T.C.
BAHÇEŞEHİR UNIVERSITY

CUSTOMER SEGMENTATION FOR CAMPAIGN MANAGEMENT BY USING FUZZY C-MEANS

M.S. Thesis

Ümit ŞAHİN

İstanbul, 2011

T.C.

BAHÇEŞEHİR UNIVERSITY
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

CUSTOMER SEGMENTATION FOR CAMPAIGN MANAGEMENT BY USING FUZZY C-MEANS

M.S. Thesis

Ümit ŞAHİN

Supervisor: **Assoc. Prof. Dr. Adem KARAHOCA**

İstanbul, 2011

T.C
BAHÇEŞEHİR UNIVERSITY
The Graduate School of Natural and Applied Sciences
Computer Engineering

Title of the Master's Thesis	: Customer Segmentation for Campaign Management by Using Fuzzy C-means
Name/Last Name of the Student	: Ümit ŞAHİN
Date of Thesis Defense	: 20.01.2011

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Asst. Prof. F. Tunç BOZBURA
Acting Director

This is to certify that we have read this thesis and that we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members:

Assoc. Prof. Dr. Adem KARAHOCA :

Asst. Prof. Dr. Mehmet Alper TUNGA :

Asst. Prof. Dr. Yalçın ÇEKİÇ :

ABSTRACT

CUSTOMER SEGMENTATION FOR CAMPAIGN MANAGEMENT BY USING FUZZY C-MEANS

ŞAHİN, Ümit

COMPUTER ENGINEERING

Supervisor: Assoc. Prof. Dr. Adem KARAHOCA

February 2011, 99 pages

Nowadays, having control of customer is vital based on product diversity in the banking sector. The old policies of the banks lost their validity because they were general and they were not responding to individual needs which determine choice criteria of the customer for selecting a bank. For this reason, banks have to develop campaign strategies to manage relations with all their customers. Therefore, banks have to segment customers to provide them valuable solutions and products. In other words, customer segmentation is used for selecting appropriate customers is needed to determine campaign management. In this study, banking customers will be segmented adaptively to present them suitable and beneficial services.

Keywords: CRM, Campaign strategies, Campaign Management, Customer Segmentation.

ÖZET

KAMPANYA YÖNETİMİ İÇİN FUZZY C-MEANS ALGORİTMASI
KULLANILARAK MÜŞTERİ SEGMENTAYONU YAPILMASI

ŞAHİN, Ümit

BİLGİSAYAR MÜHENDİSLİĞİ
Tez Danışmanı: Doç. Dr. Adem KARAHOCA

Şubat 2011, 99 sayfa

Bugünlerde, müşteriye ait ürün sahipliğinin kontrolü bankacılık sektöründe hayati bir önem arzetmektedir. Bankacılıkta kullanılan eski politikalar geçerliliğini yitirmiştir. Bunlar genel politikalardı ve müşterinin banka seçimini belirlerken dikkat ettiği bireysel ihtiyaçlarına cevap veremiyorlardı. Bu sebeple, bankalar bütün müşterileriyle olan ilişkilerini yönetmek için kampanya stratejileri geliştirmek durumundadırlar. Bu yüzden, bankalar müşterilerinin ihtiyaçlarına yönelik çözümler ve ürünler sağlamak için onları sınıflandırmak zorundadırlar. Başka bir deyişle, birbirine yakın özelliklere sahip olan müşterilerin bir araya getirilmesi, onlar için uygun kampanya stratejileri belirlenmesi için gereklidir. Bu çalışmada, bankacılık müşterilerine daha uygun ve yararlı hizmetler sunmak için onları sınıflara ayıracağız.

Anahtar Kelimeler: Müşteri İlişkileri Yönetimi, Kampanya Stratejileri, Kampanya Yönetimi, Müşteri Segmentasyonu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 PROBLEM DEFINITION

Banks are the most important supporter of the global economy, providing capital for innovation, infrastructure, job creation and overall prosperity. Banks also are company and they have to derive a profit from the services on money. Nowadays, almost all banks offer the same services to the customers. For this reason, service is not major factor for the customer. The quality of service and the requested fee for the service are the most important factors to determine customer's choice. Therefore, banks always struggle greatly in banking sector.

Change is the only constant factor in this dynamic world and banking is not an exception. In the past, banks earned money only from deposit and loan services. Today, they have a lot of services and they serve as a bridge between customer and world with these services. Therefore, they have to carefully review on both sides of the bridge. Banks always have strategic plans due to the conditions in global economy but they also know that these plans are worthless without knowing customer needs. Customer is basis of the banking sector because there is no need for bank if there is not customer's needs. All the struggles in banking sector have always been made for the customers. The challenges in banking sector can be divided into two parts.

The challenging on customer loyalty which refer people have a relation with bank is one of them. The product diversity is the most important factor to improve loyalty between customer and bank. The numbers of services which are serviced by bank to the customer are increased; the possibility of customer churn is decreased. Therefore, banks always have to observe the services of customer to protect or improve the customer loyalty.

The challenging on getting new customer which refer people have no relation with bank is the another part. Banks always try to get new customer to increase the number of customer so they also increase their profitability. The management of challenges which mentioned above is vital for the bank. Therefore, these challenges are examined in the customer relationship management.

1.2 CUSTOMER RELATIONSHIP MANAGEMENT

Customer relationship management (CRM) is more necessary today because of the increasing rate of change in the consumer market. The rapid changes in the requirements of customers are distinct from each other. CRM is the main means by which businesses can face these challenges, and it is able to help them grasp the varied demand of customers and then earn competitive advantage (Anderson ET, 2002; Anderson K & Kerr C, 2001). CRM can be defined as a dynamic process of managing a customer–bank relationship such that customers elect to continue mutually beneficial commercial exchanges and are dissuaded from participating in exchanges that are unprofitable to the company (Bergeron B, 2002).

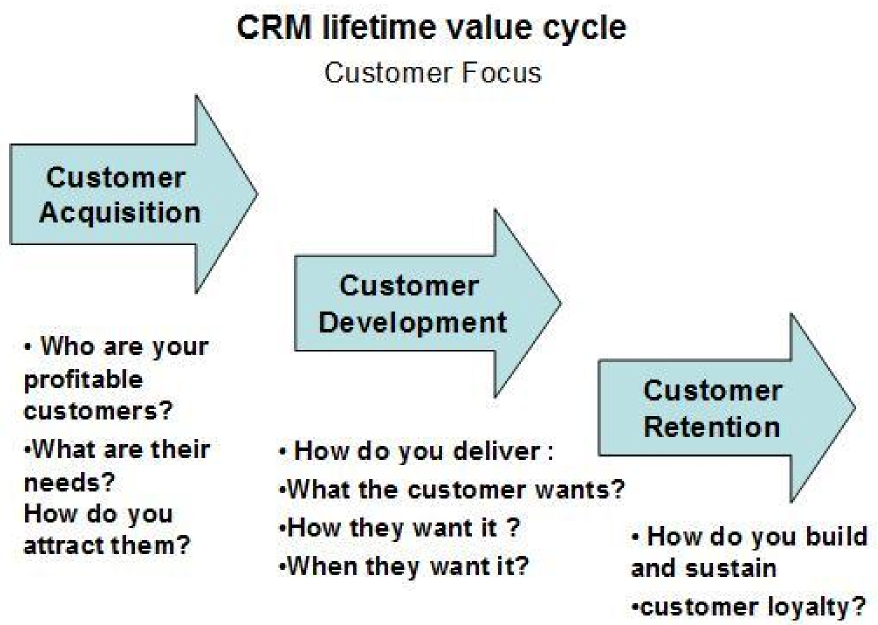


Figure 1.1 : Customer lifetime value defined by IBM (Liu, 2001).

While approaching the 2000s, CRM became one of the most important factors that influence customer for the strategic positioning of a company. CRM takes all information about who is the customer as a superficial. All of marketing activities are designed and implemented with this information. Knowing the customer's needs, metadata knowledge, good coordination and analysis on the information are basis of the CRM. Customer lifetime value is improved when all of them occur successfully.

CRM is very important for companies that have different groups of customers and use different interaction channels for them. The existing information should be updated with information from all channels so companies can access the most current information about the customers.

Knowing the customer's behavior and understanding the next step of customer is the core of CRM. Every action can be easily taken in customer retention, customer loyalty and customer profitability to maximize customer lifetime value, after customer understanding occurs successfully. Customer understanding is the most important factor to set the road map of company. The analysis of customer understanding should be done carefully otherwise the poor results may occur. These poor results can lead to decrease of customer lifetime value. Therefore, correct customer understanding has an important role in CRM.

Figure 1.2 shows an idealized CRM cycle.

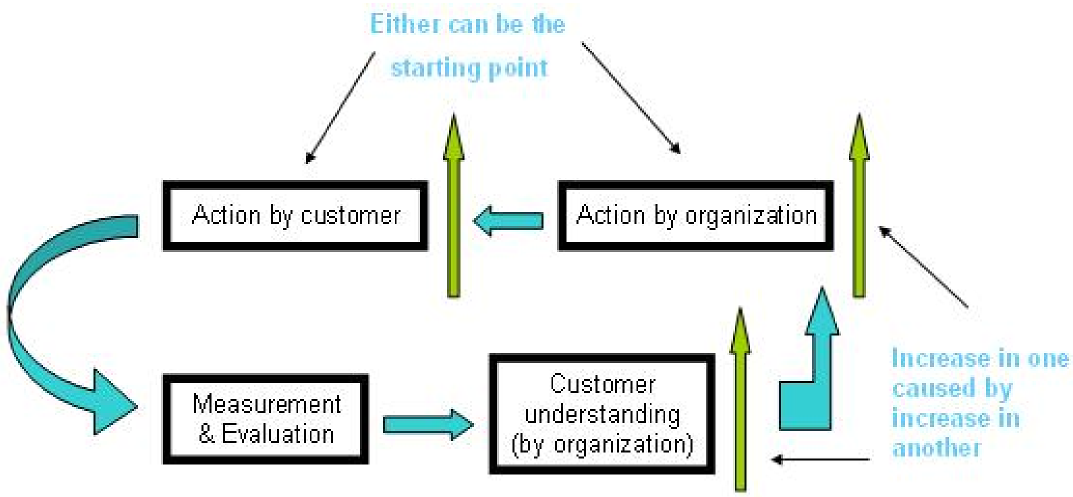


Figure 1.2 : The basic CRM cycle.

In this figure, boxes represent actions:

- The customer makes contact with company, e.g. to buy something, to obtain information for after sales support, to make a reclamation or a suggestion etc.
- The company makes contact with customer, e.g. to sell something, to obtain information about customer satisfaction for its product or service which is sold or inform people about its products or services etc.
- The company analyzes the information available from customer and company actions to understand customer needs. This analysis determines the road map of company towards the customer, both when customer contacts the company and company contacts customer.

The functionality of CRM, especially in corporate companies, is transformed into a structure that surrounds the customer in more detail when it is compared with the central coordination and integration suggested by Figure 1.2:

- Many channels of company collect the information about the customer, but some of this information is not relevant with customer knowledge.
- The all information about the customer is actively brought together to implement plans about actions of customer, such as marketing campaigns and the producing of new products. The results of these actions are not sometimes available because these actions can be taken out of organization borders or the unit which collects information does not know the unit which takes actions. Therefore, the organization of company must be created to include these situations.
- There may not be able to obtain all information about the customer because of some causes. One of them, the competition between departments of company may prevent data sharing. Another of them, it is not acceptable extracting the information out of the unit due to the data security. On the other hand, data owners have to often describe to data collectors to achieve advance customer understanding so coordination and information sharing between units is very important. Finally, internal strife and confidentiality constraints generate information-sharing barriers.

- The unit gathers information for a special purpose does not always share the data unconditionally with other units for other purposes of company.
- Hence, all the information about the customer should not bring together in a corporate warehouse. The analysis should be developed in non-integrated data while preserving privacy and confidentiality constraints.

