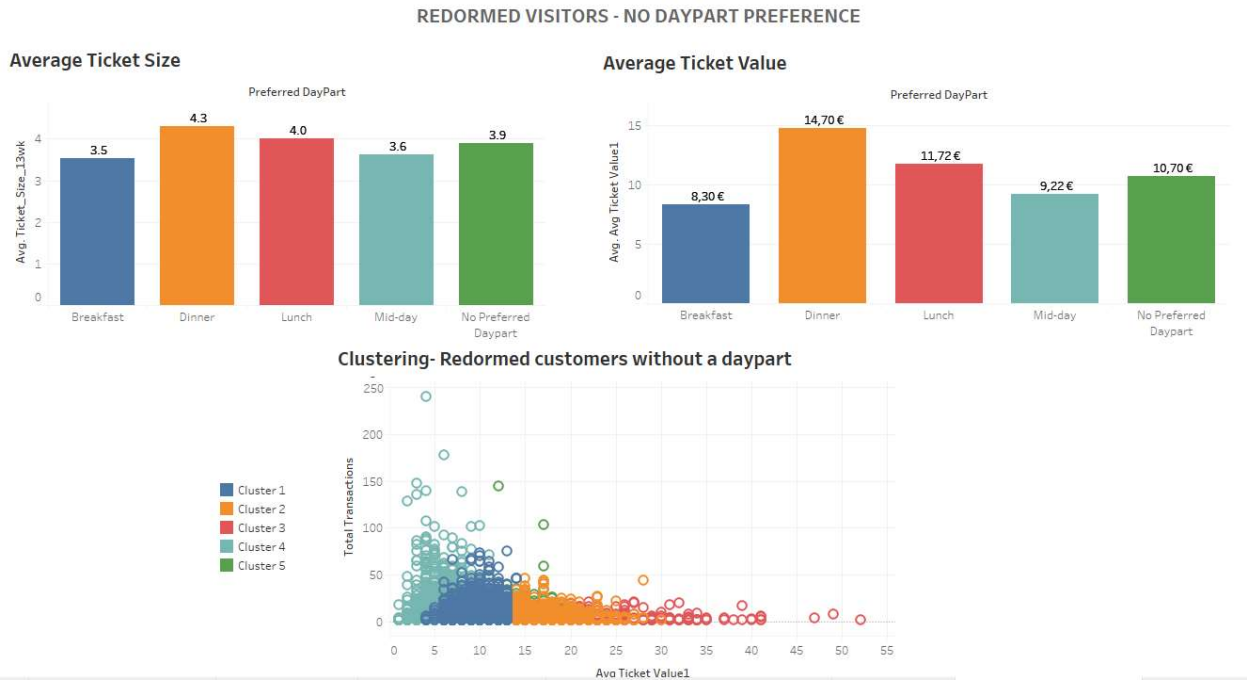


Figure 23 Redormed Group - No daypart preference

Summary Diagnostics

Number of Clusters:	5
Number of Points:	5260
Between-group Sum of Squares:	76.926
Within-group Sum of Squares:	42.211
Total Sum of Squares:	119.14

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_13wk
Cluster 1	2158	9.9495	10.293	3.7955
Cluster 2	178	26.978	5.4382	8.9925
Cluster 3	799	17.111	7.4743	3.9329
Cluster 4	1539	5.8161	14.705	2.2197
Cluster 5	586	13.181	11.234	7.2672
Not Clustered	0			

Figure 24 Summary statistics - Clustering Redormed group without a daypart preference

Unlike other groups, the average ticket value for this group was relatively high on Monday, as seen in Figure 26. This is also possibly an indicator or just one-time visits or visiting with a group. Clusters 2,3 and 6 had high values for the clustering variables and indicated a preference towards Friday or Sunday. Cluster 5 clearly visited during weekdays (Tuesday-Thursday). Overall, it can be said that for this group cluster analysis did not provide conclusive evidence for a majority of customers about the preferred day of the week. Results can be seen in Figure 25.

Figure 25 Redormed Group - No day preference

Summary Diagnostics

Number of Clusters:	6
Number of Points:	8668
Between-group Sum of Squares:	54.651
Within-group Sum of Squares:	20.38
Total Sum of Squares:	75.031

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_13wk
Cluster 1	1590	15.303	8.3981	3.9017
Cluster 2	92	35.098	4.1196	8.7192
Cluster 3	465	22.29	6.5269	6.6728
Cluster 4	3351	9.6464	11.992	3.6836
Cluster 5	2516	5.4479	18.103	2.4022
Cluster 6	654	12.792	11.758	7.654
Not Clustered	0			

Figure 26 Summary statistics - Clustering Redormed group without a day preference

5.1.4 New Customers

This is a relatively unknown group in terms of transactional history. It is highly likely that they registered at La Place because of an on-going promotional activity but this not known due to the lack of availability of data (as of now). 'Millenial' is the biggest age group, unlike other

'Recency' based groups. In general, the ticket sizes and ticket values on specific parts of the day and days of the week are lower than those of other groups. An overview of 'New' customers can be seen in Figure 27.

