VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELAGAVI-590018, KARNATAKA

A PROJECT REPORT(1cr16is116) ON

"CUSTOMER SEGMENTATION USING CLUSTERING"

Submitted in partial fulfilment of the requirement

for the award of the degree of

Bachelor of Engineering

In

Information Science & Engineering

For the academic year 2019-20

Submitted by

USN	Name
1CR1S116	TARUN SAINI

Under the guidance of

Dr. Shahina Parveen

Associate PROFESSOR, DEPARTMENT OF ISE, CMRIT BANGALORE

Department of Information Science & Engineering

CMR Institute of Technology, Bengaluru – 560 037

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to Certify that the dissertation work **"CUSTOMER SEGMENTATION USING CLUSTERING"** carried out by **Mr. Tarun Saini USN: 1CR16is116,** respectively is bonafide students of **CMRIT** in partial fulfillment for the award of **Bachelor of Engineering** in **Information Science & Engineering** of the **Visvesvaraya Technological University, Belagavi,** during the academic year **2019-20**. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said degree.

Signature of Guide	Signature of HOD	Signature of Principal
Dr Shahina parveen	**Dr. Shreekanth Mooroor Prabhu**	**Dr. Sanjay Jain**
Associate Professor	**Head of the Department**	**Principal**
Dept. of ISE	**Dept. of ISE,**	**CMRIT,**
CMRIT, Bengaluru	**CMRIT, Bengaluru**	**Bengaluru**

DECLARATION

We, the students of Information Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled " **CUSTOMER SEGMENTATION USING CLUSTERING** " has been successfully completed under the guidance of Mrs Dr. Shahina Parveen., Associate Professor, Information Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfilment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2020 - 2021. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place: Bangalore

Date: 06-Jun-2020

Team members:

TARUN SAINI(1CR16IS116)

ABSTRACT

Customer segmentation and pattern extraction is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. This project proposes an approach for determining target customers using predictive model and also discover their associative buying patterns by displaying it using k means clustering algorithm . After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

The objective of this proposed system is to develop an application which will predict the set of people to be targeted based on their purchasing capacity. So, a company can decide the right set of people to show the product. This project will try to predict the set of customer that have a higher chances of buying product base on their behavior .

ACKNOWLEDGEMENT

We take this opportunity to express our sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing us a platform to pursue our studies and carry out our final year project.

We take great pleasure in expressing our deep sense of gratitude to **Dr. Sanjay Jain,**Principal, CMRIT, Bangalore for his constant encouragement.

We would like to thank **Dr. . Shreekanth Mooroor Prabhu ,** Professor and Head, Department of Computer Science &Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

We express our sincere gratitude and we are greatly indebted to **Dr. Shahina Parveen, Associate Professor,** Department of Computer Science & Engineering, CMRIT, Bangalore, for her invaluable co-operation and guidance at each point in the project without whom quick progression in our project was not possible.

We are also deeply thankful to our project guide **Dr Sudhakar K.N, Associate Professor,** Department of Information Science & Engineering, CMRIT, Bangalore, for critically evaluating each step in the development of this project and providing valuable guidance through our mistakes.

We also extend our thanks to all the faculty of Information Science & Engineering who directly or indirectly encouraged us.

Finally, we would like to thank our parents and friends for all their moral support they have given us during the completion of this work.

TABLE OF CONTENTS Page No

CHAPTER 1

INTRODUCTION

Sometimes referred to as market segmentation, customer segmentation is a method of analysing a client base and grouping customers into categories or segments which share particular attributes. Key differentials in segmentation include age, gender, education, location, spending patterns and socio-economic group. Relevant differentials are those which are expected to influence customer behaviour in relation to a business. The selected criteria are used to create customer segments with similar values, needs and wants.
When planning a targeted marketing campaign, it is also necessary to differentiate customers within these groupings according to their preferred means of communication.

Customer segmentation is an essential tool in customer relationship management, enabling businesses to market effectively to their customers. Businesses are expected to understand their customers and demonstrate their customer insights by sending only relevant, targeted communications to their customers. Customers want to feel valued and be treated as individuals, yet for anything other than perhaps the smallest of businesses this level of customer knowledge is impossible to achieve.

Segmentation also allows businesses to channel their resources appropriately. High value customers who purchase frequently and generate more revenue usually belong in a segment which is allocated a higher level of marketing spend.

Analysing customer demographics and psychographics gives layers of insights which help anticipate customers' needs and plan new products and services. This in turn enables marketeers to target more accurately those customers or prospects who would be most interested in them.

1.1 Create the foundation for a relevant Value Proposition...

With products matching customers needs;

With adjusting pricing and discount schemes; and

With service offerings attracting and keeping the relevant customers loyal.

1.2 Adjust the Go-to-Market Strategy (GTM) in order

Utilize the best use of a multi-channel approach;

Optimize sales efforts across the full Customer Experience journey and

Align marketing and messaging toward our target customers.

1.3 segmentation in right way

Segmentation. Everyone's doing it, but are you doing it right?

Although segmentation should help target the right customers, it does not always bring the intended value. Inappropriately or incorrectly defined segments may lead management to make bad strategic decisions.

Companies try to use just one segmentation model for all/different purposes.

Companies have multiple segmentation models without a clear and consistent linkage among them.

Companies base the model on data at hand vs. data needed to make the insights relevant.

Management is not aligned on what drives and what describes the market.

Companies allow complexity of a few key customers interfere with a simple market view.

The purpose of the segmentation is "lost in translation" when executed on an operational level.

1.1 Objectives

The objectives of the project are:

1) Gain a competitive advantage
2)Break into a new market
3)Improve product development
4)Boost return on investment
5)Optimize user experiences
6)Improve business focus

1.4 Scope

The premise of market segmentation is that to maximize sales to a large population of customers, it is best to divide it into logical subgroups. The assumption is that by dividing one large, amorphous mass into subgroups, you can fine-tune your product, messaging, support, or distribution channels to meet the specific needs of unique customer groups. Thus, the goal is to use a market segmentation model to improve marketing success and optimize marketing ROI.