CRM is comprehensive issue, includes all issues associated with the customer, data mining techniques are used in some of them. Some of these issues which contain functions about data mining are customer segmentation, customer loyalty, customer retention, response analytics, contact management, call management, etc. Now, we focus on how data mining and analytics can make these issues more meaningful and effective.

1.3 DATA MINING

Today's finance world, one of the strategically important areas for many organizations is data mining. Data mining can be explained as a process of analyzing the data and summarizing it into desirable and valuable information. In the definition of data mining "the information" refers all data and information which can be used to increase revenue, reduce cost or both

Data mining serves to look for hidden pattern in a group and to find out unknown relationships in the data.

Nowadays, the value of collecting customer data is recognized by most marketers. They also realize the challenges of leveraging this knowledge to create intelligent, proactive pathways back to the customer.

One of the important issues which are concluded by data mining is recognizing and tracking patterns within data. This shows the way to businesses how to build a meaningful relationship between unrelated data. It tries to explain anticipated relations rather than simply react to customer needs.

1.3.1 How does data mining work?

One of the important issues which are concluded by data mining is recognizing and tracking patterns within data. This shows the way to businesses how to build a meaningful relationship between unrelated data. It tries to explain anticipated relations rather than simply react to customer needs.

- **Classes:** The predetermined groups in data are revealed when the stored data is examined. As an example, a supermarket could store the data that determining what customers typically buy when they come the supermarket. Daily promotions can easily be determined by customer groups which are occurred from the storage data of supermarket.
- **Clusters:** The components of data are clustered according to logical relationships or customer preferences. As an example, segments of customers or connections between customers can be identified by the stored data.
- **Associations:** Associations between customers can be provided by data mining. The most well-known example of associative mining is beer-diaper example which is applied in supermarkets.
- **Sequential patterns:** Customer trends and behavior patterns can be estimated by data mining. As an example, the likelihood of a swimsuit being purchased based on a customer's purchase of beach slippers and sea towels could be anticipated by a sports store.

Data mining can be expressed by the following five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Regression**: The most well-known statistical technique that the data mining community utilizes is regression. Essentially, a numerical data set is used by regression, and then regression improves a mathematical formula which fits the data. When you're ready to use the results that help to get a prediction about future behavior, you clearly get your new data and fill it into the improved formula. This technique only works well with continuous quantitative data such as weight, speed or age. It will be suitable to choose another technique, if you're working with categorical data where order is not important in profession, city or gender.

- **Classification**: Classification is a data mining technique used to put customer into groups according to their characteristics. This technique is capable of processing a wider variety of data than regression and is growing in popularity.

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: A decision tree is a predictive model that, as its name implies, looks like a tree structure. The braches of tree are divided into two according to response of the classification question and the last branches of the tree which are not divided are partitions of the dataset with their classification.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

The all of data mining levels will be used to reach most profit and applicable approaches for campaign management.

1.4 CAMPAIGN MANAGEMENT

Marketing decisions, such as promotions, distribution channels and advertising media, based on traditional segmentation approaches result in poor response rate and increased cost. In fact, each customer wants to be served according to her individual and unique needs.

The campaign management is the address where policies of the bank are applied. The general policies determine the target position of the bank at the end of the year and they can not be applied to every customer of bank. For example; a policy for increasing the number credit cards can not include every customer because some of customer already have credit card so customers who already have credit cards will be out of the target audience. On the other hand, the limit of credit card depends on the customer's income so you can not offer same limit to every customer. Because some of customer may not pay their loan when they spend their money which is borrowed so bank will be careful about the limit of credit card. Therefore, general policies only show the way to the aim but they are not applicable on the way. On the other hand, the specific policies are needed to reach main business goals and specific policies applied to target audience. Target audience is formed by customer segmentation. Knowledge-based marketing, which uses appropriate data mining tools and knowledge management framework, addresses this need and helps leverage knowledge hidden in databases. There are lots of areas in application of data mining for knowledge-based marketing.

In this paper, we will focus on customer segmentation for campaign management.

1.4.1 Customer Segmentation

- How can you maximize the cost-effectiveness and value of your strategies in the current economic climate?
- What criteria should you use in your segmentation strategy; should it now be based on customer needs, attitudes and experience?
- How can you collect accurate and meaningful data from your customers to carry out cost-effective segmentation?
- How can you successfully implement a micro-segmentation strategy?
- How should you apply the results of your segmentation to develop strategies for improving campaign efficiency, customer retention and customer profitability?
- How can you use your customer intelligence to deliver more value to your customers, particularly at a time when they want to spend less?

The questions are on the above are the general questions which are expected to answer when the customer segmentation is occurred.

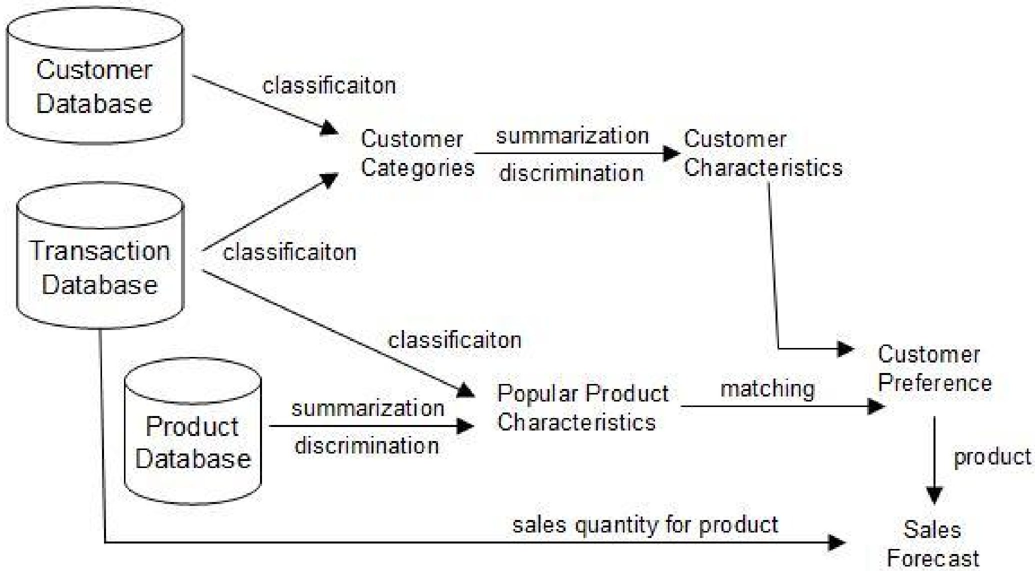


Figure 1.3 : Customer profiling system (Michael & Chandrasekar & Gek Woo & Michael, 2001).

Customer segmentation is used to define the target audience profile in general. The segmentation process is directly cleaved to the customer profile and customer profile is determined by customer value. In the other words, customer segmentation is based on customer value.

1.4.1.1 Customer value

Customer value has been studied under the name of LTV (Life Time Value), CLV (Customer Lifetime Value), CE (Customer Equity) and Customer Profitability. The previous researches define LTV as the sum of the revenues gained from company's customers over the lifetime of transactions after the deduction of the total cost of attracting, selling, and servicing customers, taking into account the time value of Money (Dwyer, 1997; Hoekstra & Huizingh, 1999; Jain & Singh, 2002).

Customer segmentation methods using LTV can be classified into three categories: (1) segmentation by using only LTV values, (2) segmentation by using LTV components and (3) segmentation by considering both LTV values and other information.

In the first method, the list of customers' LTV is sorted in descending order. The list is divided by its percentile. In this case, we segment customer list by only LTV, however, other information like socio-demographic information or transaction analysis may be used together for a better marketing practice. For instance, after segmenting a highly profitable customer group, a firm may recommend popular products to the targeted group at a discounted price. Figure 1.4 briefly depicts the concept of segmentation using only LTV.

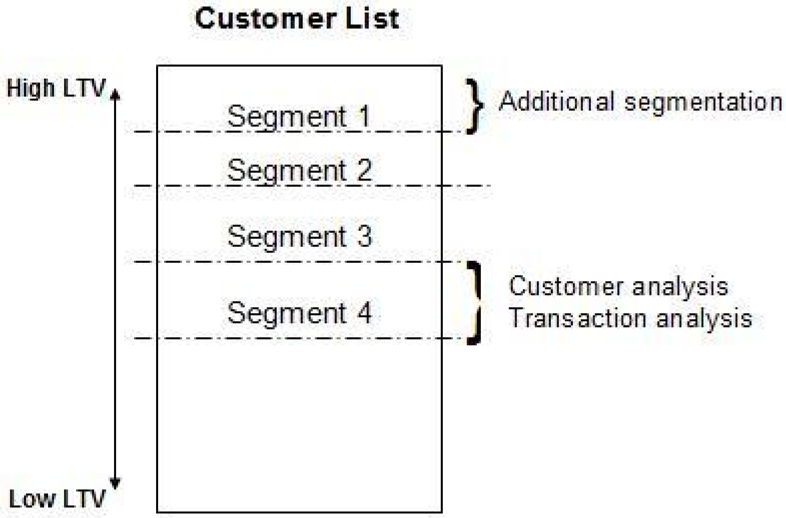


Figure 1.4 : Customer segmentation using LTV (Su-Yeon & Tae-Soo & Eui-Ho & Hyun-Seok, 2006).

The second method performs segmentation by considering components used in LTV calculation. Hwang, Jung, and Suh (2004) considered three factors: current value, potential value, and customer loyalty to calculate LTV and present the method to segment the three factors for customer segmentation. Figure 1.5 shows segmentation using factors in calculating LTV (Su-Yeon & Tae-Soo & Eui-Ho & Hyun-Seok, 2006).

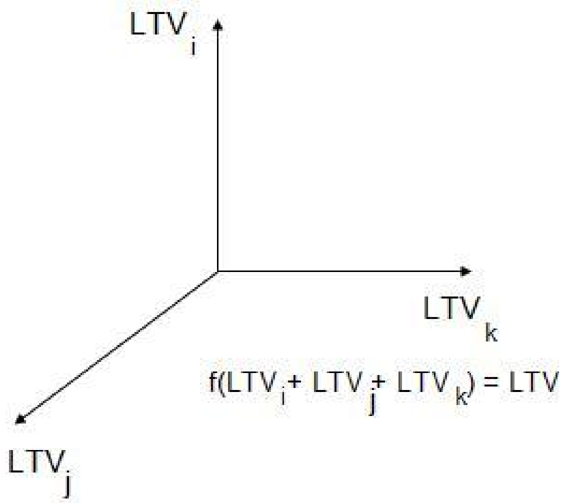


Figure 1.5 : Customer segmentation using LTV components.

The last method is to segment the customer list with LTV value and other managerial information. In this case, LTV is an axis of the segment in n-dimensional segment space and other information, such as socio-demographic information and transaction history become another axis. This approach is more meaningful for segmenting the customer list than the first method. Figure 1.6 shows a segmented customer list with LTV value and other managerial information (Su-Yeon & Tae-Soo & Eui-Ho & Hyun-Seok, 2006).

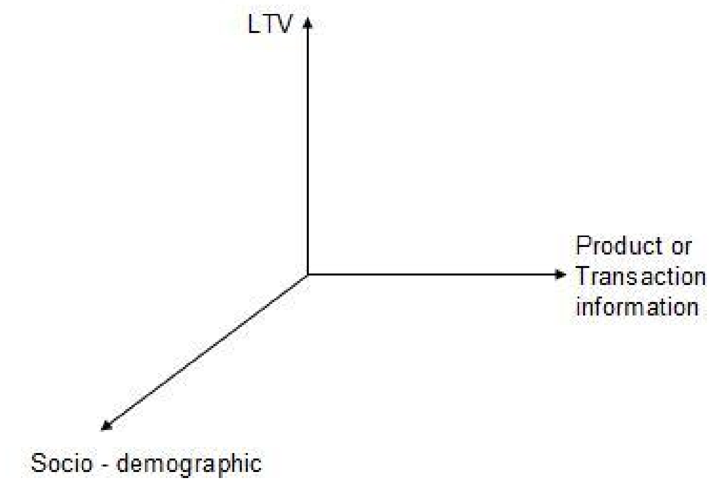


Figure 1.6 : Customer segmentation based on LTV and other information.