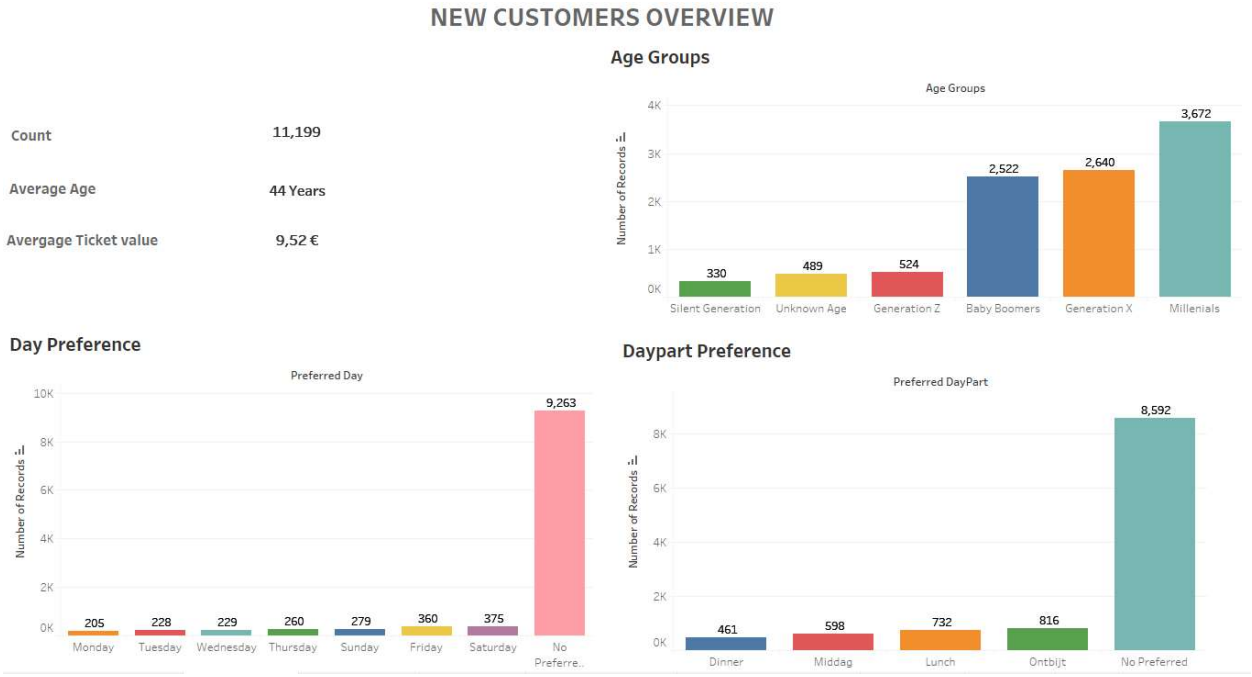


Figure 27 New customers Overview

Cluster analysis was done on the groups without a preference for a daypart and day of the week. The variables that were considered for clustering were the Average ticket size, the average ticket value and total transactions. The number of clusters were set to 5 as this gave distinct results unlike 2 clusters which was selected by Tableau automatically. Almost 8.5k customers did not prefer a part of the day to Visit La Place, which can be seen in Figure 28. Upon using K-means clustering, it was found that clusters 3,4 and 5 seemed to prefer 'Dinner' based on the high ticket size and ticket value, as seen in Figure 29. Cluster 1 had a clear preference for 'Breakfast, with the lowest ticket value, even though the ticket size indicated a preference for 'Midday'. This is concluded based on the average price of products in the transactions (Ticket value/Ticket size). Cluster 2 was difficult to conclude to have preference towards 'Midday' or 'Lunch' but more data points could have helped. So, the daypart preference for about 60% of this group was deducible based on the cluster analysis.

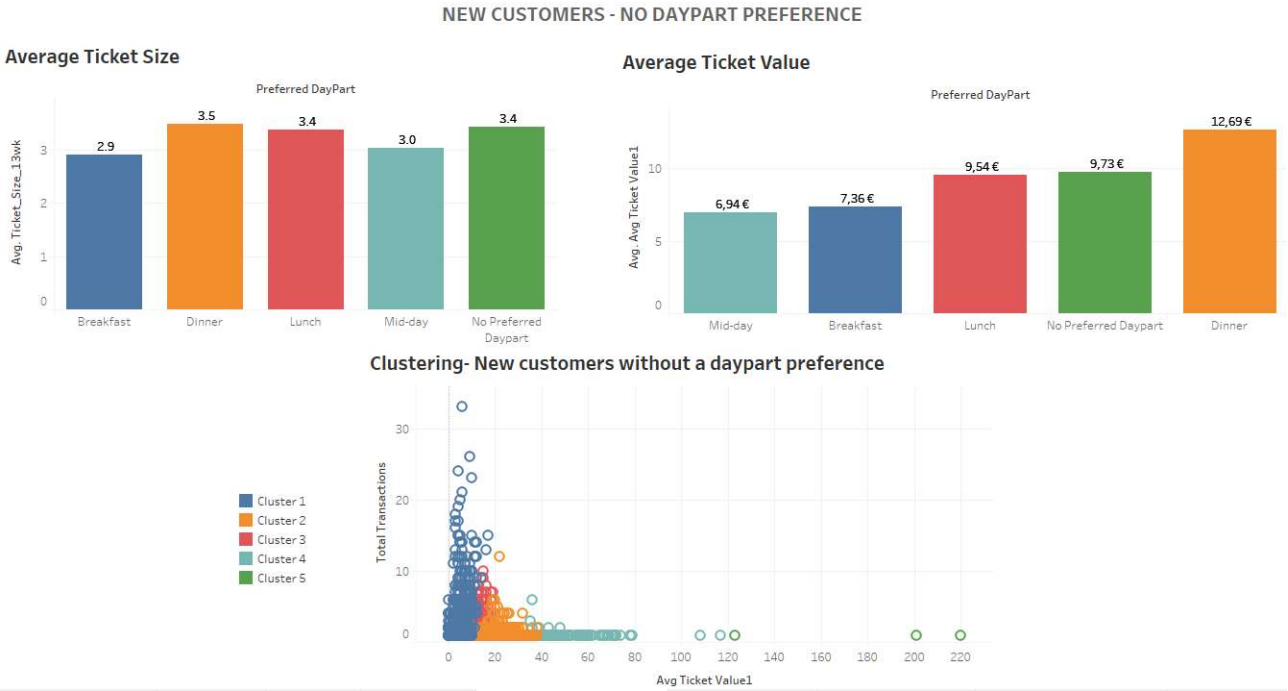


Figure 28 New customers - No daypart preference

Summary Diagnostics

Number of Clusters:	5
Number of Points:	8592
Between-group Sum of Squares:	23.739
Within-group Sum of Squares:	21.754
Total Sum of Squares:	45.493

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_13wk
Cluster 1	4030	3.1065	1.5347	1.816
Cluster 2	3165	11.153	1.297	3.9293
Cluster 3	1186	22.086	1.1796	6.1219
Cluster 4	208	43.572	1.0625	10.818
Cluster 5	3	181.33	1.0	46.667
Not Clustered	0			

Figure 29 Summary statistics - Clustering New customers without a daypart preference

Cluster analysis on customers without a preference for a day of the week showed better results, as seen in Figure 30. Clusters 3,5 and 6 clearly had a strong preference for weekends based on the high ticket value and ticket size (Figure 31). Clusters 1 and 4 can be concluded to prefer weekdays (Monday-Thursday) due to the low ticket size and ticket values. Cluster 2 does not have very conclusive values, but it is safe to say that this group prefers visiting mostly later in the week (Friday-Sunday) based on the relatively high values for the clustering variables. Overall, it was possible to draw inferences for more than 70% of this group. However, it must be

admitted that clustering New customers can always be ambiguous due to factors like a smaller number of data points, possibly one-off visits in response to an ongoing deal and so on.