Segmentation models vary from basic to complex, and the approaches to developing and applying them is a topic for an entire book itself. But here are some examples:

- **Example 1**: A telecommunications company selling mobile phone services might segment its market based on complexity of needs. One customer group might only need voice service, and very little volume at that. Another group might primarily use mobile texting services. Yet another might be a heavy user of mobile phones for

voice, text, email, and web browsing. By identifying distinct patterns in customer needs, the company can optimize product bundles and target them at the correct audiences.

- **Example 2**: A hotel chain that caters to families might simply segment its market based on income level and travel frequency. The chain might find that a group of moderate income-frequent travelers exists that is swayed by certain loyalty program rewards. It might find another important group exists—that is more swayed by on-site amenities. With this information, the hotel can optimize its offerings and loyalty programs to appeal to each group's unique needs.

CHAPTER 2

LITERATURE SURVEY

2.1 Market Segmentation, Targeting and Positioning

Market segmentation is the actual process of identifying segments of the market and the process of dividing a broad customer base into sub-groups of consumers consisting of existing and prospective customers. Market segmentation is a consumer-oriented process and can be applied to almost any type of market. In dividing or segmenting markets, researchers typically look for shared characteristics such as common needs, common interests, similar lifestyles or even similar demographic profiles. So, market segmentation assumes that different segments require different marketing programmes, as diverse customers are usually targeted through different offers, prices, promotions, distributions or some combination of marketing variables. For example, Southwest Airlines' single-minded focus on the short-haul, point-to-point, major-city routes, allowed them to prosper as their competitors floundered. The airline's focus on specific segments allowed them to do a better job of deciding what their target segment really valued (for example, convenience, low price, on-time departures and arrivals, among other things).

Once the customer segments have been identified and profiled, the marketer must decide which segment to target. Diverse customers will have different expectations. For instance, there may be customers who will value a differentiated, high quality service, whilst others may be more price-sensitive. Notwithstanding, not all firms have the resources to serve all customers in an adequate manner. Trying to serve the entire market could be a recipe for disaster. The overall aim of segmentation is to identify high-yield segments. These are likely to be the most profitable groups of customers, or may hold potential for growth. Hence, the most lucrative segments will usually become target markets. In the tourism industry, the business traveller is usually considered as an attractive segment. However, there are different types of business travellers:

- The Hard Money Travellers (or the independent business travellers), these include the business individuals travelling at their own expense;

- The Soft Money Travellers (or corporate business travellers), these include business individuals travelling on an expense account;
- The Medium Money Travellers (or the conference or incentive business travellers), these include business individuals travelling within a group;
- The Interim Travellers, these include business travellers who are combining personal travel with a business trip;
- The Frequent Short Travellers, these include business travellers who consistently fly a short-haul route;
- The Periodic Travellers, these include sales persons who make a round of stops on a steady itinerary.

However, these six groups are said to be only part of some other travel groupings which have often been identified as principal sources of revenue for the tourism industry. Travel and tourism marketers must analyse these various segments. They must then select at least one segment and decide how to service them, in terms of fare prices, facilities, frequencies and special features.

Having defined segmentation and discussed about its benefits, the next question to address is; how could businesses segment their markets? The traditional variables that may be used for market segmentation can be grouped into five main categories: (i) Demographic; (ii) Geographic (iii) Psychographic; (iv) Behavioural and / or (v) Product-Related Factors.

2.2 Tying and assessing clustering technique

> Data is the goldmine in today's ever competitive world. Everyday large amount of information is encountered by organizations and people. An indispensible means to handle this data is to categorize or classify them into a set of groups, partitions or clusters. "Basically classification systems are either supervised or unsupervised ,depending on whether they assign new inputs to one of the finite number of discrete supervised classes or unsupervised categories respectively"

Partitional clustering is highly disimilar to hierarchical approach which yields an incremental level of clusters with iterative fusions or divisons, partitional clustering assignes a set of objects into K clusters with no hierarchical structue[3]. Research from very recent years acknowleges that partitional algorithms are a favoured choice when dealing with large datasets. As these algorithms have comparatively low computational requirements[21]however when it comes to the coherence of clustering ,this approach is less effective then agglomerative aproach. These algorithms deduce the shape of clusters as hyper-ellipsoidal and basically experiment with cutting data into n number of clusters so that partitioning of data optimizes a given criterion. Centroid based techniques as used by K-MEANS and ISODATA assign some points to clusters so that the mean squared distance of points to the centroid of the choosen cluster is minimized

K-means is undoubtedly a very popular partitioning algorithm. It has been discovered, rediscovered and studied by many experts from different fields, by steinhaus(1965),Ball and Hall (1965),Lloyd(proposed 1957 – published 1982) and Macqueen(1967). It is distance-based and by definition data is partitioned into pre-determined groups or clusters. The distance measures used could be Euclidean or cosine. Originally a fixed K cluster centroids [11, 12] are marked at random; k-means reassigns all the points to their closest centroids and re-computes centroids of newly created groups. This iteration continues till the squared error converges. Following steps can summarize the function of k-means. 1. Initialize a K partition based on previous information. A cluster prototype matrix A=[ai…..aj] is created. Where a1,a2,a3 …are cluster centers. Data set D is also initialized. 2. In the next step assignment of each data point in the dataset(di) to its nearest cluster(ai) is performed. 3. Cluster matrix can be recalculated considering the current updated partition or until

ai,aj,ak….. Show no further change. 4. Repeat 2 and 3 until convergence has been reached [20]. K-means is probably the most wildly studied algorithm this is the reason why there exists too many variations and improved versions of k-means yet it can show some sensitivity towards noise and outliers present in data sets. Even if a point is at a distance from the cluster centroid, it could still be enforced to the centre and can result in distorted cluster shape. K- means does not clearly defines a universal method of deciding total number of partitions in the beginning , this algorithm relies heavily on user to provide in advance ,the number of k clusters. Also, k-means is not applicable to categorical data. Since k-means presumes that user will provide initial assignments it can produce replicated results upon every iteration.(the k-means ++ addresses this problem by attempting to choose better starting clusters [12,13]. K-MEDIODS Unlike the k-means, in the k-medioids or partition around medoids (PAM)[13,14] method , a medoid represents any cluster . This characteristic object called the medoid, is the most centrally located point within the cluster [13]. Medoid show better results against outliers as compared to centroids [2]. K-means finds the mean to define accurate centre of the cluster which can result in extreme values but k-medoid calculates the cluster centre using an actual point. Primarily this algorithm attempts to minimize the average dissimilarity of objects against their closest object. The following steps can sum up this algorithm:

1. Initialize: a random k is selected of the n data points as the medoid
2. Assign: each data point should be associated with the closest medoid.
3. Update: for every m medoid and data point d, swapping of m and d can be done compute average dissimilarity of d to all the data points associated with m.

Steps 2 and 3 can be repeated multiple times until there is no further change left in assignments. PAM uses a greedy search resulting in failure in finding an optimum solution.

Cluster analysis is a very crucial paradigm in the entire process of data mining and paramount for capturing patterns in data. This paper compared and analyzed

some highly popular clustering algorithms where some are capable of scaling and some of the methods work best against noise in data. Every algorithm and its underlying technique have some disadvantages and advantages and this paper has comprehensively listed them for the reader. Every paradigm is capable of handling unique requirements of user application. An extensive research and study has been done in the field of data mining and there exist popular real life examples such as Netflix, market basket analysis studies for business giants, biological breakthroughs which use complex combinations of various algorithms resulting in hybrids also and subsequently cluster analysis in the future will unveil more complex data base relationships and categorical data. There is an alarming need of some sort of benchmark for the researchers to be able to measure efficiency and validity of diverse clustering paradigms. The criteria should include data from diverse domains .

2.4 Euclidean Distance and Elbow Method

Euclidean distance matrices (EDM) are matrices of squared distances between points. The definition is deceivingly simple: thanks to their many useful properties they have found applications in psychometrics, crystallography, machine learning, wireless sensor networks, acoustics, and more. Despite the usefulness of EDMs, they seem to be insufficiently known in the signal processing community. Clustering is a data mining technique used to analyse data that has variations and the number of lots. Clustering was process of grouping data into a cluster, so they contained data that is as similar as possible and different from other cluster objects. SMEs Indonesia has a variety of customers, but SMEs do not have the mapping of these customers so they did not know which customers are loyal or otherwise. Customer mapping is a grouping of customer profiling to facilitate analysis and policy of SMEs in the production of

goods, especially batik sales. Researchers will use a combination of K-Means method with elbow to improve efficient and effective k-means performance in processing large amounts of data. K-Means Clustering is a localized optimization method that is sensitive to the selection of the starting position from the midpoint of the cluster. So choosing the starting position from the midpoint of a bad cluster will result in K-Means Clustering algorithm resulting in high errors and poor cluster results.

Imagine that you land at Geneva airport with the Swiss train schedule but no map. Perhaps surprisingly, this may be sufficient to reconstruct a rough (or not so rough) map of the Alpine country, even if the train times poorly translate to distances or some of the times are unknown. The way to do it is by using Euclidean distance matrices (EDM): for a quick illustration, take a look at the "Swiss Trains" box. An EDM is a matrix of squared Euclidean distances between points in a set.1 We often work with distances because they are convenient to measure or estimate. In wireless sensor networks for example, the sensor nodes measure received signal strengths of the packets sent by other nodes, or time-ofarrival (TOA) of pulses emitted by their neighbours

function RANKCOMPLETEEDM(W, De , d) 2: DW ← DeW . Initialize observed entries 3: D1 1 >−W ← μ . Initialize unobserved entries 4: repeat 5: D ← EVThreshold(D, d + 2) 6: DW ← DeW . Enforce known entries 7: DI ← 0 . Set the diagonal to zero 8: D ← (D)+ . Zero the negative entries 9: until Convergence or MaxIter 10: return D 11: end function 12: function EVTHRESHOLD(D, r) 13: U, [λi] n i=1 ← EVD(D) 14: Σ ← diag λ1, . . . , λr, 0, . . . , 0 | {z } n−r times 15: D ← UΣU > 16: return D 17: end function

At the end of this tutorial, we hope that we succeeded in showing how universally useful EDMs are, and that we inspired readers coming across this material for the first time to dig deeper. Distance measurements are so common that a simple, yet sophisticated tool like EDMs deserves attention. A good example is the semidefinite relaxation: even though it is generic, it is the best performing algorithm for the specific problem of ad-hoc microphone array localization.

1. Determine the number of clusters K and the number of maximum iterations

. 2. Perform the initialization process K midpoint cluster, then the equation of centroid count feature:

∑ Equation 1 is done as much as p dimensions from i = 1 to i = p

3. Connect any observation data to the nearest cluster. Euclidean distance spacing measurements can be found using equation 2. √( ) ( )

4. Reallocation of data to each group based on comparison of distance between data with each group's centroid .

5. Recalculate the cluster midpoint position. is the value of the membership of point xi to the centres of the group c1, d is the shortest distance from the data xi to the group K after being compared, and c1 is the centre of the group to 1.

The objective function used by this method is based on the distance and the value of the data membership in the group. The objective function according to MacQueen (1967) can be determined using equation. ∑∑ ( ) n is the amount of data, k is the number of groups, ai1 is the membership value of the data point xi to the c1 group followed a has a value of 0 or 1. If the data is an anngota of a group, the value ai1 = 1. If not, the value ai1 = 0.