1.4.1.2 Calculating customer value

Customer value is calculated on six variables. These variables are defined as follows:

- From revenue to profitability
  - Pure spend-based models: straight forward and easy
  - Introducing cost elements: allocation methods
  - Detailed individual customer costing

- From past to future value
  - Historic value: recent history
  - Historic value: total cumulative history(lifetime)
  - Future (potential) value: different predictive methods
  - Net present value(NPV) : discounted future value

- Lifetime value (LIV) or CLV (Customer lifetime value)

- Predictive techniques
    - Regressions(linear, multiple, logistic)
    - Trees (classification, decision)
    - Advanced techniques(neural Networks, genetic algorithms)

- Definitions of "lifetime"
    - Strictly referring to a future period
    - Need for consistent predictability and actionable/manageable
    - Techniques to calculate LOS (length of service) : survival analysis(LIFEREG, PHREG procedures in SAS)

- Alternative models and techniques
    - The RFM(Recency, frequency, monetary value) model as a segmentation tool
    - Rapid identification techniques ("golden questions")

1.4.1.3 A framework for building managerial strategies based on customer value

Customer segmentation is built on customer value in three phases. The data is prepared to define the customer value and set up marketing strategies in Phase 1. Phase 2 explains the evaluation of the customer value from three ways – current value, potential value and customer loyalty.

After completion of the customer segmentation, a segment analysis is obtained from segmentation results. The current value, the potential value and the customer loyalty are used to analyze the characteristics of each segment in Phase 3. This phase presents these three customer values form the basis of marketing strategies. A conceptual framework of this study is shown in Figure 1.7 We will explain this research according to the framework**.**

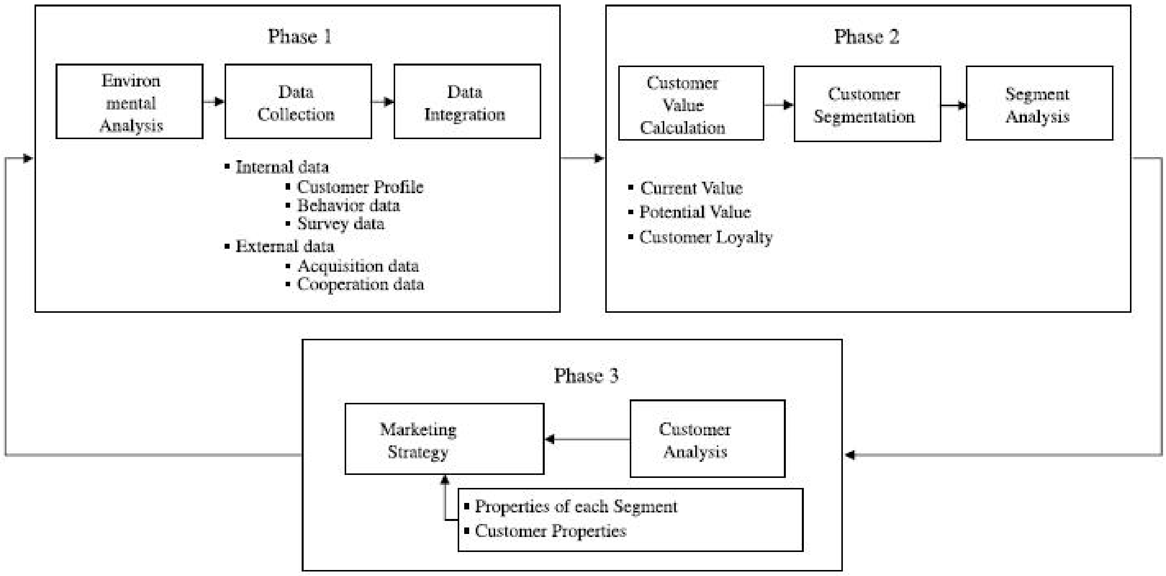


Figure 1.7 : Framework for customer segmentation based on LTV (Micheal & Andreas & Anemon, 2007).

The conceptual structure, is shown in Figure 1.8, presents how it should be a good customer segmentation and customer profiling. The important components of this Figureure, customer segmentation and customer profiling, are the overall objectives of companies in the field of marketing. The account and transaction data sets are used as input sets for customer segmentation. The customer profiles, target audiences, are the output of this conceptual structure. Customer segmentation is affected by RFM values and Repayment Behaviours which are needed for behavioral scoring. Credit card usage and customer attributes are the determining factors on customer profiling.

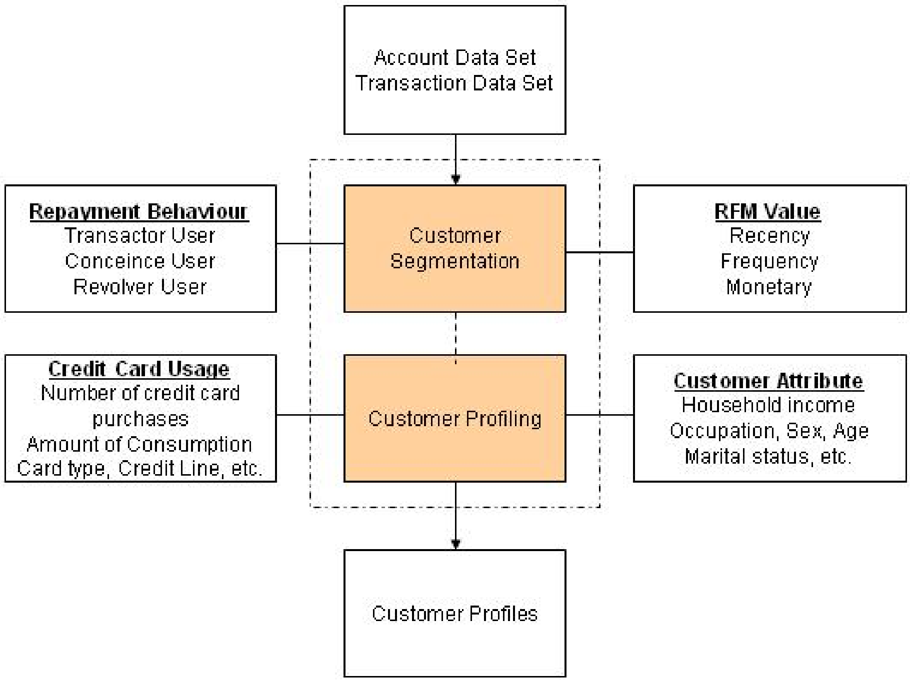


Figure 1.8 : A conceptual framework of customer behavioral modeling (Nan-Chen, 2004).

1.4.1.4 Analyzing the behaviour of customers

To develop relationships with customers and win new customers, banks constantly seek ways of differentiating their offerings and developing more appropriate services for customer demands and expectations in different segments.

Customers are divided into segments according to customer data. Therefore, segmentation analysis is created on past transaction data. The results produced are based on the assumption that the customer behavior follows patterns similar to past patterns and will repeat in the future. Therefore, there could not be a better time than now to recognize the importance of an effective new marketing strategy using data mining techniques. To increase the amount of purchases while improving customer satisfaction is a major goal (Nan-Chen, 2004).

Segmentation analysis is a comprehensive method to communicate better with customers and is the first step for the classification of individual customers. As a result of the segment analysis customer groups are identified in the data, and putting customers into segments according to their affinities or similar characteristics.

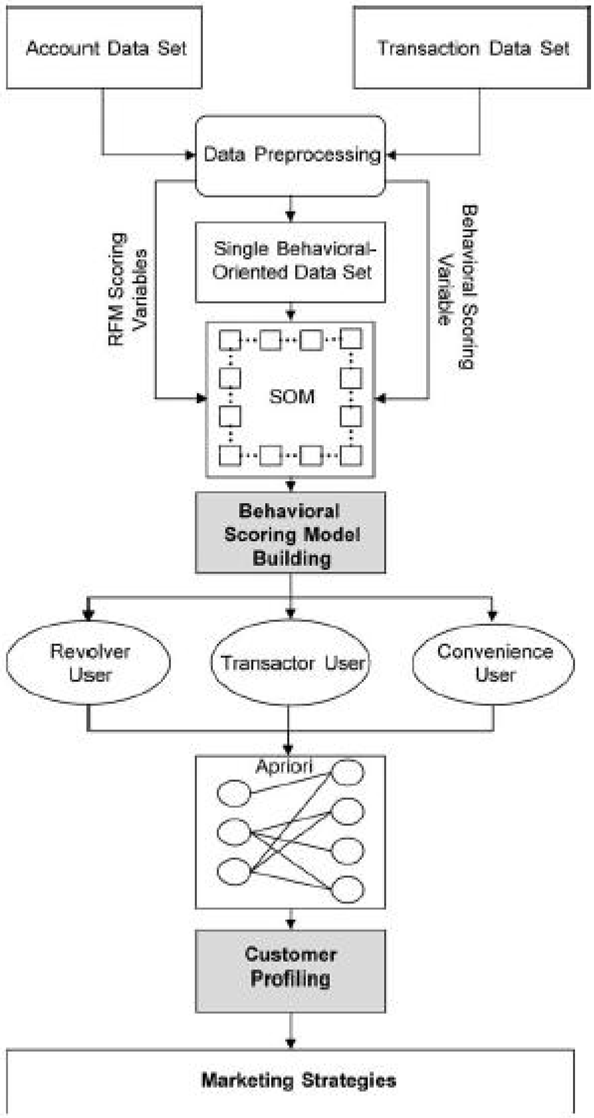


Figure 1.9 : Two-stage behavioral scoring modeling (Nan-Chen, 2004).

1.5 CRISP - DM

Data mining contains a lot of algorithms to explore the contents of data and to show hidden relationships in data. The information is tested with algorithms to reach the best knowledge. The outcome is used to measure the success of the process. There are 4 typical phases of data mining process. These are: Raw Data acquisition, preprocessing, feature extraction and modeling. The major impact on the outcome of the project happens thanks to managing the interactions of these phases.

The combination of the algorithms are developed and used in passing through different phases and the running data through the chain constitutes the traditional approach of implementing the data process, as it is presented in Figure 1.10.

The importance of data processing algorithms seems in Figure 1.10. The information is moved smoothly through the chain from the first step to the last step on a single run, and the algorithms are usually implemented within the same application. It is normal that the data analyst takes care of all the phases, and not much attention is always paid to the storage format of the data. This creates a lot of difficulties that detract from the quality of the data mining process. Therefore, there are many challenges for implementing the approaches a variety of tools to the data, the analyst needs to be expert on the subject, and results in incoherent storage of the research results and data.

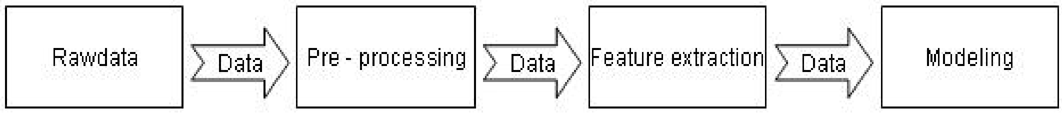


Figure 1.10 : The traditional data mining process.

The methodology of CRISP-DM is shown in a hierarchical process model in Figure 1.11. The structure of this process consists of four different levels. These are, respectively: phase, generic task, specialized task and process instance.

In the first level, the process of data mining has come together in phases. There are a few tasks for each phase. The name of second level is generic, because it is aimed to cover all applications of data mining. These tasks are expected to be as comprehensive and consistent as possible. The meaning of the comprehensive is that containing the entire process of data mining and its all possible practices. The meaning of the consistent is that the process should be applicable for the new modeling techniques will be released in the future.

In the third level, it is shown how the action will be taken in special circumstances. For example, a generic task named clean data might be in the second level. The third level shows how to vary this task in different cases, the examining of numeric and categorical values for arrangement or the selecting of the best algorithm for solving the problem.