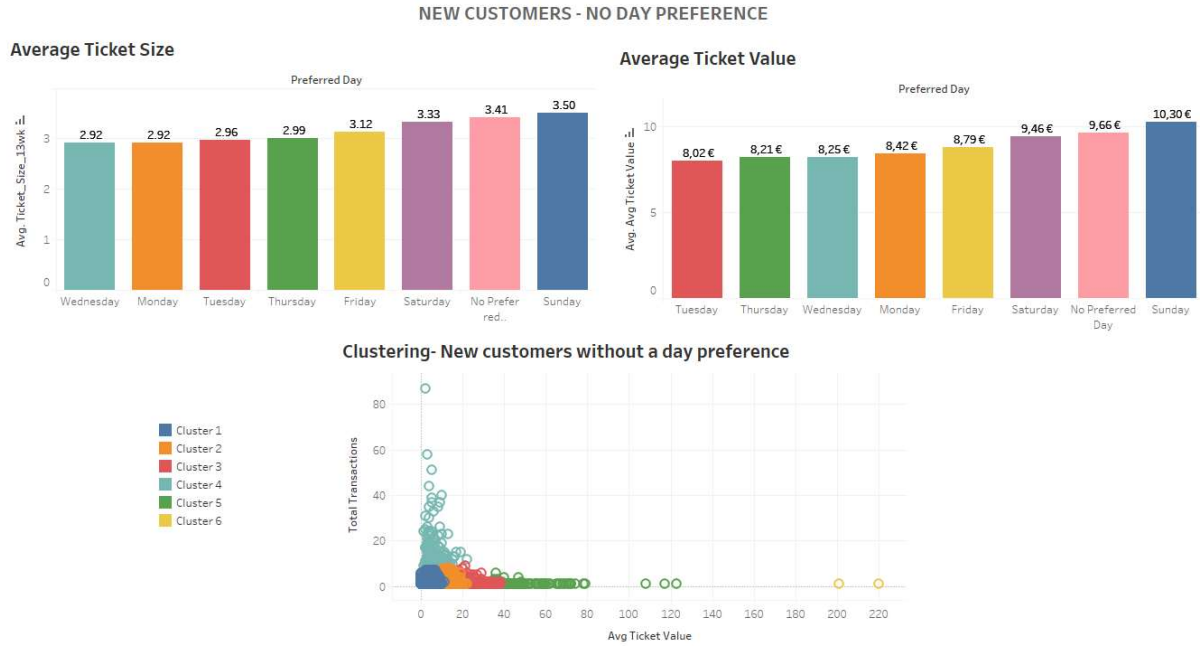


Figure 30 New customers - No day preference

Summary Diagnostics

Number of Clusters:	6
Number of Points:	9263
Between-group Sum of Squares:	29.178
Within-group Sum of Squares:	9.68
Total Sum of Squares:	38.858

		Centers		
Clusters	Number of Items	Sum of Avg Ticket Value	Sum of Total Transactions	Sum of Ticket_Size_13wk
Cluster 1	4016	2.9315	1.3394	1.7759
Cluster 2	3495	10.831	1.5831	3.855
Cluster 3	1309	21.811	1.2628	6.0443
Cluster 4	226	6.4115	14.035	2.7393
Cluster 5	215	43.679	1.1163	10.906
Cluster 6	2	210.5	1.0	54.5
Not Clustered	0			

Figure 31 Summary statistics - Clustering New customers without a day preference

Chapter 6. EVALUATION

Evaluation is an integral part of the DSRM and is necessary to receive feedback about the developed artifact. This section discusses the setup used for evaluation, the questions and the results.

6.1 Setup

Ideally, the evaluation should have been filled by employees of the marketing team at La Place. However, due to organizational changes and lack of awareness among the employees about the method being developed, it was not possible to ask the marketing team to fill in an evaluation. Hence, due to lack of time and availability of resources to evaluate the designed method, a short and simple questionnaire was developed. Since this research was conducted mainly for the benefit of managers and analysts at La Place, it was thought of using an evaluation model focused on 'users' at its core. Therefore, the unified theory of acceptance and use of technology (UTAUT) and some general questions were used as a reference model to build a questionnaire [32]. This questionnaire focused on several aspects like ease of using the artifact, effort to learn to use it and perceived usefulness of the artifact. The details of the questionnaire can be found in APPENDIX F.

6.2 Respondents

There were two respondents to the questionnaire. One of them is the Director of Business Analytics at La Place and supervised this research. He has deep knowledge of the processes at a Place and was an ideal candidate to give his opinions about the developed artifact. Moreover, he is update with recent trends and technologies in data science and machine learning, thereby being able to provide additional inputs for the improvement of the artifact. The other candidate chosen for evaluating the artifact was the supervisor from E-tail Genius. He has been associated with La Place through previous projects and values La Place as an esteemed client. His knowledge over the years, especially in the business applications of data science and machine learning techniques, proved to be valuable to overcome the lesser number of respondents to the questionnaire. He was able to provide a neutral perspective as he was not directly involved in the development of the artifact.

6.3 Questions setup for evaluation

Based on the UTAUT model, the evaluation questions were setup. The questions were sent to the respondents in the form of a Google form, which can be found in APPENDIX E. To make it easier for interviewees to evaluate and to assess in short amount of time, only 8 important questions were included covering the most important aspects for adoption of a new technology in an organization. The questions were divided mainly into two sets, open and closed questions. Moreover, it is well known that inclusion of both these types of questions in the evaluation process helps to learn unexcepted and important things [33].

- *Open Questions* these type of questions were setup to know the usability of the developed artifact and the overall impression about this research. The questions were framed of this type because it allows respondents to give free form of answers without limited choices.

- *Closed Questions* these types of questions were setup to understand the usefulness of functionalities of the dashboard which have a limited set of possible answers. The 5-point Likert Scale[10] is used to measure the results of these questions.

6.4 Evaluation Results

This section summarizes the evaluation results of the developed artifact based on the feedback received from the respondents. The results are divided into two sub-sections, namely; qualitative and quantitative results. The qualitative results are from open-end questions and the quantitative results are from closed-end questions.

6.4.1 Qualitative results

The qualitative results help understanding what went well and what did not from the respondents' point of view. Only one of the two respondents answered the open-ended questions thereby resulting in the sample size being just one for the qualitative results. The following questions were used for the qualitative results:

[10] https://en.wikipedia.org/wiki/Likert_scale

1. *What were the good things you liked in this classification scheme?*

One of the respondents mentioned that this classification combines a scheme which is easily understood together with ML techniques. This could be useful for managers and analysts to draw insights from the data which were probably previously unknown to them.

2. *What improvements would you like to see in this classification scheme?*

The same respondent suggested that this classification scheme should be extended to include other preferences of customers like product preferences. He mentioned that this scheme would definitely help La Place in understanding their customers better.

He also mentioned that it is worth investing time in such developments exploring the potential of using ML and data mining techniques.

6.4.2 Quantitative results

The quantitative results are more measurable and depend largely on the scales for measuring them. Both the respondents answered the closed-ended questions. The following questions were used for the qualitative results:

1. *On a scale of 1 to 5 (with 1 being the least and 5 being the highest) please rate the usefulness (utility) of this classification scheme in your opinion.*

One of the respondents thought this was mid-way on this scale of usefulness. The other respondent rated this the highest for usefulness. This is mainly because the knowledge of internal business processes at La Place. One of the respondents who was well acquainted with how things work at La Pace probably felt there was more scope to improve the usefulness of this work. For the other respondent, this was probably completely new, and he was fascinated by it.

2. *How much effort (effort to learn) did you have to take in order to understand this classification scheme?*

One of the respondents mentioned that he had to put some effort to understand the classification scheme. The other respondent mentioned that he needed very little time to learn. It was intended to keep this study simple and easy to understand without asking too much from someone trying to understand it. Based on the responses, it can be claimed that this fairly achieved.

3. *Do you think such a classification scheme will help La Place in better understanding its customers?*

Both the respondents were sure that this classification scheme would help in better understanding customers. This research provides a good basis and a starting point for La Place towards better customer relationship management. Also, it is assumed that there is no prior CRM tools or similar practices in place at La Place.

4. *Do you think with such a classification scheme, customers will benefit by being informed about the appropriate deals?*

One of the respondents was unsure if this scheme would ensure that customers would be informed timely about appropriately deals. The other respondent thought this scheme would be able to timely inform the right group of customers about upcoming deals. This is mainly because unless this scheme is implemented within the company and the results are checked for its effectiveness, there will always be uncertainty about the usefulness about such a study.