6. If there is a change in the cluster midpoint position or number of iterations

CHAPTER 3

SYSTEM REQUIREMENTS SPECIFICATION

A System Requirement Specification (SRS) is an organization's understanding of a customer or potential client's system requirements and dependencies at a particular point prior to any actual design or development work. The information gathered during the analysis is translated into a document that defines a set of requirements. It gives the brief description of the services that the system should provide and the constraints under which, the system should operate. Generally, SRS is a document that completely describes what the proposed software should do without describing how the software will do it. A two-way insurance policy assures that both the client and the organization understand the other's requirements from that perspective at a given point in time.

SRS document itself states in precise and explicit language those functions and capabilities a software system (i.e., a software application, an ecommerce website and so on) must provide, as well as states any required constraints by which the system must abide. SRS also functions as a blueprint for completing a project with as little cost growth as possible. SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it.

Requirement is a condition or capability to which the system must conform. Requirement Management is a systematic approach towards eliciting, organizing and documenting the requirements of the system clearly along with the applicable attributes. The elusive difficulties of requirements are not always obvious and can come from any number of sources.

3.1 General Description

In this section of the presented thesis, the introduction of software product under consideration has been presented. It presents the basic characteristics and factors influencing the software product or system model and its requirements.

3.1.1 Product Perspective

Segments should be easily measurable after targeting. The size, profiles and purchasing power of the segments can be measured. Sometimes, segmentation variables are difficult to measure. For example, if any company wants to launch any product which is useful for left-handed people. In America, there are more than 3 million people are left-handed, which is somehow equal to the entire population of Canada. So, if they target their market towards left-handed segment, then the major problem may be that it will be difficult to identify and measure..

Markets segments should be effectively reached and served to the targeted market. Marketers have to make their products accessible for customers. Suppose, a fragrance company gets that heavy users of one of its brand are single men and women who stay out late and socialize a lot. Unless this group shops from company's outlets, its members will be difficult to identify. The market segments should be large or profitable to serve. Segments should be huge possible homogeneous groups worth pursuing with an altered marketing program. For example, it would not pay to automobiles manufactures, if they manufacture cars especially for those customers whose heights are more than 6.5 feet.

3.1.2 User Characteristics

The user should have at least a basic knowledge of windows and web browsers, such as install software like Python 3.6.8 etc. and executing a program, and the ability to follow on screen instructions. The user will not need any technical expertise in order to use this program.

3.2 Functional Requirement

- The data representing customer behaviour is to be provided .
- with the machine accuracy.
- The program will be able to produce multiple cluster and different was of assessing the customer .
- The program will be able to choose the right set of customer to be targeted based on the input s.
- There is a way to tackle each customer as per requirement .

3.3 Non-Functional Requirement

- **Usability:** The user is facilitated with the control section for the entire process in which they can put in the customer data under consideration, the data is similar to wicha the model is constructed command generation etc. can be effectively facilitated by mean of user interface

  The data is similar to previous data given then there is no problem the user has to just feed the new data and run the program to obtain the result.

- **Security and support:** Application will be permissible to be used only in secure network so there is less feasibility of insecurity over the functionality of the application. On the other hand, the program recides in the user machine and can easily protect it by pc security .

- **Maintainability:** The installation and operation manual of the project will be provided to the user.

- **Extensibility:** The project work is also open for any future modification and hence the work could be define as the one of the extensible work.

3.5 External Interface Requirement

An interface description for short is a specification used for describing a software component's interface. IDLs are commonly used in remote procedure call software. In these issues, the machines on moreover last part of the "link" might be utilizing

Dissimilar Interface Description recommends a bridge between the two diverse systems. These descriptions are classified into following types:

- **User Interface:** needs a little understanding of python and running the code.

- **Restoration with Text Removal Software Interface:** The Operating Systems can be any version of Windows, Linux, UNIX or Mac.

- **Hardware Interface:** In the execution of this project, the hardware interface used is a normal 32/64 bit operating system supported along with better integration with network interface card for better communication with other workstations. For better and precise outcome, a high definition camera with calibrated functioning with

CHAPTER 4

SYSTEM ANALYSIS

Analysis is the process of finding the best solution to the problem. System analysis is the process by which we learn about the existing problems, define objects and requirements and evaluates the solutions. It is the way of thinking about the organization and the problem it involves, a set of technologies that helps in solving these problems. Feasibility study plays an important role in system analysis, which gives the target for design and development.

4.1 Feasibility Study

All systems are feasible when provided with unlimited resource and infinite time. But unfortunately, this condition does not prevail in practical world. So it is both necessary and prudent to evaluate the feasibility of the system at the earliest possible time. Months or years of effort, thousands of rupees and untold professional embarrassment can be averted if an ill- conceived system is recognized early in the definition phase. Feasibility & risk analysis are related in many ways. If project risk is great, the feasibility of producing quality software is reduced. In this case three key considerations involved in the feasibility analysis are:

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

4.1.1 Economic Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

4.1.2 Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

4.1.3 Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcome, as he is the final user of the system.

4.2 Analysis

4.2.1 Performance Analysis

For the complete functionality of the project work, the project is run with the help of healthy networking environment. Performance analysis is done to find out whether the proposed system. It is essential that the process of performance analysis and definition must be conduct in parallel.

4.2.2 Technical Analysis

System is only beneficial only if it can be turn into information systems that will meet the organization's technical requirement. Simply stated this test of feasibility asks whether the system will work or not when developed & installed, whether there are implementation. Regarding all these issues in technical analysis there are several points to focus on:

Changes to bring in the system: All changes should be in positive direction, there would be increased level of efficiency and better customer service.

Required skills: Platforms & tools used in this project are widely used. Therefore, the skilled work force is readily available in the industry.

Acceptability: The structure of the system is kept feasible enough so that there should not be any problem from the user's point of view.

4.2.3 Economic Analysis

Economic analysis is perform to evaluate the development cost weighed against the ultimate income or benefits derived from the developed system. For running this system, we need not have any routers, which are highly economical. Therefore, the system is economically feasible enough.