The phases of first level and tasks in second and third levels separately are defined and they are shown in a particular directory. In practice, a lot of tasks should be applied in a several order and to apply specific actions will need to return to the previous tasks. There is no need to reach all possible data mining solutions in this process model because this is difficult and complex process for achieving the solution.

The process instance is the last level of data mining process. This level is composed of existing decisions, actions and results of data mining processes. In an organized process instance, tasks which are defined at the higher levels indicate a particular status so there are not used in general tasks.

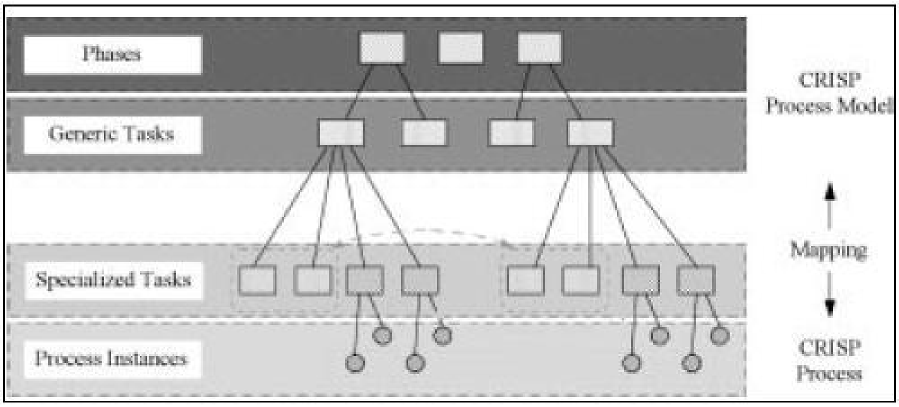


Figure 1.11 : The four level breakdown of the CRISP-DM methodology (Pete & Randy & Julian, Thomas &Thomas & Rüdiger, 1999).

1.5.1 The CRISP-DM reference model

The reference model of CRISP-DM shows the whole life cycle of data mining process. It is composed of the phases of a project. These phases are consisted of tasks and the relations between the tasks. At this description level, all relationships are not likely to find. Essentially, the aim, the experience and willing of the user and the knowing the content of the data are the main elements showing of the relationship between data mining tasks.

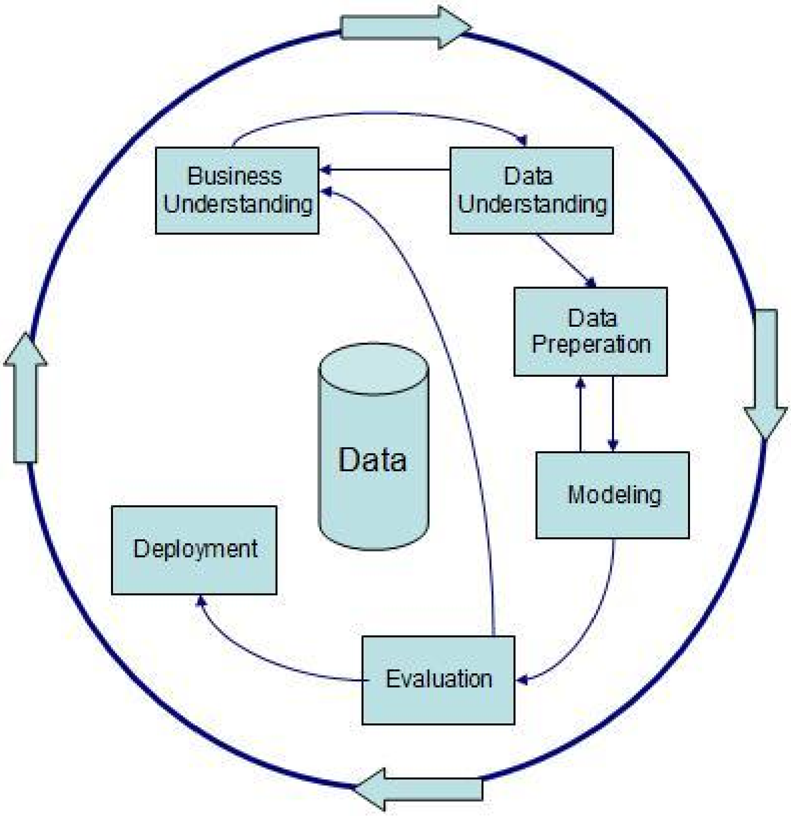


Figure 1.12 : Phases of the CRISP-DM reference model (Pete & Randy & Julian, Thomas &Thomas & Rüdiger, 1999).

The life cycle of a data mining project contains six phases as it is shown in Figure 1.12. There is not a certain sequence for phases. The switching between phases is always required. The tasks in each phase are determined according to results of next stage. The connections between the phases are shown by arrows.

Data mining always controls and improves itself as it is shown in the outer circle of Figure 1.12. Even if a solution were found, the cycle continues. The information acquired in the process and deployed solutions can lead to new questions or may reveal the need of a new solution. Future data mining processes are created based on previous experience.

In the following, each phase is described briefly:

Business understanding

In the first stage, project goals and needs of business are understood. Then, the demand is converted to the data mining problem definition and a road map is created to achieve the objectives.

Data understanding

Understanding the data is started with data collection and data processing processes. Once you understand the data, you can measure the quality of the data, can find extreme values or hidden information in data.

Data preparation

The data preparation phase contains all processes until the data is to be inserted into the modeling. Data preparation process can be applied many times. The main tasks of data preparation phase are data cleaning, selection variables and transformation of the data.

Modeling

In the modeling phase, many different models are chosen and tested. The various models are available for the same data mining problem type. Some of data mining techniques require editing of the data structure. Therefore, it is often necessary to return to the previous stage.

Evaluation

At this stage, a good-quality model was developed for data analysis but it is needed to test the quality of model and check the achievement of business objectives reviewing all the steps of process before the deployment phase. It should be decided on the model to be implemented when the phase is complete.

Deployment

The completion of the model does not mean that the project is complete. The aim of this model not only gains knowledge. At the same time is to obtain a regular data. There are working models in real-time data. These models typically encountered in web pages. The deployment phase may reveal very different results. These results may be preparing a report or may be revising the whole data mining process. In many cases, the model is removed from deployment by customer, not data analyst because differentiation of customer data overrides model.

2. LITERATURE SURVEY

Despite the importance of data mining techniques to customer relationship management (CRM), there is a lack of a comprehensive literature review for it. For this thesis, review of the literature was investigated in Web of Science (ISI Web of Knowledge).

Firstly, CRM and Data Mining are investigated. 214 topics are revealed from research results. These topics are examined in three different points of view: document type, subject area and publication year. The record counts of document type in research are shown in Figure 2.1.

Field: Document Type	Record Count	% of 214	Bar Chart
PROCEEDINGS PAPER	162	75.7009 %	
ARTICLE	51	23.8318 %	
REVIEW	1	0.4673 %	
Field: Document Type	Record Count	% of 214	Bar Chart

Figure 2.1 : The document types of research "CRM" and "Data Mining".

The record counts of subject areas in research are shown in Figure 2.2.

Field: Subject Area	Record Count	% of 214	Bar Chart
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	78	36.4486 %	
COMPUTER SCIENCE, INFORMATION SYSTEMS	76	35.5140 %	
ENGINEERING, ELECTRICAL & ELECTRONIC	54	25.2336 %	
OPERATIONS RESEARCH & MANAGEMENT SCIENCE	48	22.4299 %	
COMPUTER SCIENCE, THEORY & METHODS	33	15.4206 %	
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	31	14.4860 %	
MANAGEMENT	31	14.4860 %	
BUSINESS	27	12.6168 %	
TELECOMMUNICATIONS	27	12.6168 %	
ECONOMICS	16	7.4766 %	
COMPUTER SCIENCE, SOFTWARE ENGINEERING	14	6.5421 %	
ENGINEERING, INDUSTRIAL	14	6.5421 %	
BUSINESS, FINANCE	10	4.6729 %	
COMPUTER SCIENCE, CYBERNETICS	10	4.6729 %	
AUTOMATION & CONTROL SYSTEMS	9	4.2056 %	
INFORMATION SCIENCE & LIBRARY SCIENCE	8	3.7383 %	
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	5	2.3364 %	
ENGINEERING, MULTIDISCIPLINARY	5	2.3364 %	
ENGINEERING, MANUFACTURING	4	1.8692 %	
ENVIRONMENTAL SCIENCES	4	1.8692 %	
CHEMISTRY, ANALYTICAL	3	1.4019 %	
INSTRUMENTS & INSTRUMENTATION	3	1.4019 %	
MATHEMATICS, APPLIED	3	1.4019 %	
ROBOTICS	3	1.4019 %	

Figure 2.2 : The subject areas of research "CRM" and "Data Mining".

The record counts of publication years in research are shown in Figure 2.3.

Field: Publication Year	Record Count	% of 214	Bar Chart
2008	33	15.4206 %	
2009	32	14.9533 %	
2006	31	14.4860 %	
2007	30	14.0187 %	
2004	24	11.2150 %	
2005	19	8.8785 %	
2003	13	6.0748 %	
2002	12	5.6075 %	
2010	8	3.7383 %	
2001	6	2.8037 %	
2000	4	1.8692 %	

Figure 2.3 : The publication years of research "CRM" and "Data Mining".

There have been a lot of studies about using data mining techniques with CRM elements. In this thesis; we focus on customer segmentation is about determining the segment of customer to refer right campaign. There are two approaches to review literature of this study. One of them is research of "Data Mining Techniques" and "Customer Segmentation". 122 topics are revealed from research results. These topics are examined in three different points of view: document type, subject area and publication year. The record counts of document types in shown in Figure 2.4.

Field: Document Type	Record Count	% of 122	Bar Chart
PROCEEDINGS PAPER	66	54.0984 %	
ARTICLE	55	45.0820 %	
REVIEW	1	0.8197 %	

Figure 2.4 : The document types of research "Data Mining Techniques" and "Customer Segmentation".

The record counts of subject areas in research are shown in Figure 2.5.

Field: Subject Area	Record Count	% of 122	Bar Chart
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	58	47.5410 %	
OPERATIONS RESEARCH & MANAGEMENT SCIENCE	42	34.4262 %	
COMPUTER SCIENCE, INFORMATION SYSTEMS	41	33.6066 %	
ENGINEERING, ELECTRICAL & ELECTRONIC	39	31.9672 %	
MANAGEMENT	18	14.7541 %	
COMPUTER SCIENCE, THEORY & METHODS	17	13.9344 %	
BUSINESS	12	9.8361 %	
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	12	9.8361 %	
ENGINEERING, INDUSTRIAL	7	5.7377 %	
COMPUTER SCIENCE, SOFTWARE ENGINEERING	6	4.9180 %	
TELECOMMUNICATIONS	6	4.9180 %	
COMPUTER SCIENCE, CYBERNETICS	5	4.0984 %	
ECONOMICS	5	4.0984 %	
INFORMATION SCIENCE & LIBRARY SCIENCE	5	4.0984 %	
ENGINEERING, MULTIDISCIPLINARY	4	3.2787 %	
MATHEMATICS, APPLIED	4	3.2787 %	
AUTOMATION & CONTROL SYSTEMS	3	2.4590 %	
ENGINEERING, MANUFACTURING	3	2.4590 %	
BUSINESS, FINANCE	2	1.6393 %	
STATISTICS & PROBABILITY	2	1.6393 %	

Figure 2.5 : The subject areas of research "Data Mining Techniques" and "Customer Segmentation".

The record counts of publication years in research are shown in Figure 2.6.

Field: Publication Year	Record Count	% of 122	Bar Chart
2009	30	24.5902 %	
2007	17	13.9344 %	
2008	15	12.2951 %	
2005	13	10.6557 %	
2006	13	10.6557 %	
2010	8	6.5574 %	
2004	7	5.7377 %	
2002	5	4.0984 %	
2003	5	4.0984 %	
1999	3	2.4590 %	
2000	2	1.6393 %	
2001	2	1.6393 %	
2011	2	1.6393 %	

Figure 2.6 : The publication years of research "Data Mining Techniques" and "Customer Segmentation".