5. *Do you think it is worth investing time and effort to use Machine learning and data mining techniques to better understand customer behavior?*

Both respondents agreed that it is worth investing time and effort to use Machine learning and data mining techniques to better understand customer behavior. This was a common opinion while conducting this research at La Place. Considering that the company will soon undergo a takeover by a larger company with more IT expertise, this opinion might change in the future.

6. *Do you think it is a nice idea to continue the development of such a classification scheme in the future?*

Both respondents agreed and were convinced that it is indeed a nice idea to continue the development of such a classification scheme in the future . This was based on the current knowledge and expertise of the respondents. However, it is highly likely that once the business takeover takes place, there might be even better ideas and techniques which could supersede the proposed classification scheme.

Overall, mixed results were obtained from the evaluation. It would have been better if more people would have undertaken this evaluation but due to the lack of time and other business constraints, only two respondents were available. More respondents would have resulted in a more diverse set of results and increased the opportunities for future improvements.

Chapter 7. CONCLUSION AND DISCUSSION

The main objective of this study was to design a method to group customers together into small and distinct groups based on their transactional history. These groups could then be approached with different promotions and deals based on their characteristics. In order to do so, this study defines the design rules as well as presents simple and well-known techniques that can be used to classify and cluster customers. This chapter summarizes the overall findings from this study by answering the earlier defined research questions. Besides that, this section supports the reasoning behind using this 'hybrid' approach and suggestions for future work.

7.1 Conclusions

RSQ1: Which are the most widely used customer classification and clustering techniques?

Based on the literature review conducted as a background for this research, it was found that the LRFM model is the most commonly used classification method. It is easy to understand, well-known and a proven model. It is an improvement over the RFM model and considers the Lifetime (L) of customers into account. Classification based on LRFM results in different groups which can be easily sorted based on their importance and relevance to answer a specific question. The main reason for choosing this model as an inspiration for this study was because of its ability to classify customers based on each of the four attributes independently as well as combined. Some other popular classification techniques include association, regression, neural networks and decision tress.

Clustering is also a popular data mining technique in customer segmentation and used in over 50% of the available literature [1]. K-means clustering algorithm was found the most popular clustering algorithm. The working of this algorithm is easy to understand, and its parameters can be easily altered depending upon requirements. The in-built clustering feature with Tableau made it the perfect choice for this study. Also, the clustering results were easy to interpret and contributed to the overall goal of identifying small, distinct groups. Some other popularly used clustering techniques are the k-means ++ algorithm and c-means algorithm.

RSQ2: How can classification and clustering techniques be combined to group customers?

Using classification and clustering independently has its own benefits. Many researchers used them independently depending upon the business needs and the context [3] [17] [29]. Some researchers found the combination of the two better for their needs [18] [4].

This research proposes a generic method in Chapter 3 using which both these techniques can be combined. Each technique can be used for its own benefits and can be sufficient in its

own way. But with the help of the case study in Chapter 4, it has been shown that combining the two techniques is possible and clustering is normally used after classification to get relatively smaller groups from which inferences can be drawn. This does not rule out using classification after clustering.

RSQ3: What is the business value of incorporating results of these techniques in marketing strategies?

Many researchers have indicated that it is possible to combine classification and clustering [18] [27]. However, not many of them mentioned what were the quantifiable benefits of combining these techniques. This could be for many reasons, primarily such as not revealing to competitors how they do it and what financial benefits can be gained from it. Very few researchers pointed out what could be the value of combining these techniques [9]. This research indicated how different plans and measures can be taken to reach out to different groups of customers with the help of a customer analysis for marketing purposes.

In case of implementing the proposed method at La Place, it is evident from the Chapter 5, that the 'Frequent' visitors contribute the highest to the total revenue of La Place. Therefore, efforts should be made to increase these visitors and incentives should be given to them to ensure that they keep visiting frequently. It can be helpful if the promotions are designed to attract more customers to visit regularly. For this to happen, new marketing strategies need to be thought of.

Currently, there is no specific strategy used at La Place to target specific groups of customers for specific promotions and deals. The deals are targeted at the entire customer base. Sometimes there are very attractive deals like the one that was in place earlier this year around the month of April when a coffee and an apple pie was on promotion for just a Euro. The normal price is around Euro 5.50. This clearly caused a lot of customers to visit just for the deal and they probably never returned. It is impossible to track these customers as they did not sign-up for the loyalty programme. This ends up not being a very profitable deal for La Place. So, there should be an attempt to design an approach of minimizing such one-time visits and maximizing overall visits for all transactions and not just promotional deals.

Moreover, the knowledge of other customer groups such as the 'New' and the 'Redormed' can be crucial in turning them into 'Frequent' visitors. The 'Monitor' group needs to be addressed differently as they are a high potential group of turning into 'Frequent' visitors. A completely different campaign can be run to try and make the 'Lost' customers to return. This is a business decision that depends on various factors like the cost of reaching these customers, size of the group, perceived response rate and other detailed analysis like availability of nearby outlets for these customers, their previous experience with La Place and so on.

It cannot be said just yet if this study will help bring down the costs of marketing and promotional activities, but what it can help achieve is a direction in reaching out to specific groups of customers depending on what is being offered to them and what is already know about them.

RSQ4: How can we predict future visitors based on the results of these techniques?

Marketing and forecasting were largely unrelated at La Place at the time of conducting this research. However, the proposed classification scheme can cater to both. Its application to marketing has already been discussed. For forecasting and predicting future visitors, the random forest classifier method was used as discussed in Chapter 3. The code was implemented in Python and can be found in APPENDIX D.

Random forest was implemented to predict how many of the 94k Frequent visitors are classified as 'Frequent' based on our classification scheme. The model was trained using 70% of the data and tested using 30% of the data. The features used for predicting were related to the last 13 week transactional behavior and these features included the number of transactions in the last 13 weeks, the amount spent, the number of products consumed, the age of the customer and the lifetime. The model was able to predict with an accuracy of 98.03% that Frequent visitors would return. A sample output of the random forest classifier can be seen in Figure 32.

	truth	label	output_std
453711	1	1	1.00000
620874	1	1	0.97000
756986	1	1	0.99000
454709	1	1	1.00000
475236	1	1	0.96000
635405	1	1	0.88000
716347	1	1	1.00000
723562	1	1	0.97000
524634	1	1	1.00000
677061	1	1	0.71206
768195	1	1	0.88000
478570	1	1	1.00000
382974	1	1	0.98000
462825	1	1	0.79000
405431	1	1	0.94000
783569	1	1	0.98000
783451	1	1	0.88000
359150	1	1	0.98000
362382	1	0	0.36000
638342	1	1	1.00000
364414	1	1	0.95000
793779	1	1	0.92000

Figure 32 Sample output from Random forest classifier

This is a rather simple implementation of random forest for future prediction of visitors based on most recent transactional behavior. This can be extended based on past transactional behavior and inclusion of other attributes like knowledge about upcoming promotions and weather data, to get more refined and complete results. The more the data with which this model is trained, the better it will perform in order to make future predictions.