CHAPTER 5

SYSTEM DESIGN

Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the developments of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing. Design is the perfect way to accurately translate a customer's requirement in the finished software product. Design creates a representation or model, provides details about software data structure, architecture, interfaces and components that are necessary to implement a system. The logical system design arrived at because of systems analysis is converted into physical system design.

5.1 System Development methodology

System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps, which are used for product progress. The development method followed in this project is waterfall model.

5.1.1 Model Phases

The waterfall model is a sequential software development process, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Requirement initiation, Analysis, Design, Implementation, Testing and maintenance.

Requirement Analysis: This phase is concerned about collection of requirements of the system. This process involves generating document and requirement review.

System Design: Keeping the requirements in mind the system specifications are translate in to a software representation. In this phase, the designer emphasizes on-algorithm, data structure, software architecture etc.

Coding: In this phase, programmer starts his coding in order to give a full sketch of product. In other words, system specifications are only convert in to machine-readable compute code.

Implementation: The implementation phase involves the actual coding or programming of the software. The output of this phase is typically the library, executables, user manuals and additional software documentation

Testing: In this phase, all programs (models) are integrated and tested to ensure that the complete system meets the software requirements. The testing is concerned with verification and validation.

Maintenance: The maintenance phase is the longest phase in which the software is updated to fulfil the changing customer need, adapt to accommodate change in the external environment, correct errors and oversights previously undetected in the testing phase,

5.1.2 Advantages of Waterfall Model

- Clear project objectives.
- Stable project requirements.*
- Progress of system is measurable.
- Strict sign-off requirements.
- Helps you to be perfect.
- Logic of software development is clearly understood.
- Production of a formal specification.
- Better resource allocation.

- Improves quality. The emphasis on requirements and design before writing a single line of code ensures minimal wastage of time and effort and reduces the risk of schedule slippage.

-

- Less human resources required as once one phase is finished those people can start working on to the next phase

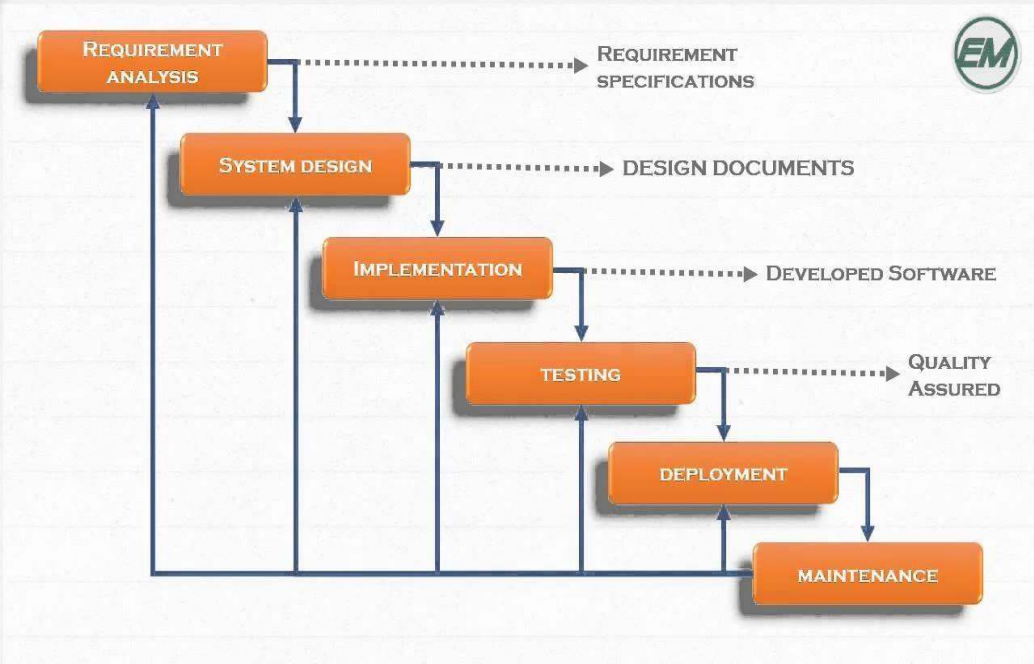

Fig 5.1 Waterfall Model

5.2 Design Using UML

Designing UML diagram specifies, how the process within the system communicates along with how the objects with in the process collaborate using both static as well as dynamic UML diagrams since in this ever-changing world of Object Oriented application development, it has been getting harder and harder to develop and manage high quality applications in reasonable amount of time. Because of this challenge and the need for a universal object modelling language

everyone could use, the Unified Modelling Language (UML) is the Information industries version of blue print. It is a method for describing the systems architecture in detail. Easier to build or maintains system, and to ensure that the system will hold up to the requirement changes.

5.3 Data Flow Diagram

The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generate by the system.