In this research, data mining techniques which are used for customer segmentation have been identified. Decision Trees, Regression, K-means and rough-set data mining techniques approaches are the most widely used of them. In this thesis, these techniques are applied to the data.

Another approach of these studies is research of "Data Mining", "Customer Segmentation" and "Fuzzy C-means and Rough Sets". In this research, all topics are related our thesis. Therefore, all of them are examined one by one. 32 topics are revealed from research results. These topics are shown in Table 2.1.

Table 2.1 : The topics of research "Data Mining", "Customer Segmentation" and "Fuzzy C-Means and Rough Sets".

Title	Author(s)	Published
Customer Segmentation Architecture Based on Clustering Techniques	Lefait G, Kechadi T	2010
Customer Segmentation based on a Novel Hierarchical Clustering Algorithm	Cao SQ, Zhu QY, Hou ZW	2009
Channels of Contact with Revenue: Is the Telephone Irreplaceable?	Manai G, Dwyer M, Cleary D	2009
Gaining Insight to Customer Behavior based on Time Series Data	Wei-Ping X	2009
Identifying target green 3C customers in Taiwan using multiattribute utility theory	Wang ML, Kuo TC, Liu JW	2009
Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn	Bose I, Chen X	2009
Chameleon based on clustering feature tree and its application in customer segmentation	Li JF, Wang KL, Xu LD	2009
A case study of applying data mining techniques in an outfitter's customer value analysis	Huang SC, Chang EC, Wu HH	2009
Knowledge creation in marketing based on data mining	Zhang GZ, Zhou FM, Wang F, et al.	2008
Combining several SOM approaches in data mining: Application to ADSL customer behaviours analysis	Fessant F, Lemaire V, Clerot F	2008
Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques	Coussement K, Van den Poel D	2008
Web personalisation through incremental individual profiling and support-based user segmentation	Hofgesang PI	2007
Customer segmentation using overlapping cluster algorithm	Qian F	2007
Applying knowledge engineering techniques to customer analysis in the service industry	Ha SH	2007
Characterizing ADSL customer behaviours by network traffic data-mining	Fessant F, Francois J, Clerot F	2007
Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression	McCarty JA, Hastak M	2007
Mining of mixed data with application to catalog marketing	Hsu CC, Chen YC	2007
A purchasing sequences data mining method for customer segmentation	Wang H, Wang SH	2006

Table 2.1 : The topics of research "Data Mining", "Customer Segmentation" and "Fuzzy C-Means and Rough Sets" (continued).

Title	Author(s)	Published
Customer intelligence system based on improving LTV model and data mining	Chen YZ, Zhao MH, Zhao SL, et al.	2006
Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps	Verdu SV, Garcia MO, Senabre C, et al.	2006
Business performance management system for CRM and sales execution	Ettl M, Zadrozny B, Chowdhary P, et al.	2005
Predicting customer retention and profitability by using random forests and regression forests techniques	Lariviere B, Van den Poel D	2005
Temporal analysis of clusters of supermarket customers: conventional versus interval set approach	Lingras P, Hogo M, Snorek M, et al.	2005
Visualization method for customer targeting using customer map	Woo JY, Bae SM, Park SC	2005
A methodology for dynamic data mining based on fuzzy clustering	Crespo F, Weber R	2005
Visualization techniques and automation of cluster analysis with self-organizing maps for decision support in customer segmentation	Richartz J	2004
A human-computer interactive method for projected clustering	Aggarwal CC	2004
Something approaching science? Cluster analysis procedures in the CRM era	Nairn A, Bottomley P	2003
Combining data mining and optimization for campaign management	Vercellis C	2002
Redefining clustering for high-dimensional applications	Aggarwal CC, Yu PS	2002
CRM in a real-world insurance company	Pedrazzi G, Turra R, Zanasi A	2000
Data mining for database marketing at Garanti Bank	Alis OF, Karakurt E, Melli P	2000

3. MATERIALS & METHODS

3.1 MATERIALS

3.1.1 Program

In this study, the programs are used:

- SPSS Clementine Version 12: SPPS Clementine, data mining techniques, with remarkable achievements in many areas has been signed. Also it is leader in the Gartner Magic Quadrant report of 2008 for data mining solution.

- KNIME (Konstanz Information Miner) Version 2.3.0: KNIME is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform.

- MS SQL Server 2005: It is a relational model database server produced by Microsoft.

3.1.2 Data

In this thesis, data which are used are taken from a bank. Therefore, data which identifies customer are masked. There are only demographic and operational data of customers.

There are two data sets for 50.000 customers:

- Demographic data set,
- Operational data set – 5 monthly operational information.

3.1.2.1 Data preparation

The all of data which are in the separate resources are collected in one source by SPSS. The data of customer are converted to rawdata format according to Customer_ID reference in database.

There are two data sets: operational and demographic data sets in this study. The operational data sets have five separate monthly data. These two data sets have been brought together in four steps. It is shown in Figure 3.1.

First of all, the operational data set is separated five different data sets according to month. “Year” and “Month” are filtered from these five data sets.

In the second step, new data are derived from demographic data set:

- “Age” is derived from birth date.
- “Year of Customer” is derived from initial date.

In the third step, five different operational and demographic data sets are merged in rawdata format according to Customer_ID.

In the last step, new data set which is occurred in third step is written to the database.

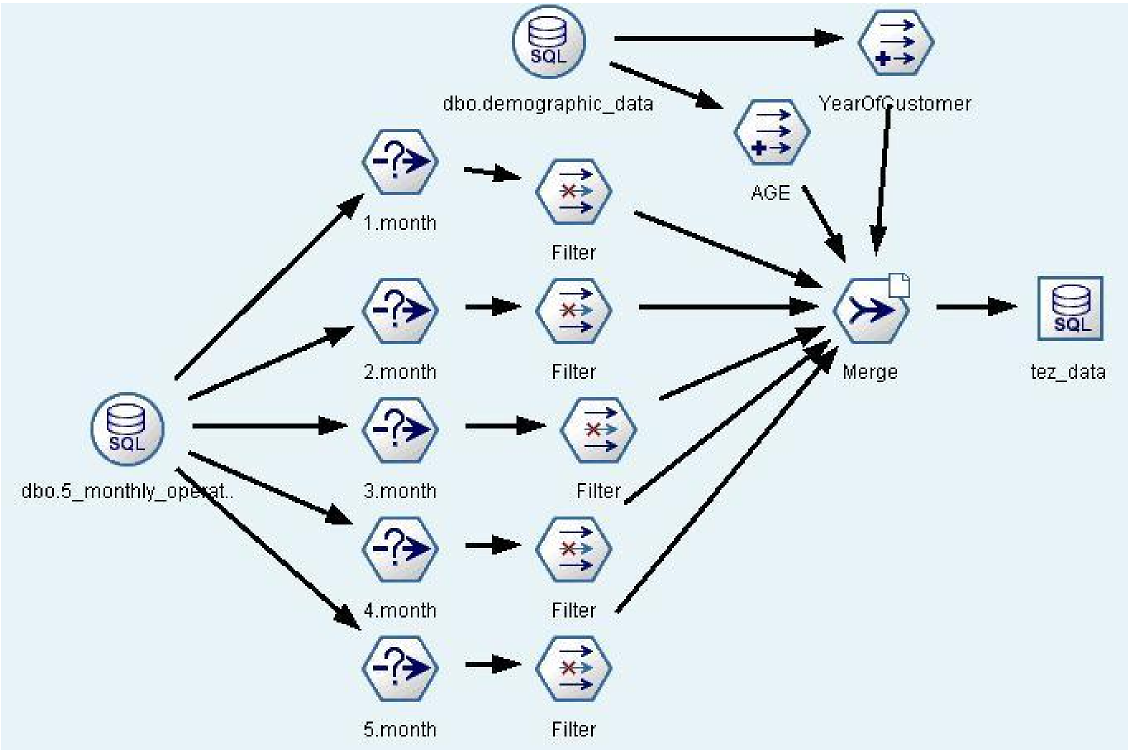


Figure 3.1 : Preparing Data.

The all data belong to customer is prepared in a single raw data in database for every customer. Thus, analysis is obtained will be easier and faster.

The variables of customer are collected under two groups at the end of the data preparation phase. First group is demographic data: these data contain personal information such as birth date, job, marital Status, educational background and gender. Some of these data are relevant both the bank and the customer such as customer type, branch office, city and initial date. The second group is operational data which are written such "_1", "_2", "_3", "_4" and "_5" at the end of the parameters as it is shown in Table 3.2. The parameter names containing "_1", "_2", "_3", "_4" and "_5" show 5-monthly data and these numbers indicate the month of data.

The all variables without "_2", "_3", "_4" and "_5" monthly data in data are shown in Table 3.1. The all variables in data are shown in appendix A.

Table 3.1 : The variables of data.

Data Name	Value Type	Comment
Must_No	Int	Customer ID
Müşteri Tipi	Nvarchar(1)	Customer Type
Doğum Tarihi	DateTime	Birth Date
Muta Şb	Varchar(4)	Branch Offiec of Bank Customer
Meslek Kod	Small Int	Job Code
meslek	Nvarchar(255)	Job
KamuOzel	Varchar(1)	Work Place Information - (Is it state or private?)
Medeni Durum	Varchar(2)	Marital Status
Eğitim Durum	Nvarchar(1)	Educational Background
Cinsiyet	Nvarchar(1)	Gender
İlk tanımlama tar	DateTime	Initial Date (for customer)
İLİ	Nvarchar(50)	City
MİY_SICIL_NO	Varchar(5)	Customer Relations Officer ID
ABONE24_TUT_1	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_1	Float	The Amount of Gold Account
ALTIN_FLAG_1	Bit	The Flag of Gold Account
BK_FLAG_1	Bit	The Falg of Personal Loan
BK_ORT_TUT_1	Float	The Average Monthly Amount of Personal Loans
BORDRO24_FLAG_1	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_1	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_1	Bit	The Flag of Debit Card
DIALOG_ISL_TUT_1	Float	The Amount of Transactions on The Call Center
DTH_FLAG_1	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_1	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_1	Int	The Number of Loan Delay
INTRNT_ISL_TUT_1	Float	The Amount of Transactions on The Internet
KK_FLAG_1	Bit	The Flag of Credit Card
KK_LIMIT_1	Float	The Limit of Credit Card
KKB_BNK_LMT_1	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_1	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_1	Int	The Credit Score of Customer
KMH_FLAG_1	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_1	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_1	Float	Cumulated Net Profit
MENKUL_ORT_TUT_1	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_1	Bit	The Flag of Investment Account
NAR_BKY_TUT_1	Float	The Amount of Nar
NAR_FLAG_1	Bit	The Flag of Nar which is a kind of Product for Retail Marketing
OTMTK_TLMT_FLAG_1	Bit	The Flag of Direct Debit
SGK_FLAG_1	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_1	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_1	Float	The Amount of Insurance Commission
SIGORTA_FLAG_1	Bit	The Flag of Insurance

Table 3.1 : The variables of data (continued).

Data Name	Value Type	Comment
TAKIP_FLAG_1	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_1	Float	The Amount of Loan Installment
TP_CALIS_TUT_1	Float	The Total Amount of Work
TP_URUN_ADET_1	Int	The Number of Total Product
TP_VARLIK_1	Float	Total Assets
TP_VDL_TUT_1	Float	Total amount of Term Deposits
VDL_ORT_TUT_1	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_1	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_1	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_1	Bit	The Flag of Demand Deposit Account

3.1.2.2 Data understanding

Data Sources

Data sources are built on 50.000 customers. There are two data sets: Demographic information and Operational information. Operational information is hold monthly and 5 monthly data are used in this application.

Demographic Input Data

- Customer Type,
- Birth Date,
- Branch Office,
- Profession,
- Work Place Information,
- Marital Status,
- Educational Background,
- Gender,
- Getting Customer Date,
- City.