How can classification and clustering techniques be used in effective target group-based marketing schemes?

This research indicates how it is possible to understand customer preferences based on their transactional history. The knowledge gained from such a study can help in exploring untapped insights about customers. Depending upon the nature of data available within an organization, the proposed artifact can be modified without much effort in order to answer specific business questions.

Many organizations are yet to explore the potential of using machines learning techniques like classification and clustering. A good starting point for them would be to start with an approach like the one proposed here. An understanding of most important customers based on their past behavior could be crucial in retaining them. This would reduce the cost and effort of trying out different things which possibly do not yield positive results.

Marketing and promotional activities are generally expensive and time consuming. What does not help is that often marketing teams do not know their target audience well and end up building something too generic without focusing on special customer groups. Using customer classification and clustering techniques, it can be possible to find the existence of special customer groups, understand their characteristics and find what makes them unique. Once there exists knowledge about such groups, targeted approaches can be made to reach out to each unique group in a different way. This is the main purpose of CRM systems, trying to build personal connections with customers.

Another vital asset offered by this research is the prediction of future behavior of customers. More advanced mathematical and statistical techniques/models can be combined with such a study. This can be used to used not only for marketing but also for demand forecasting and planning. So, the benefits of such a research are multifold.

7.2 Discussion

The most important findings of this research and the contributions are discussed in this section.

7.2.1 Classification vs Clustering OR Both combined?

An important argument to be considered while conducing this research was whether classification or clustering could alone help achieve the set objective. Classification alone can help classify new customers based on predefined labels. This needs a base model to be trained and

tested. Only after enough efforts, will this model be able to accurately classify new customers. If a model like LRFM is used, it can lead to many groups, which could easily become unmanageable for the company. However, labels or classes are very important as they help focus on specific groups, knowing their importance. On the other hand, clustering by itself is not very capable of explaining much. The results strongly depend upon the chosen algorithm, the input variables and the number of clusters. It requires some prior knowledge about the input variables being used. This is very helpful in interpreting the results. It can be said that in this context, classification or clustering alone would not have been sufficient and therefore, a combination of the two techniques was used. Although, a claim about which is better of the two techniques could only be made after comparing the results of both separately.

The benefits of combining both classification and clustering is that it combines the advantages of both and overcomes the limitations of just using on technique independently. Moreover, the combination helps in getting more detailed information. For example, the classification scheme used here helped find 7 groups based on the recency of transactions, as seen in (Figure 8, Figure 10, Figure 17, Figure 22 and Figure 27). This helped in further clustering selected groups (Figure 11, Figure 14, Figure 18, Figure 20, Figure 23,Figure 25, Figure 28 and Figure 30). If only one of the two techniques were used, it would probably be too difficult to interpret the details of the results and the mechanism behind those techniques. On the other hand, clustering algorithms group customers automatically based on the inputs, whereas classification relies on human interference to tell the computer what it should do and then empowers the computer to do it automatically. The perfect combination is a mix of both, classification with human knowledge helping focus on specific groups, and clustering with the power of algorithms to find patterns within groups which were unknown based on classification. Combining the two techniques helps in getting a more detailed level of information than just using one technique by itself.

7.2.2 What's in it for the Marketing team?

One of the most important outcomes from this classification scheme is the set of customers to be focused upon based on their recent transactional behavior. Different approaches are needed to target different groups. The eventual goal should be to maintain and increase frequent visitors. Also, new customers should be regularly added, and attempts should be made to revive the ones who were lost/inactive.

The most important group is clearly the 'Frequent' visitors. As seen in Appendix C, almost 40% of the 'Frequent' visitors have a likelihood of falling out as they visited only once in the last 13 weeks, whereas the average visits of this group were 3.8 times. This is an indicator that special

attention needs to be paid to these visitors. A deal or a promotion should be designed to appeal them to visit again in the next 13 weeks. It is also seen that more than 60% of the 'Frequent' visitors have an 'OLD' Lifetime. There is clearly evidence of a higher loyalty by customers aged in the 'Baby boomer' category. This is probably the most loyal group of La Place and schemes should be regularly made to respect and honor their loyalty. There is nothing more influencing than the word-of-mouth and trust of customers.

The 'Monitor' group is probably the most sensitive group to tap as it has great potential to turn into 'Frequent' visitors. This group has a relatively lower average age and the biggest group of customers belong to the 'Millenial' age group. This group is highly likely to use social media and technology in their daily lives and this communication channel should be explored to reach them. Moreover, results from surveys and feedbacks received from these customers could be a great indicator as to why they chose not to return. An appealing deal with the caption *'We missed you'* could go a long way in tempting them to return. After all, customers like to be valued and a personal touch would be appreciated.

The 'Redormed' group has the highest likelihood to fallout as they were possibly one-off visitors on a vacation or a quick brunch/meal with family. But the fact that they still used the membership card to record their transaction indicates that they haven't forgotten La Place. They are generally high spenders indicating they do not visit alone. In order to turn them into frequent visitors, they should be targeted probably with family deals or deals which encourage them to visit with their loved ones. A deal with the caption *'Nice to see you back'* would possibly be appreciated.

It is almost impossible to predict which of the customers from the 'Inactive' and 'Lost' group would return. What does not help even further, is that customers are no longer encouraged to use their loyalty card. So, even if they do visit and do not use their cards, their transactional history is not recorded. It is important to first understand the reason behind such a big 'Lost' group. One such major reason is the shutting down of many stores post 2015. It is true that slowly more stores will be opened at new locations, but it remains a business call if they would like to target such a big group of 'Lost' customers. This can be expected to be an expensive project, but whether it will yield high return on investment depends on several factors based on the history of La Place and the research that goes into addressing this issue.

7.2.3 Contributions

This research makes overall contributions to science and specifically to business applications of machine learning techniques. Firstly, it provides extensive knowledge about various classification and clustering techniques. Many companies use these techniques individually or combined.

However, based on the literature review conducted prior to carrying out this research, a research gap was observed in terms of using machine learning techniques to help marketing teams design promotions and deals. None of the companies revealed how they designed promotions catering to the needs of specific groups. Everything was more theoretical and there was lack of a solid and concrete framework/method for companies to follow, willing to explore opportunities of using machine learning techniques like classification and clustering in marketing.

The main purpose behind this study was to propose a new method that would help La Place in creating new promotional strategies based on the knowledge about specific groups. While formulating the design framework and the design method, the emphasis was on designing something generic. Yes, it should help solve the business problem at La Place, but it also contributes to science and fill in the identified research gap. Classification and clustering helped in identifying specific customer groups. This proposed method is robust and can be adapted by other companies without much effort. This research opens horizons for companies at different scales to be able to understand their customers better. Many CRM tools readily available in the market also help achieve similar goals, but it requires expertise to handle such systems. Moreover, this method was designed to keep it simple and the results are explainable to people with all knowledge levels, mainly intended for marketing teams, analysts and managers.

Additionally, using techniques like random forests or regression, companies can build models based on past transactional behavior of customers. These models can be trained and used for classification and regression. Such models can then be used for classification and forecasting.