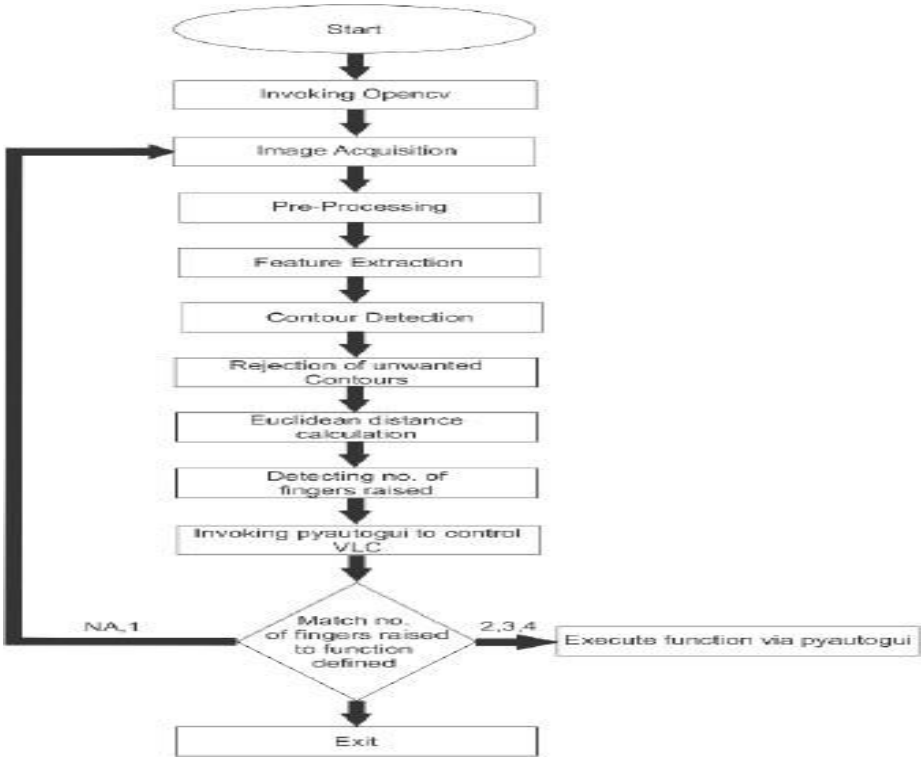

Fig 5.2 Data Flow Model

The data flow diagram essentially shows how the data control flows from one module to another. Unless the input filenames are correctly given the program cannot proceed to the next module. Once the user gives, the correct input filenames parsing is done individually for each file. The required information is taken in parsing and an adjacency matrix is generated for that. From the adjacency matrix, a lookup table is generated giving paths for blocks. In addition, the final sequence is computed with the

lookup table and the final required code is generated in an output file. In case of multiple file inputs, the code for each is generated and combined together..

5.5 Use Case Diagram

A use case defines a goal-oriented set of interactions between external entities and the system under consideration. The external entities, which interact with the system, are its actors. A set of use cases describe the complete functionality of the system at a particular level of detail and the use case diagram can graphically denote it.

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.

In software and systems engineering, a use case is a list of steps, typically defining interactions between a role (known in Unified Modelling Language (UML) as an "actor") and a system, to achieve a goal. The actor can be a human, an external system, or time.

In systems engineering, use cases are use at a higher level than within software engineering, often representing missions or stakeholder goals. The detailed requirements may then be capture in Systems Modelling Language (SysML) or as contractual statements.

The Sequence of activities that are carried out are the same as the other diagrams .Use case for this module indicates the users interaction with the system as a whole rather than individual modules .All the encryption mechanisms are carried out via the login page that redirects the user to the particular functionality that he or she wishes to implement.

5.6 Activity Diagram

An activity diagram shows the sequence of steps that make up a complex process. An activity is shown as a round box containing the name of the operation. An outgoing solid arrow attached to the end of the activity symbol indicates a transition triggered by the completion.

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams are intended to model both computational and organisational processes (i.e. workflows). Activity diagrams show the overall flow of control.

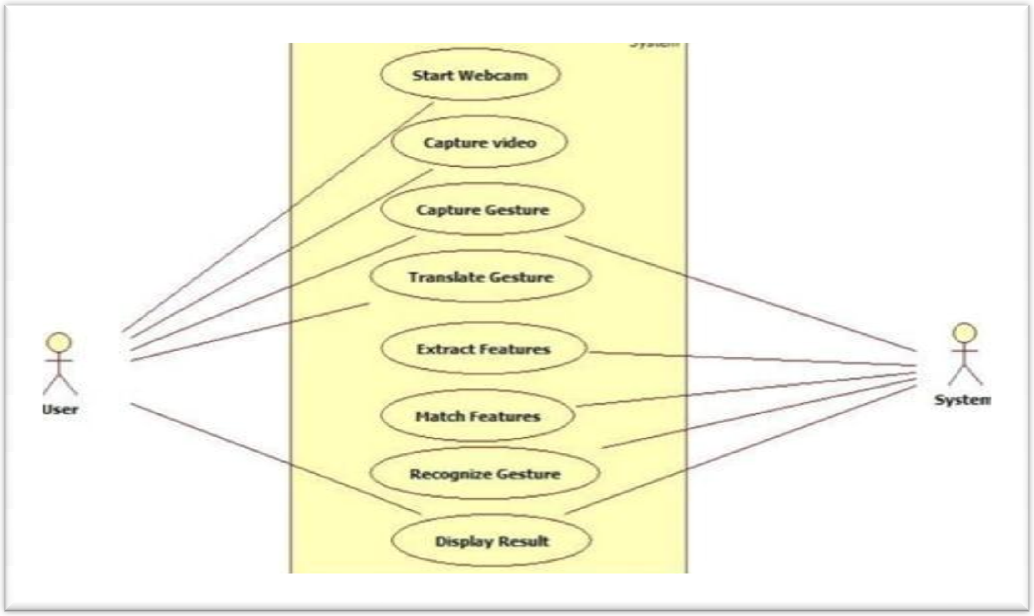

Fig 5.4 Use Case Diagram

Activity diagrams are constructed from a limited number of shapes, connected with arrows. The most important shape types:

- rounded rectangles represent actions;
- diamonds represent decisions;
- bars represent the start (split) or end (join) of concurrent activities;
- a black circle represents the start (initial state) of the workflow;
- An encircled black circle represents the end (final state).

The basic purposes of activity diagrams are similar to other four diagrams. It captures the dynamic behaviour of the system. Other four diagrams are used to show the message flow from one object to another but activity diagram is used to show message flow from one activity to another.

Activity is a particular operation of the system. Activity diagrams are not only use for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part.