Operational Input Data

- Sight Deposit Account
- Deposit Account
- Gold Deposit Account
- Foreign Currency Account
- Credit-Deposit Account
- Order Account
- Credit Card
- Investment Account
- Insurance
- Personal Loan
- Salary
- Internet Branch Office
- Call Center
- Cumulative Net Earnings

There are 224 fields as input data in this application. All data can be analyzed easily by Data Audit Module of SPSS.

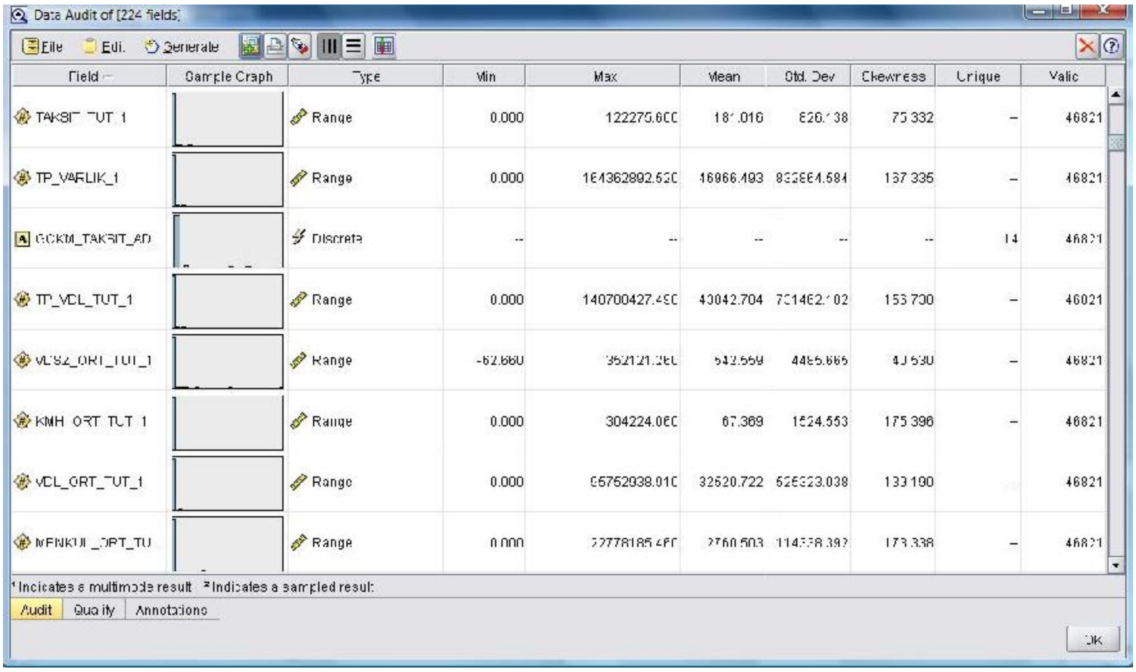


Figure 3.2 : Data Audit.

The data analysis will be examined in two ways according to type of variables. First of them is the analysis of range type. The minimum, maximum and mean ranges of these variables and standard deviation of these variables without "_1", "_2", "_3" and "_4" monthly data are shown in Table 3.2. The analysis of all variables is shown in appendix A.

Table 3.2 : The analysis of data

Parameters	Min	Max	Mean	Std. Dev
Birth Date	03.02.1900 00:00	17.11.2004 00:00	--	--
Initial Date	01.01.1900 00:00	27.11.2007 00:00	--	--
Age	6	110	52.410	11.312
Year of The Customer	3	110	14.268	6.963
TAKSIT_TUT_5	0.000	108.577.890	188.994	796.915
TP_VARLIK_5	0.000	191.569.794.180	44.314.144	908.813.941
TP_VDL_TUT_5	0.000	167.644.576.830	41.347.289	806.155.750
VDSZ_ORT_TUT_5	-586.390	410.666.410	605.418	4.798.216
KMH_ORT_TUT_5	0.000	190.563.470	67.850	1.101.719
VDL_ORT_TUT_5	0.000	100.521.671.100	30.664.054	519.667.665
MENKUL_ORT_TUT_5	0.000	22.538.401.410	2.559.275	109.403.719
BK_ORT_TUT_5	0.000	891.700.040	4.541.385	18.687.654
KUMULE_NET_KAZANC_5	-564.150	28.804.810	66.039	283.923
DTH_ORT_TUT_5	0.000	66.792.062.000	11.367.794	320.514.383
SGRT_KOM_TUT_5	0.000	2.583.330	13.741	47.819
INTRNT_ISL_TUT_5	0.000	29.100.622.500	1.968.100	132.453.211
DIALOG_ISL_TUT_5	0.000	204.400.100	49.054	1.386.483
BORDRO24_MAAS_TUT_5	0.000	201.355.980	200.278	1.254.543
SGK_MAAS_TUT_5	0.000	77.698.960	200.519	1.291.657
ABONE24_TUT_5	0.000	9.468.200	44.428	190.282
TP_CALIS_TUT_5	-3.277.010	189.854.857.420	47.932.465	898.619.890
ALTIN_BKY_TUT_5	0.000	412.883.730	226.535	4.443.716
NAR_BKY_TUT_5	0.000	260.563.260	54.155	1.718.943

Another way for the data analysis is the analysis of categorical variables. All the values of these parameters are included in is shown in following tables.

The values of Customer Type are shown in Table 3.3

Table 3.3 : The Customer Type Values

Values	Meaning
B	**B**ireysel - Individual
Ö	**Ö**zel - Private

The values of State_Private are shown in Table 3.4.

Table 3.4 : The State_Private Values

Values	Meaning
B	**B**ilinmiyor - Unknown
K	**K**amu - State
Ö	**Ö**zel - Private

The values of Gender are shown in Table 3.5.

Table 3.5 : The Gender Values

Values	Meaning
B	**B**ilinmiyor - Unknown
E	**E**rkek - Male
K	**K**adın - Female

The values of Marital Status are shown in Table 3.6.

Table 3.6 : The Marital Status Values

Values	Meaning
BE	**BE**kar - Single
BI	**BI**linmiyor - Unknown
DI	Dul - Widow
EV	**EV**li - Married

The values of Educational Status are shown in Table 3.7.

Table 3.7 : The Educational Status Values

Values	Meaning
B	**B**ilinmiyor - Unknown
D	**D**oktora - Doctorate
İ	**İ**lkokul - Primary School
L	**L**ise - High School
M	**M**aster - Graduate Student
O	**O**rtakokul - Middle School
Ü	**Ü**niversite - University

The values of City are shown in Table 3.8.

Table 3.8 : The City Values

Adana	Bilecik	Erzurum	Karaman	Mersin	Tekirdağ
Adıyaman	Bingöl	Eskişehir	Kars	Muğla	Tokat
Afyon	Bitlis	Gaziantep	Kastamonu	Muş	Trabzon
Ağrı	Bolu	Gazimağusa	Kayseri	Nevşehir	Tunceli
Aksaray	Burdur	Giresun	Kilis	Niğde	Uşak
Amasya	Bursa	Girne	Kırıkkale	Ordu	Van
Ankara	Çanakkale	Gümüşhane	Kırklareli	Osmaniye	Yalova
Antalya	Çankırı	Hakkâri	Kırşehir	Rize	Yozgat
Ardahan	Çorum	Hatay	Kocaeli	Sakarya	Zonguldak.
Artvin	Denizli	Iğdır	Konya	Samsun	
Aydın	Diyarbakır	Isparta	Kütahya	Siirt	
Balıkesir	Düzce	İstanbul	Lefkoşa	Sinop	
Bartın	Edirne	İzmir	Malatya	Sivas	
Batman	Elazığ	Kahramanmaraş	Manisa	Şanlıurfa	
Bayburt	Erzincan	Karabük	Mardin	Şırnak	

The values of Profession are shown in Table 3.9.

Table 3.9 : The Profession Values

Akademisyen / Üniversite Öğretim Görevlisi	Hatalı Girilmiş	Pilastik Sanatlar (Ressam Heykeltıraş … vb.)
Analist / Programcı	Hemşire / Ebe	Polis / Güvenlik Görevlisi
Astsubay	Hizmet / Ticaret	Reklam / Halkla ilişkiler
Aşçı / Garson / Barmen	Hostes	Sağlık Personeli
Avukat	İmalat / Sanayi	Sahne Sanatları (Bale Tiyatro … vb.)
Büyük Sanayici	İnsan Kaynakları	Sekreter / Yönetici Asistanı
Çalışmayan	İşçi	Serbest Meslek / Esnaf (Bakkal Nalbur … vb.)
Çiftçi / Balıkçı	Küçük Sanayici	Serbest Meslek / Zanaatkâr (Terzi … vb.)
Dealer / Broker	Manken / Fotomodel	Sosyal ve İdari Bilimci
Diğer	Memur	Sporcu / Antrenör
Diplomat	Mevsimlik İşçi	Subay (Teğmen Yüzbaşı Binbaşı Albay … vb.)
Diş Hekimi	Mimar / İçMimar / Dekoratör	Şöför
Eczacı	Muhasebeci / Mali Müşavir	Teknisyen
Emekli	Mühendis	Temel Bilimci (Fizikçi Kimyager … vb.)
Ev Hanımı	Noter	Tercüman / Çevirmen
Finans	Öğrenci	Tıp Doktoru
Fotoğrafçı	Öğretmen / Eğitmen	Veteriner
Gazeteci (Basın)	Pazarlama / Satış	Yazar / Şair
Hâkim / Savcı	Pilot / Kaptan	Yerel Yönetim.

The examining of the distribution of categorical variables is necessary to better analyze the data. Distribution of the categorical variables is as follows:

- Distribution of the age is shown below.

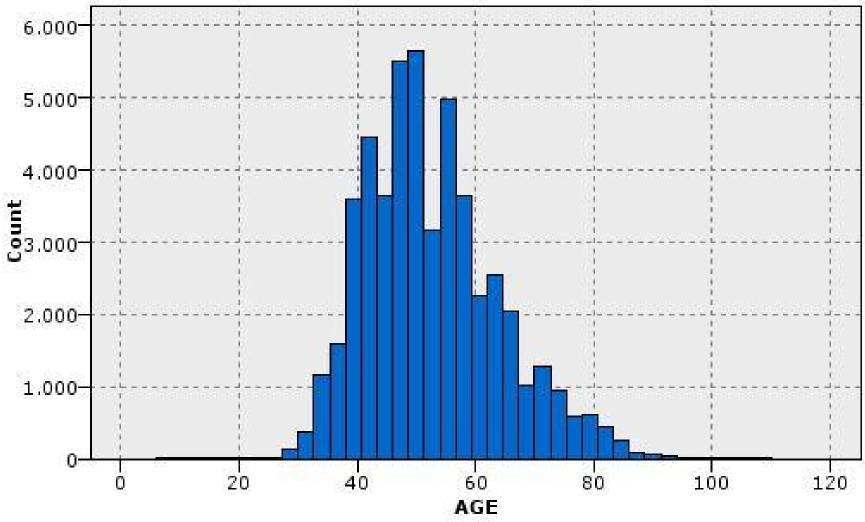


Figure 3.3 : Age

- Distribution of the profession is shown below.

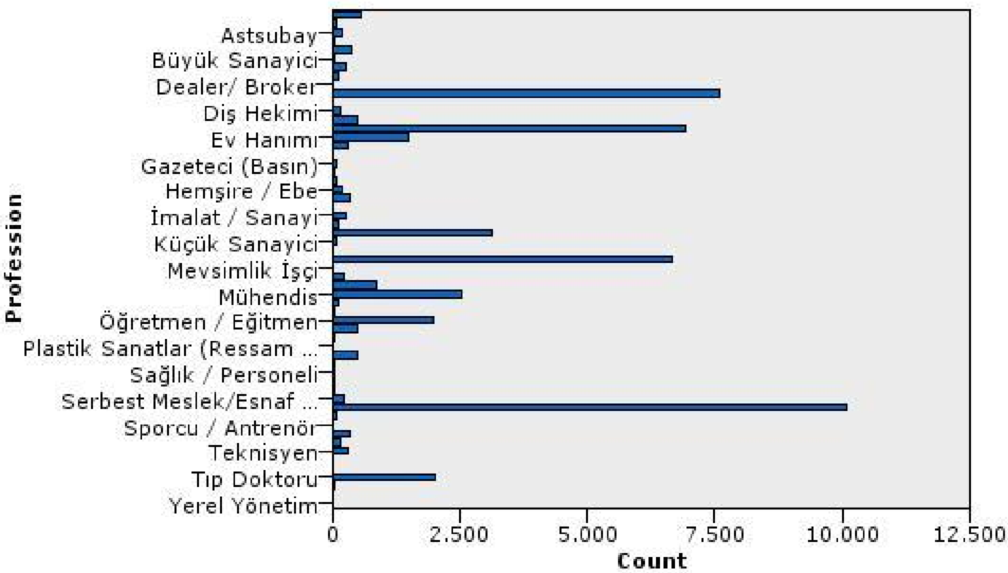


Figure 3.4 : Profession

- Distribution of the educational status is shown below.