7.2.4 Limitations

This research does have some limitations. A methodological paper was used to evaluate the limitations of this study [11]. The first limitation of this research is that it depends on customers' transactional history. Transactional history is used at a macro-level. Specific details of transactions such as products consumed or combinations of products, were not drilled into. Including these details would lead to useful insights about assortment preferences of customers. Second limitation is that customers do not use their loyalty/membership cards while in the store and this means that for the same customers, there is only partial data recorded. So, this could mean that a bias is injected while calculating a customer's reference. Third limitation is that this research did not consider the impact of promotions on actual sales. We think that promotions are likely to significantly affect customers preferences and spending. Finally, this study does not compare different classification and clustering algorithms for their performance and effectiveness. This is a time consuming task but could be useful in revealing if the right approach

is being used for a given context. Also, more involvement from other teams within the organization would have benefited the development of the proposed method.

7.2.5 Validity and Reliability

A critical validity concern in design science research such as ours is about generalizability: to what extent the method we propose in the context of La Place could possibly work and be useful in other organizational contexts [11]. In order to prove that the method proposed in this research would be effective, it needs to be put into practice. It must be mentioned that the method proposed is here is ideally suited for companies interested in learning from transactional history of customers. This does not remain limited to HORECA companies but can easily be adopted within supermarkets, retail chains and similar companies/industries which build a brand loyalty with their customers. Although, this research lacks validation by evaluating its implementation in a real-world scenario, it is safe to say that the proposed method has been designed with utmost care in order to yield fruitful results for companies which did not focus so much on customer loyalty and targeted marketing.

When the results of this research were analyzed, they were found to be reliable enough based on the supervisor's business knowledge at La Place and can be seen in Chapter 5 . This is mainly because of the data pre-processing that was done before implementing the techniques. The correct set of customers were identified into different classes and these were verified by the prior business knowledge of the supervisor at La Place. Also, the model was motivated from the LRFM model, which remains one of the most widely used and trusted classification models. K-means clustering algorithm has always been reliable in terms of performance and interpretation of results. It is easy to understand the technical details of these algorithms which helps in making them more reliable.

Overall, there is enough credibility in terms of the validity and reliability of this research, but this could only be proven by time once it is put into practice and improvements could be made based on received feedback.

7.3 Recommendations for future work

No research is complete and has the potential of being extended to enhance it. Upon analyzing the results of this study and gathering feedback through evaluation from experts, there are some

suggestions and recommendations to improve the proposed method and extend it. Some of these key recommendations for future work are listed below:

- The inclusion of product level preferences for each customer would be a very interesting attribute in the classification scheme. This would make the outcome of this method very big and complex but can provide a different angle to find new groups of customers.
- Inclusion of other attributes like the preferred type of store and the location of store (highway/city) could also be a good indicator of the existence of previously unknown groups.
- Use of already known customer data, like their feedback about their previous visits through surveys, Net Satisfaction rate (NSAT) and perception about La Place, can be explored to gain insights about their experience with La Place.
- A sentiment analysis (Big Data) of social media data about the opinion of people about La Place and in specific to promotions and deals [34] [12]. This can be combined with actual in-store data for a defined intervention period. For example, a popular ongoing promotion like the coffee deal earlier this year in April which had a tremendous response in terms of in-store visit and social media activity.
- The knowledge of upcoming promotions in advance and the response to past promotions must be done to understand if customers behave as they were perceived to. This can be an indicator for an improvement to the proposed method.
- The analysis of in-store traffic during an intervention period can be an important input as different types of customers visit during an ongoing promotion, not just loyalty members [35].

In general, if this method is implemented at La Place, an intervention period should be defined to check the extent to which results of the classification scheme holds true. This can then be used to improve the method in combination with the other recommendations listed above.

7.4 Advice to La Place

Upon designing the proposed method and implementing it on the historical data, important findings were revealed which were previously unknown at La Place. The effectiveness of this method is mainly contributed by the combination of the approach used, which combines the classification techniques and then followed by clustering. This combined approach looks like the way to go rather than using each technique independently. Moreover, this scheme is relatively easy to understand and replicable. Considering that soon La Place will undergo major business

transformation, this study can be used as a building block for future work. The tools and the IT landscape may change, but enough care was taken to design a method which can be replicated without much re-work and effort.

References

[1] D. A. Kandeil, A. A. Saad and S. M. Youssef, "A two-phase clustering analysis for B2B customer segmentation," in *In 2014 International Conference on Intelligent Networking and Collaborative Systems*, IEEE, 2014, pp. 221--228.

[2] E. W. Ngai, L. Xiu and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert systems with applications,* vol. 36, no. 2, pp. 2592--2602, 2009.

[3] G. Zhang, Yun Chen, H. Dengfeng and F. Chuan, "Customer segmentation based on survival character," *In 2007 International Conference on Wireless Communications, Networking and Mobile Computing,* pp. 3391--3396, 2007.

[4] M. Lewis, "The influence of loyalty programs and short-term promotions on customer retention," *Journal of marketing research,* vol. 41, no. 3, pp. 281--292, 2004.

[5] T. K. Das, "A customer classification prediction model based on machine learning techniques," in *In 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2015.

[6] D. Rajagopal, "Customer data clustering using data mining technique," *arXiv preprint arXiv:1112.2663,* 2011.

[7] A. Ansari and A. Riasi, "Customer clustering using a combination of fuzzy c-means and genetic algorithms," *International Journal of Business and Management,* vol. 11, no. 7, p. 59, 2016.

[8] K. K. Tsiptsis and A. Chorianopoulos, Data mining techniques in CRM: inside customer segmentation, John Wiley \& Sons, 2011.

[9] S.-Y. Kim, T.-S. Jung, E.-H. Suh and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert systems with applications,* vol. 31, no. 1, pp. 101--107, 2006.

[10] M. Tarokh and K. Sharifian, "Applications of data mining in improving customer communication management," *Iranian Industrial Management Studies Quarterly,* vol. 6, no. 17, pp. 153-181, 2010.

[11] R. Wieringa and M. Daneva, "Six strategies for generalizing software engineering theories," *Science of computer programming,* vol. 101, pp. 136--152, 2015.

[12] A. I. Canhoto, M. Clark and P. Fennemore, "Emerging segmentation practices in the age of the social customer," *Journal of Strategic Marketing,* vol. 21, no. 5, pp. 413--428, 2013.

[13] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Systems with Applications,* vol. 40, no. 16, pp. 6491--6507, 2013.

[14] R. K. Yin, *Doing case study research,* Thousand Oaks, CA: Sage, 2009.

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems,* vol. 24, no. 3, pp. 45-77, 2007.

[16] J. F. Wolfswinkel, E. Furtmueller and C. P. Wilderom, "Using grounded theory as a method for rigorously reviewing literature," *European journal of information systems,* vol. 22, no. 1, pp. 45--55, 2013.

[17] G. Lefait and T. Kechadi, "Customer segmentation architecture based on clustering techniques," in *2010 Fourth International Conference on Digital Society*, IEEE, 2010, pp. 243--248.

[18] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Systems with Applications,* vol. 36, no. 3, pp. 4176--4184, 2009.