CHAPTER 6

PROPOSED SYSTEM

6.1 Data Obtaining

The initial move is to capture the in xl format from the required source fro which the study need to be conducted ,billing statements ,bank data network data and phone

6.3 Classification

Classification was a continuous ongoing process in our analysis. We performed 4-folds cross validation to ensure that we were not over-fitting. Instead of typical 10-cross validation, we took 4- folds by considering each subject as test case for one iteration and the remaining as training set. While performing cross validation to retain the temporal aspect of the data, we did not randomize our folds. In case of 10-folds, the accuracy of the classifier would increase, but that would be over fitting since a section of the test subject's data would already be there in the training set. We can identify overfitting by looking at validation metrics, like loss or accuracy. Usually, the validation metric stops improving after a certain number of epochs and begins to decrease afterward. The training metric continues to improve because the model seeks to find the best fit for the training data.

Fig 6.1 RNN Flow

units, respectively.

Proposed System Implementation Code

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.cluster import KMean


# Read the dataset

df =pd.read_csv(r'C:\Users\tarun saini\Desktop\Mall_Customers.csv')

df


df.shape

x = df.Income

y = df.SpendingScore

plt.scatter(x,y)

plt.show()
```

```
x = df.iloc[:, [3,4]].values

# Using the elbow method to find the optimal number of clusters

ls = []

for i in range(1,11):

    km = KMeans(n_clusters = i)

    km.fit(x)

    ls.append(km.inertia_) # Sum of squared distances of samples to their closest cluster center

plt.plot(range(1,11), ls)

plt.title('The Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel('SSE')

plt.show()

ls

# Fitting K-Means to the dataset

km = KMeans(n_clusters =5)

y_kmeans = km.fit_predict(x)
```

```
# Visualising the cluster

plt.scatter(x[y_kmeans == 0 , 0], x[y_kmeans == 0, 1], s = 50, c = 'red', label =
'Cluster 1') # Variable as two values,

# first zero is cluster 0 and second 0 is Income 1 is SpendingScore

plt.scatter(x[y_kmeans == 1 , 0], x[y_kmeans == 1, 1], s = 50, c = 'blue', label =
'Cluster 2')

plt.scatter(x[y_kmeans == 2 , 0], x[y_kmeans == 2, 1], s = 50, c = 'green', label
= 'Cluster 3')

plt.scatter(x[y_kmeans == 3 , 0], x[y_kmeans == 3, 1], s = 50, c = 'cyan', label =
'Cluster 4')

plt.scatter(x[y_kmeans == 4 , 0], x[y_kmeans == 4, 1], s = 50, c = 'magenta',
label = 'Cluster 5')

plt.scatter(km.cluster_centers_[:,0], km.cluster_centers_[:,1], s = 100, c
='yellow', label = 'Centroids')

plt.title('Clusters of Customer')

plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1-100)')

plt.legend()
```

```
plt.show()


x[y_kmeans == 0]