    - “B” represents the unknown,
    - “D” represents the doctorate,
    - “İ” represent the primary school,
    - “L” represents the high school,
    - “M” represents the postgraduate,
    - “O” represents the middle school,
    - “Ü” represents the graduate.

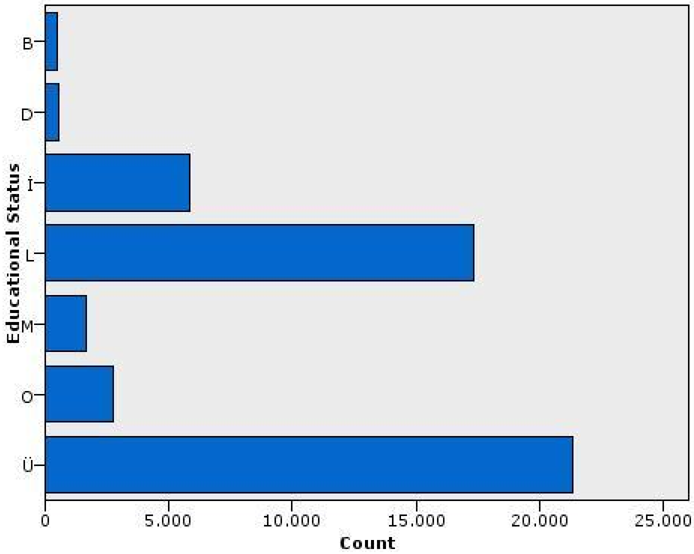


Figure 3.5 : Educational Status

- Distribution of the gender is shown below.

    - “B” represents the unknown,
    - “E” represents the male,
    - “K” represents the female.

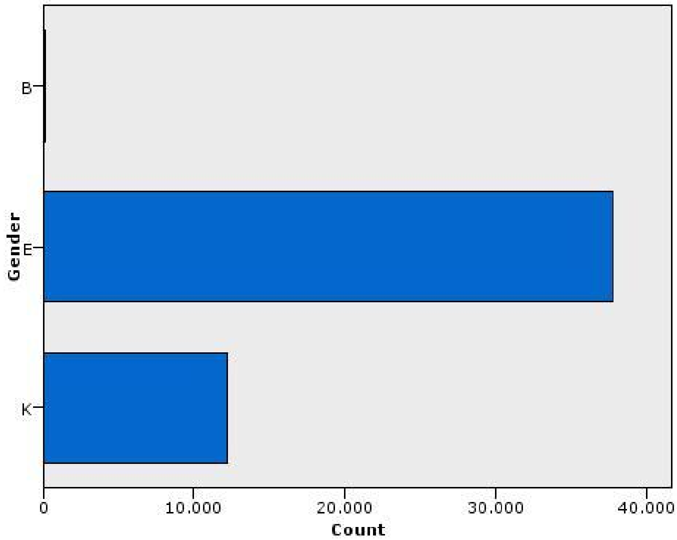


Figure 3.6 : Gender

- Distribution of the city is shown below.

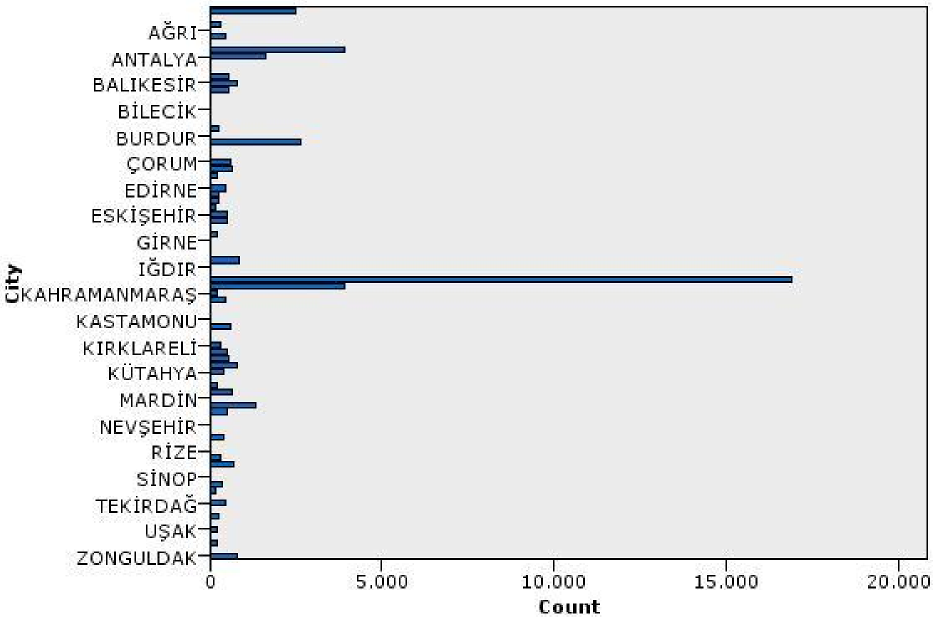


Figure 3.7 : City

- Distribution of the marital status is shown below.

    - "BE" represents the single,
    - "BI" represents the unknown,
    - "DI" represent the widow,
    - "EV" represents the married.

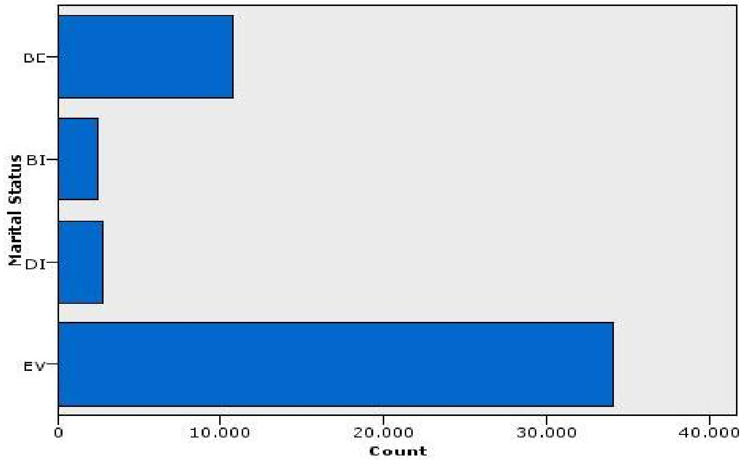


Figure 3.8 : Marital Status

3.2 METHODS

3.2.1 Fuzzy Sets Theory and C-means Algorithm

Theory of fuzzy sets examines the classical set theory for modeling vagueness and uncertainty. A fundamental question concerning fuzzy sets theory is its connections and differences. There have been many studies on this topic.

The purpose of the clustering is to separate N patterns to the c clusters with a high similarity in the cluster and low similarities between clusters while achieving an optimized function. An unsupervised learning occurs in when clusters in the data set are not separated in this way. In the partitive approach it should be provided to obtain the desired number of cluster c. All clusters are represented by their center of gravity in traditional c-means algorithm.

The conventional c-means algorithm proceeds by partitioning N objects x_k into c non-empty subsets. During each partition, the centroids or means of clusters are computed as

$$v_i = \frac{\sum_{x_k \in u_i} x_k}{|c_i|} \tag{3.1}$$

where $|c_i|$ is the number of objects in cluster U_i. The process is repeated until convergence, i.e., there are no more new assignments of objects to the clusters.

The fuzzy sets have been integrated into the c-means framework to achieve the fuzzy c-means (FCM) algorithm. FCM provides the efficient use of overlapping parts.

3.2.1.1 Fuzzy C-Means

FCM clustering algorithm, created by Dunn (1973) and later developed by Bezdek (1981), is an unsupervised clustering algorithm with multiple applications, ranging from feature analysis, to clustering and classifier design. FCM clustering techniques have unusual differences with hard clustering algorithms such as the K-means scheme in that a single data object may be mapped simultaneously to multiple clusters rather than being assigned exclusively to a single cluster. FCM clustering algorithm consists of two main titles, calculating the cluster centroids and assigning the data points to these centroids on the basis of their Euclidean distance, and determining the cluster memberships of each sample point. This is an extension of the c-means algorithm, as proposed by Bezdek (J.C. Bezdek, 1981), in the sense that we allow for partial membership of patterns to clusters. It partitions a set of N patterns (x_k) into c clusters by minimizing the objective function

$$J = \sum_{k=1}^{N} \sum_{i=1}^{c} (u_{ik})^m \left\| x_k - v_i \right\|^2, \tag{3.2}$$

where m>1 is the fuzzifier, $u_{ik} \in [0;1]$ is the membership of the k the pattern to cluster center v_i, and $\|.\|$ is the Euclide an distance, such that

$$v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m} \tag{3.3}$$

and

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/m-1}} \tag{3.4}$$

$\forall i$, with $d_{ik} = \|x_k - v_i\|^2$, subject to $\sum_{i=1}^{c} u_{ik} = 1$, $\forall k$, and $0 < \sum_{k=1}^{N} u_{ik} < N$. The object assignment and mean computation are repeated until $|u_{ik}(t) - u_{ik}(t-1)| < \varepsilon$, at iteration t. Note that for $u_{ik} \in \{0,1\}$ the objective function of Eq.(2) boils down to the hard c-means case, whereby a winner-take-all strategy is applied in place of membership values in Eq.(3).

4. FINDINGS

In this paper, all findings which are formed by using CRISP-DM methodology are determined for a bank. All steps of the CRISP-DM methodology are following step-by-step in hierarchical structure.

Here, findings will be obtained under the two headings. The first heading is clustering phase which is a process that creates the proper groups. The clustering phase is the basis of the second phase. The second heading is prediction phase which is the process of customer segmentation studies.

4.1 CLUSTERING PHASE

In this phase, all the product variables are used as inputs for clustering algorithms. According to these variables, we will try to determine the proper product groups. The two algorithms are chosen: k-means and fuzzy c-means algorithms. This two clustering algorithms are compared to achieve the best result for this phase. We made such a study to reach this result as it is shown Figure 4.1. The number of clusters which are chosen for these models is four.

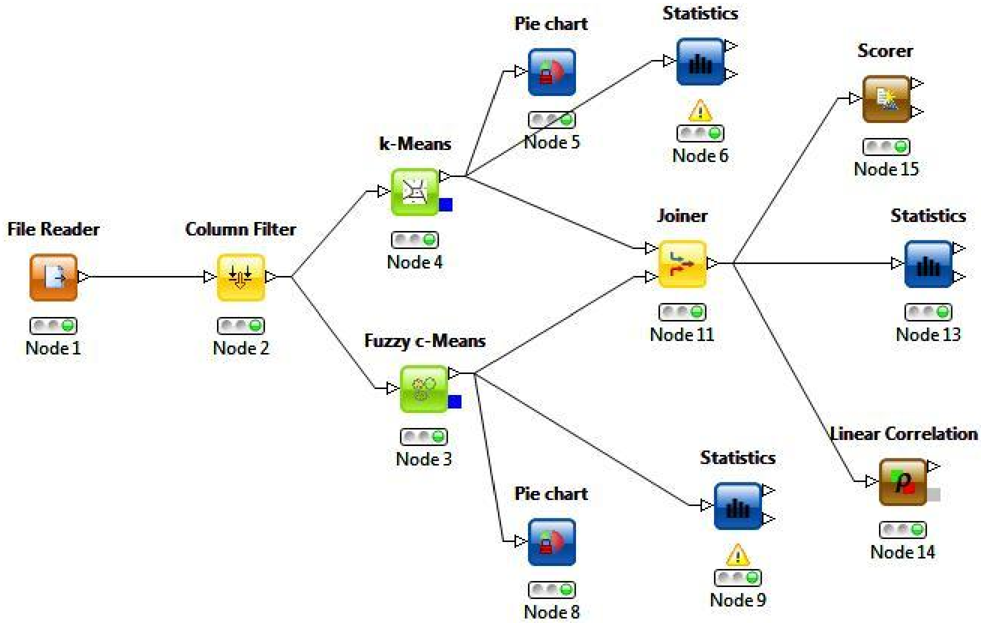


Figure 4.1 : The Comparison of K-Means and Fuzzy C-Means Algorithms.