[19] Y. H. Tao and C. C. R. Yeh, "Simple database marketing tools in customer analysis and retention," *International Journal of Information Management,* vol. 23, no. 4, pp. 291--301, 2003.

[20] H. H. Wu, E. C. Chang and C. F. Lo, "Applying RFM model and K-means method in customer value analysis of an outfitter," in *Global Perspective for Competitive Enterprise, Economy and Ecology*, S. Smith and A. Trappey, Eds., London, Springer London, 2009, pp. 665--672.

[21] J. T. Wei, S. Y. Lin, C. C. Weng and H. H. Wu, "A case study of applying LRFM model in market segmentation of a children's dental clinic," *Expert Systems with Applications,* vol. 39, no. 5, pp. 5529--5533, 2012.

[22] W. J. Reinartz and V. Kumar, "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing," *Journal of marketing,* vol. 64, no. 4, pp. 17--35, 2005.

[23] C. Calvo-Porral and J.-P. Lévy-Mangin, "From "foodies" to "cherry-pickers": A clustered-based segmentation of specialty food retail customers," *Journal of Retailing and Consumer Services,* vol. 43, pp. 278--284, 2018.

[24] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," *Procedia Computer Science,* vol. 3, pp. 1327--1332, 2011.

[25] N. R. Barraza, S. Moro, M. Ferreyra and A. de la Peña, "Information theory based feature selection for customer classification," in *In Simposio Argentino de Inteligencia Artificial (ASAI 2016)-JAIIO 45 (Tres de Febrero, 2016).*, 2016, November.

[26] K. Coussement, F. A. Van den Bossche and K. W. De Bock, "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees," *Journal of Business Research,* vol. 67, no. 1, pp. 2751--2758, 2014.

[27] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," *Journal of business research,* vol. 60, no. 6, pp. 656--662, 2007.

[28] J. Xiao, X. Jiang, C. He and G. Teng, "Churn prediction in customer relationship management via GMDH-based multiple classifiers ensemble," *IEEE Intelligent Systems,* vol. 31, no. 2, pp. 37--44, 2016.

[29] E. Apeh, B. Gabrys and A. Schierz, "Customer profile classification: To adapt classifiers or to relabel customer profiles?," *Neurocomputing,* vol. 132, pp. 3--13, 2014.

[30] A. M. Hughes, *Strategic Database Marketing,* Chicago: Probus Publishing Company, 1994.

[31] H.-H. Wu, S.-Y. Lin and C.-W. Liu, "Analyzing patients' values by applying cluster analysis and LRFM model in a pediatric dental clinic in Taiwan," *The Scientific World Journal,* 2014.

[32] V. Venkatesh, M. G. Morris, G. B. Davis and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly,* pp. 425--478, 2003.

[33] S. Farrell, "Open-ended vs. closed-ended questions in user research," *Evidence-Based User Experience Research Training, and Consulting,* 2016.

[34] A. Blom, F. Lange and R. L. Hess Jr, "Omnichannel-based promotions' effects on purchase behavior and brand image," *Journal of Retailing and Consumer Services,* vol. 39, pp. 286--295, 2017.

[35] L. D. Epstein, A. A. Flores, R. C. Goodstein and S. J. Milberg, "A new approach to measuring retail promotion effectiveness: A case of store traffic," *Journal of Business Research,* vol. 69, no. 10, pp. 4394--4402, 2016.

APPENDIX

Appendix A. Details about K-means clustering algorithm

"Not Clustered" category

When there are null values for a measure, Tableau assigns values for rows with null to a *Not Clustered* category. Categorical variables (that is, dimensions) that return * for ATTR (meaning that all values are not identical) are also not clustered.

Scaling

Tableau scales values automatically so that columns having a larger range of magnitudes don't dominate the results. For example, an analyst could be using inflation and GDP as input variables for clustering, but because GDP values are in trillions of dollars, this could cause the inflation values to be almost completely disregarded in the computation. Tableau uses a scaling method called *min-max normalization*, in which the values of each variable is mapped to a value between 0 and 1 by subtracting its minimum and dividing by its range.

Clustering constraints

Clustering is available in Tableau Desktop but is not available for authoring on the web (Tableau Server, Tableau Online). Clustering is also not available when any of the following conditions apply:

- When a cube (multidimensional) data source is being used.
- When there is a blended dimension in the view.
- When there are no fields that can be used as variables (inputs) for clustering in the view.
- When there are no dimensions present in an aggregated view.

When any of those conditions apply, it is not possible to use the Clustering feature in Tableau Desktop.

In addition, the following field types cannot be used as variables (inputs) for clustering:

- Table calculations
- Blended calculations
- Ad-hoc calculations
- Generated latitude/longitude values
- Groups
- Sets
- Bins
- Parameters
- Dates
- Measure Names/Measure Values

Summary Diagnostics and Cluster statistics

1. **Number of Clusters**

The number of individual clusters in the clustering.

2. **Number of Points**

The number of marks in the view.

3. **Between-group sum of squares**

A metric quantifying the separation between clusters as a sum of squared distances between each cluster's center (average value), weighted by the number of data points assigned to the cluster, and the center of the data set. The larger the value, the better the separation between clusters.

4. **Within-group sum of squares**

A metric quantifying the cohesion of clusters as a sum of squared distances between the center of each cluster and the individual marks in the cluster. The smaller the value, the more cohesive the clusters.

5. **Total sum of squares**

Totals the between-group sum of squares and the within-group sum of squares. The ratio (between-group sum of squares)/(total sum of squares) gives the proportion of variance explained by the model. Values are between 0 and 1; larger values typically indicate a better model. However, you can increase this ratio just by increasing the number of clusters, so it could be misleading if you compare a five-cluster model with a three-cluster model using just this value.

6. **# Items**

The number of marks within the cluster.

7. **Centers**

The average value within each cluster (shown for numeric items).

8. **Most Common**

The most common value within each cluster (only shown for categorical items).

Cluster Descriptions

Analysis of variance (ANOVA) is a collection of statistical models and associated procedures useful for analyzing variation within and between observations that have been partitioned into groups or clusters. In this case, analysis of variance is computed per variable, and the resulting analysis of variance table can be used to determine which variables are most effective for distinguishing clusters.

Relevant analysis of variance statistics for clustering include:

i. ***F-statistic***

The F-statistic for one-way, or single-factor, ANOVA is the fraction of variance explained by a variable. It is the ratio of the between-group variance to the total variance. The larger the F-statistic, the better the corresponding variable is distinguishing between clusters.

ii. ***p-value***

The p-value is the probability that the F-distribution of all possible values of the F-statistic takes on a value greater than the actual F-statistic for a variable. If the p-value falls below a specified significance level, then the null hypothesis (that the individual elements of the variable are random samples from a single population) can be rejected. The degrees of freedom for this F-distribution are (k - 1, N - k), where k is the number of clusters and N is the number of items (rows) clustered. The lower the p-value, the more the expected values of the elements of the corresponding variable differ among clusters.