# The cluster centers are stored in the cluster_centers_ attribute, and we plot them as triangles

print("The centers of clusters are:\n", km.cluster_centers_) # Kluster center is a point to which every other point is related to
```

CHAPTER 7

RESULTS AND DISCUSSIONS

Here in the project we have obtained the different cluster based on the the salary and there spending score. Different colours are used to identify different cluster which shows different behaviour .

cluster 1 the people in this cluster have a good salary but doesn't prefer buying

cluster 2 the people in this cluster have a low salary but doesn't prefer buying .

cluster 3 the people in this cluster have a good salary and also prefer buying

cluster 4 the people in this cluster have a low salary and still prefer buying.

cluster 5 the people in this cluster have a average salary and also cant bye costly but mediocre things .

Customer segmentation has highly improved the performance of different businesses over the years, the fact that the consumer platform is changing dramatically has proven to be a headache for various businesses. That's why segmenting the customers into different groups has some incredible benefits for improving the market results

Screenshots Section

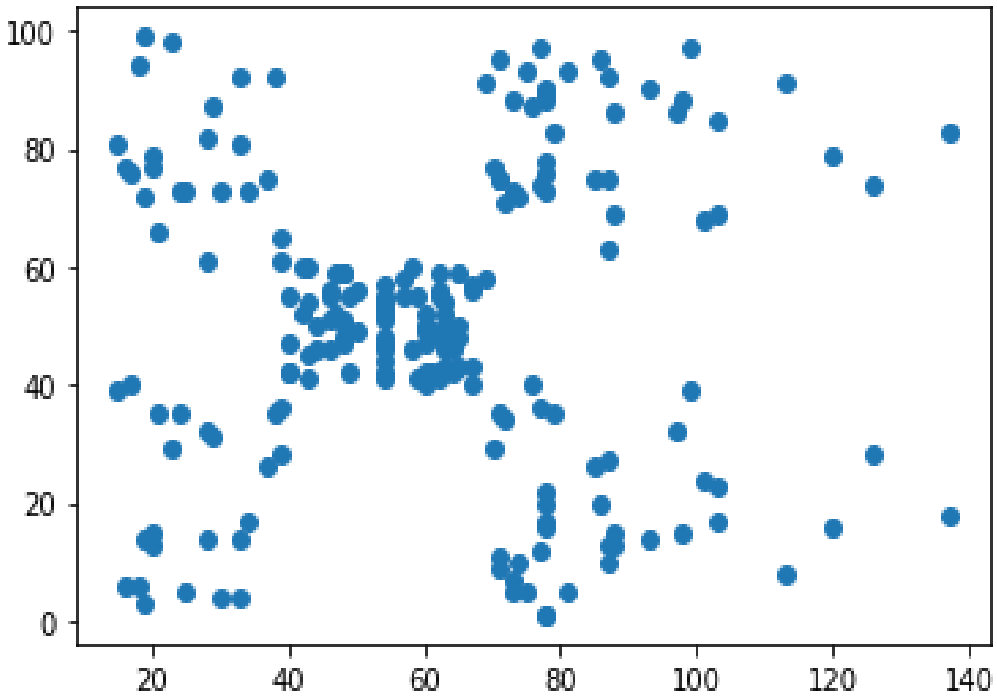

Fig 7.1 data representation on a xy plane

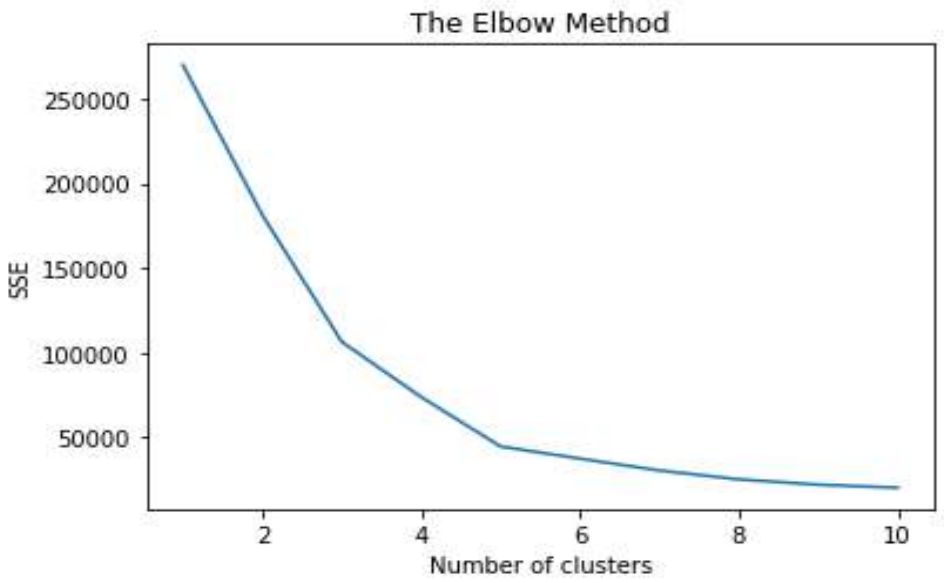

Fig 7.2 Elbow method

Fig 7.3 clusters formed .

CHAPTER 8

TESTING

Customer Segmentation basically realise on the clustering algorithm ,It is sometimes useful to construct input data where there is a known, and perhaps obvious, answer by construction. For a clustering algorithm, you might construct data with N clusters such that the maximum distance between any two points in the same cluster is smaller than the minimum distance between any two points in different clusters. Another option would be to generate a number of different data sets plotable as 2-d scatter diagrams with clusters obvious to the eye, then compare the result from your algorithm with this structure, perhaps moving the clusters together to see when the algorithm fails to see them.

You might be able to do better given knowledge of your particular clustering algorithm, but the above might at least have some chance of flushing obvious bugs from cover.

. In the testing stage following goals are tried to achieve: -

- To affirm the quality of the project.
- To find and eliminate any residual errors from previous stages.
- To validate the software as a solution to the original problem.
- To provide operational reliability of the system.

8.1 Quality Assurance

Quality assurance consists of the auditing and reporting functions of management. The goal of quality assurance is to provide management with the data necessary to be informed about product quality, thereby gaining insight and confidence that the product quality is meeting its goals. This is an “umbrella activity” that is applied throughout the engineering process. Software quality assuranceencompasses: -

- Analysis, design, coding and testing methods and tools
- Formal technical reviews that are applied during each software engineering
- Multi

-tiered testing strategy

- Control of software documentation and the change made to it.
- Measurement and reporting mechanisms.

8.2 Quality Factors

An important objective of quality assurance is to track the software quality and assess the

impact of methodological and procedural changes on improved software quality. The factors that affect the quality can be categorized into two broad groups:

- Factors that can be directly measured.
- Factors that can be indirectly measured
- Effectiveness or efficiency in performing its missio
- Its ability to undergo changes
- Its adaptability to a new environment.
- Duration of its use by its customer.

8.3 Functional Test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is cantered on the following items:

Valid Input identified classes of valid input must be accepted.

Invalid Input: identified classes of invalid input must be rejected.

Functions: identified functions must be exercised.

Output:identified classes of application outputs must be exercised. Systems/Procedures: Interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

CHAPTER 9

CONCLUSION AND FUTURE

9.1 Conclusion

Customer segmentation is a way to improve communication with the customer, to know the wishes of the customer, customer activity so that appropriate communication can be built. Customer Segmentation needed to get potential customers used to increase profits. Potential customer data can be used to provide service the characteristics of customer including ecommerce services as a media buying and selling online. This paper discusses several components to do customer segmentation, which is: Customer segmentation is an activity to divide• customers or item into groups that have the same characteristics. Data that needed for customer segmentation are• internal data and external data. The internal data include demographic data and data purchase history, while the external data include cookies and server logs. Internal data can be obtained from a database when customer do registration or transactions and external data can be obtained from web server or other source. Methods of Customer Segmentation can be classified into Simple technique, RFM technique, Target technique, and Unsupervised technique. On Target technique, researcher focus on one variable, it can be product or purchase. Unsupervised technique was used when clustering process reseacher have many variable Process of Customer Segmentation can be• simplified into defining business objective, collecting data, data preparation, analyzing var

REFERENCES

[1] [1] Al-Qaed F, Sutcliffe A. Adaptive Decision Support System (ADSS) for B2C E-Commerce. 2006 ICEC Eighth Int Conf Electron Commer Proc NEW E-COMMERCE Innov Conqu Curr BARRIERS, Obs LIMITATIONS TO Conduct Success Bus INTERNET. 2006:492-503.

[2] [2] Mobasher B, Cooley R, Srivastava J. Automatic Personalization Based on Web Usage Mining. Commun ACM. 2000;43(8).

[3] [3] Cherna Y, Tzenga G. Measuring Consumer Loyalty of B2C e-Retailing Service by Fuzzy Integral: a FANP-Based Synthetic Model. In: International Conference on Fuzzy Theory and Its Applications iFUZZY.; 2012:48-56.

[4] [4] Magento. An Introduction to Customer Segmentation. 2014. info2.magento.com/.../ An_Introduction_to_Customer_Segmentation...