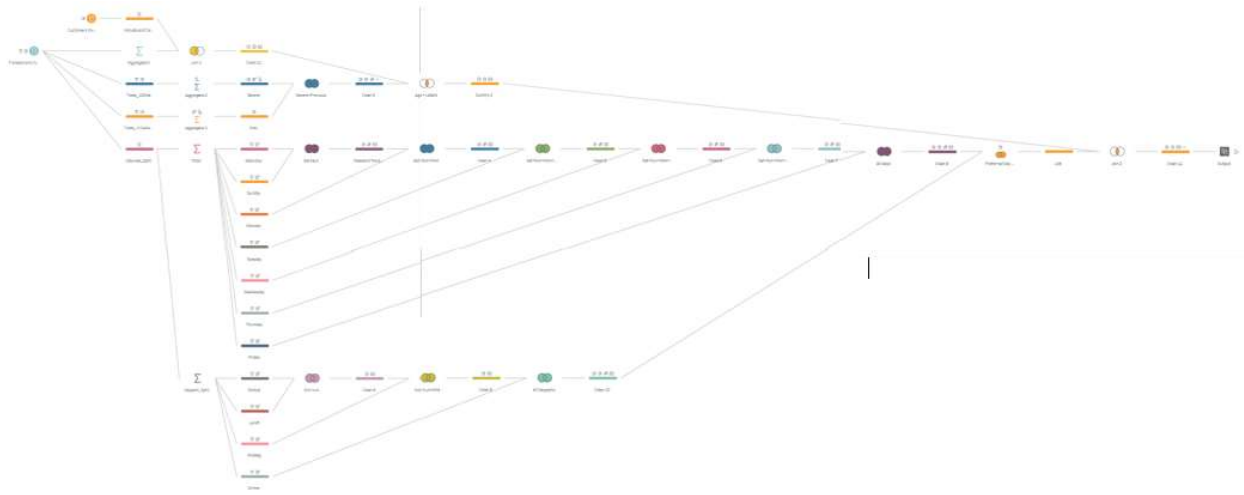
iii. ***Model Sum of Squares and Degrees of Freedom***

The Model Sum of Squares is the ratio of the between-group sum of squares to the model degrees of freedom. The between group sum of squares is a measure of the variation between cluster means. If the cluster means are close to each other (and therefore close to the overall mean), this value will be small. The model has k-1 degrees of freedom, where k is the number of clusters.

iv. ***Error Sum of Squares and Degrees of Freedom***

The Error Sum of Squares is the ratio of within-group sum of squares to the error degrees of freedom. The within-group sum-of-squares measures the variation between observations within each cluster. The error has N-k degrees of freedom, where N is the total number of observations (rows) clustered and k is the number of clusters. The Error Sum of Squares can be thought of as the overall Mean Square Error, assuming that each cluster center represents the "truth" for each cluster.

Appendix B. Tableau Prep Flow



Appendix C. Additional Tables

1. Lifetime + Recency

	Lifetime Labels		
Label	MIDDLE	NEW	OLD
Frequent	32.98%	5.67%	61.35%
Inactive	56.06%		43.94%
Lost	0.06%		99.94%
Monitor	43.76%	7.47%	48.76%
New		99.97%	0.03%
Redormed1	32.94%		67.06%
Redormed2	0.12%		99.88%

2. Customers with high likelihood to fall out based on last 13 weeks transactional history

Customers with Only 1 transaction:

Label	
Frequent	36,195
New	7,116
Redormed1	5,382
Redormed2	2,962
Grand Total	51,655

Total customers:

Label	
Frequent	94,811
New	11,199
Redormed1	7,666
Redormed2	4,222
Grand Total	117,898

Appendix D. Code for Random Forest

```
import pandas as pd

df = pd.read_excel (r' C:\Users\Amit Das\Desktop\Thesis - MBIT\Research Topics – MBIT\
Customer_Classification_Output.xlsx')
```

```
df.Label.unique()

df['Label'] = df['Label'].apply(lambda x: '1' if x == 'Frequent' else '0')

df.head(10)

df["Label"]=pd.to_numeric(df["Label"])

frequent = df[df["Label"]>0]

others = df[df["Label"]==0]

frequent.shape

others.shape

frequent_features = frequent.iloc[:,1:5]

frequent_targetvariable = frequent.iloc[:,0]

from sklearn.model_selection import train_test_split

X_train_freq, X_test_freq, y_train_freq, y_test_freq = train_test_split(frequent_featu
res, frequent_targetvariable, stratify=frequent_targetvariable, test_size=0.3)

others_features = others.iloc[:,1:5]

others_targetvariable = others.iloc[:,0]

X_train_oth, X_test_oth, y_train_oth, y_test_oth = train_test_split(others_features, o
thers_targetvariable, stratify=others_targetvariable, test_size=0.3)

X_train = pd.concat([X_train_freq,X_train_oth])

X_test = pd.concat([X_test_freq,X_test_oth])

y_train = pd.concat([y_train_freq,y_train_oth])

y_test = pd.concat([y_test_freq, y_test_oth])

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(

    n_estimators = 100,  # 1000 trees

)

rf.fit(X_train, y_train)

y_predicted = rf.predict(X_test)

results = pd.DataFrame({

    'truth'      : y_test,                          # True labels

    'label'      : y_predicted,                     # Labels shown to models

    'output_std' : rf.predict_proba(X_test)[:,1]   # Random forest's scores

}, columns = ['truth', 'label', 'output_std'])

results

from sklearn.metrics import accuracy_score

accuracy_score(y_test, y)
```

Appendix E. Questionnaire used for Evaluation

Customer classification scheme feedback (Master Thesis - Amit Das)

Thank you for taking your time in understanding how the prototype works! As a part of evaluating the prototype of my Master Thesis , I request you to fill in this questionnaire after understanding how the prototype works. This questionnaire would take approximately 5-10 minutes to fill in. Your responses will be kept anonymous and will be used merely for research purposes and future improvement of the prototype.

* Required

1. **What is your name?** *

2. **What is your function/role at La Place?** *

3. **On a scale of 1 to 5 (with 1 being the least and 5 being the highest) please rate the usefulness of this classification scheme in your opinion.** *
   Mark only one oval.

	1	2	3	4	5	
Not at all Useful	◯	◯	◯	◯	◯	Extremely Useful

4. **How much effort did you have to take in order to understand this classification scheme?** *
   Mark only one oval.
   - ◯ Not a lot
   - ◯ Some effort
   - ◯ A lot of effort

5. **Do you think such a classification scheme will help La Place in better understanding its customers?** *
   Mark only one oval.
   - ◯ Yes
   - ◯ No
   - ◯ Maybe

6. **Do you think with such a classification scheme, customers will benefit by being informed about the appropriate deals?**
   Mark only one oval.

   - Yes
   - No
   - Maybe

7. **Do you think it is worth investing time and effort to use Machine learning and data mining techniques to better understand customer behavior?** *
   Mark only one oval.

   - Yes
   - No
   - Maybe

8. **What were the good things you liked in this classification scheme?** *

9. **What improvements would you like to see in this classification scheme?** *

10. **Do you think it is a nice idea to continue the development of such a classification scheme in the future?**
    Mark only one oval.

    - Yes
    - No
    - Maybe