As a result of clustering algorithms, clusters are as follows. "Cluster" indicates the results of k-means algorithm. "Winner Cluster" indicates the results of fuzzy c-means algorithm.

Row ID	S Cluster	I Cluster_Count	S Winner Cluster	I Winner Cluster_Count
Row0	cluster_3	14490	cluster_2	23187
Row1	cluster_1	13910	cluster_0	13271
Row2	cluster_2	11014	cluster_1	7199
Row3	cluster_0	10586	cluster_3	6343

Figure 4.2 : The Results of the Clustering Algorithms.

The accuracy statistics of the results of two algorithms are shown in Figure 4.3.

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives
cluster_2	9277	1737	25076	13910
cluster_0	4243	6343	30386	9028
cluster_1	0	13910	28891	7199
cluster_3	0	14490	29167	6343

Figure 4.3 : The Accuracy Statistics of K-Means and Fuzzy C-Means Algorithms.

The linear correlation of the results of two algorithms are shown in Figure 4.4.

Row ID	D Cluster	D Winner Cluster
Cluster	1	0.686
Winner Cluster	0.686	1

Figure 4.4 : The linear correlation between the results of the clustering algorithms.

We know that Fuzzy C-Means algorithm is better than K-Means algorithm. It is also shown in this study. The distribution of Fuzzy C-Means clusters is shown in Figure 4.5.

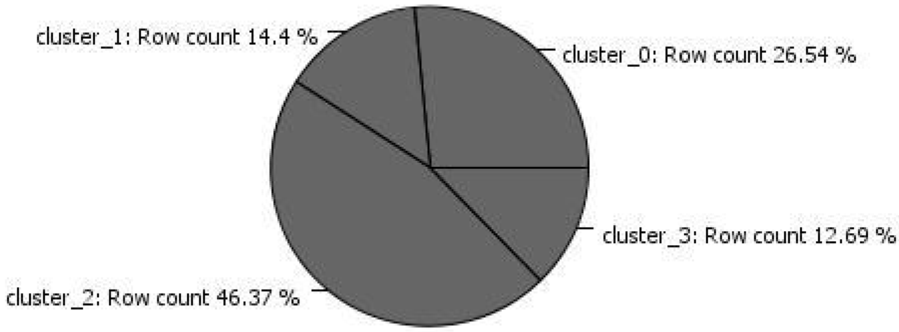


Figure 4.5 : The distribution of Fuzzy C-Means Clusters.

Minimum, maximum, mean and standard deviation values for all clusters of Fuzzy C-Means are summarized in Figure 4.6.

Row ID	D cluster_0	D cluster_1	D cluster_2	D cluster_3
Minimum	0.018	0.042	0.026	0.007
Maximum	0.677	0.767	0.896	0.817
Mean	0.206	0.239	0.393	0.162
Std. deviation	0.206	0.187	0.336	0.24

Figure 4.6 : The Statistics Table of Fuzzy C-Means.

4.2 PREDICTION PHASE

Customer segmentation is needed to easily control and manage subgroups which are formed by segmentation. There are several different types of customer segmentation and all have different procedures. Some of them are as follows:

- Demographic Segmentation,
- Value Segmentation,
- Behavioral Segmentation,
- Psychographic Segmentation,
- Geographic Segmentation, etc.

In this study, demographic and value segmentation approaches will be applied to data.

4.2.1 Demographic Segmentation

Demographic segmentation consists of dividing the market into groups based on variables such as age, gender, family size, income, occupation, education, religion, race and nationality. Some of them; Age, Marital Status, Educational Status, Gender and State_Private variables are in our data set. We will be implementing the demographic segmentation with these variables.

The cumulative net earning variable shows the profitability of customer from the beginning of the year to that month. Therefore, this variable is our target. Firstly, all demographic inputs are put in the C&R Tree, Neural Net, GenLin and CHAID Models. C&R Tree constructs a predictive decision tree using the classification and regression algorithm. Neural Net model constructs a predictive neural net model. GenLin constructs a predictive generalized linear model. CHAID constructs a predictive decision tree using the CHAID algorithm. All of these models try to examine the impacts of demographic variables on cumulative net earning. It is shown in Figure 4.7.

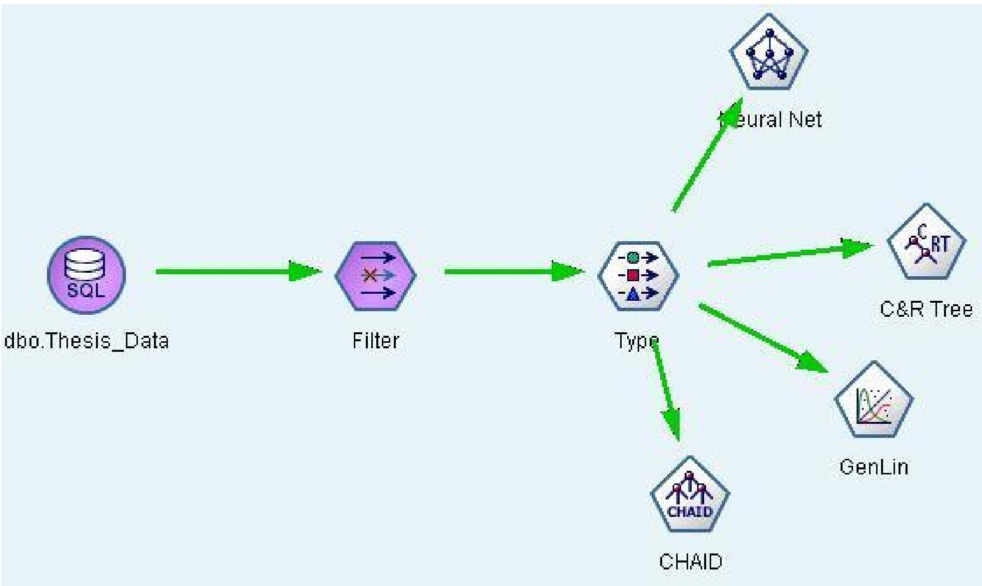


Figure 4.7 : C&R Tree, Neural Net, GenLin and CHAID models.

The results of these models emerge four different table for importance of input variables on target – cumulative net earning.

In the first table, the variable importance table of C&R Tree model is shown Figure 4.8. Age is the most important variable on target. The following important variables are State_Private, Educational Status and Gender, respectively. There is very little difference between these variables.

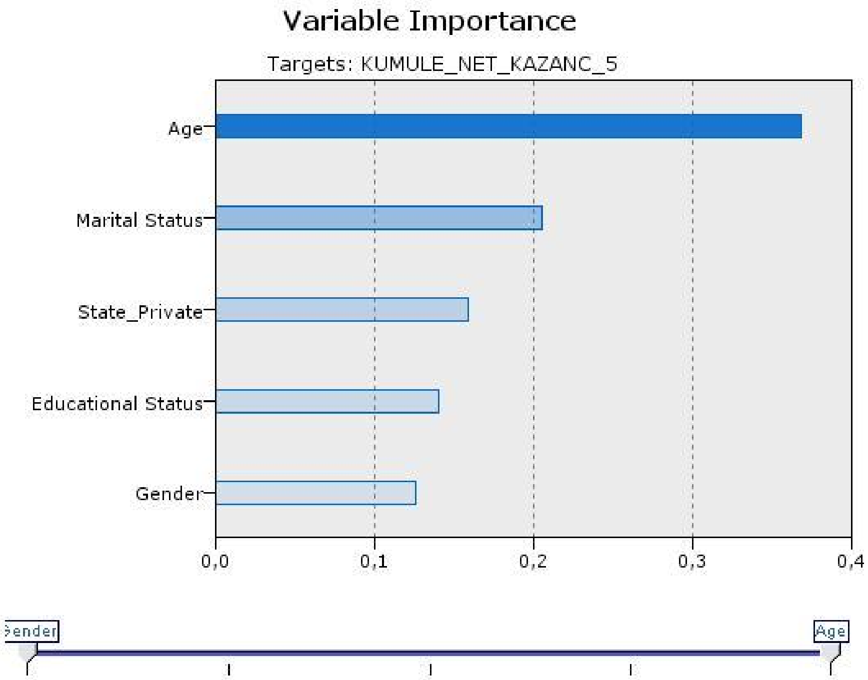


Figure 4.8 : The importance of demographic variables in C&R Tree Model

In another table, the variable importance table of Neural Net model is shown Figure 4.9. Educational Status and Marital Status are the most important variables on target. The following important variables are Gender, Age and State_Private, respectively. There is very little difference between Age and State_Private variables.

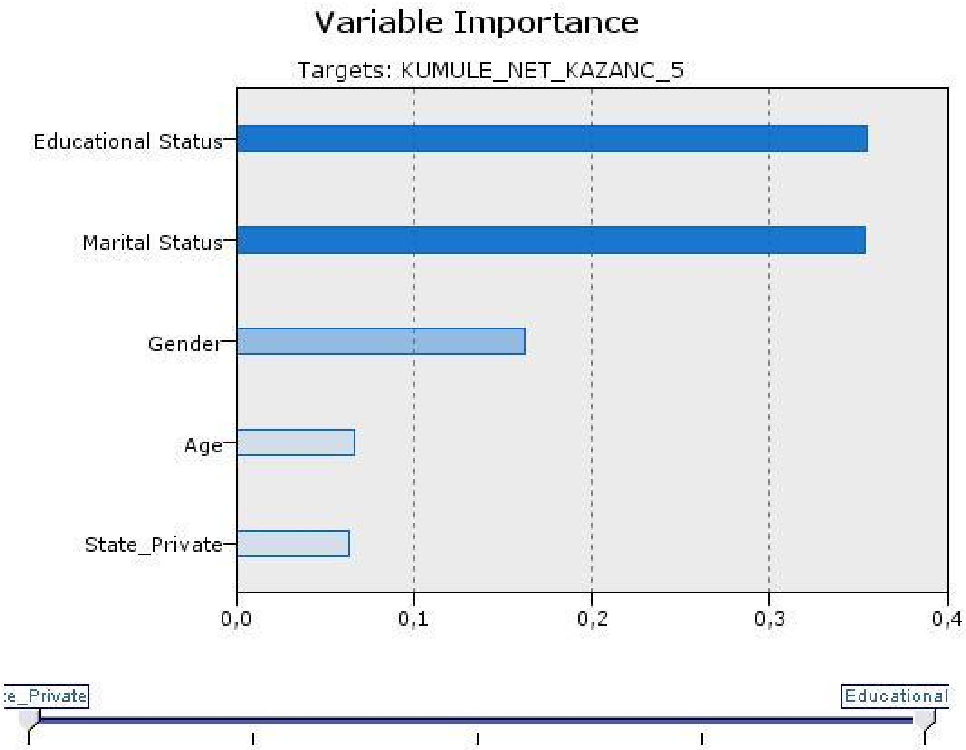


Figure 4.9 : The importance of demographic variables in Neural Net model

In the third table, the variable importance table of GenLin model is shown Figure 4.10. Profession is the most important variable on target. The following important variables are Age, Marital Status, Gender, State_Private and Educational Status respectively.

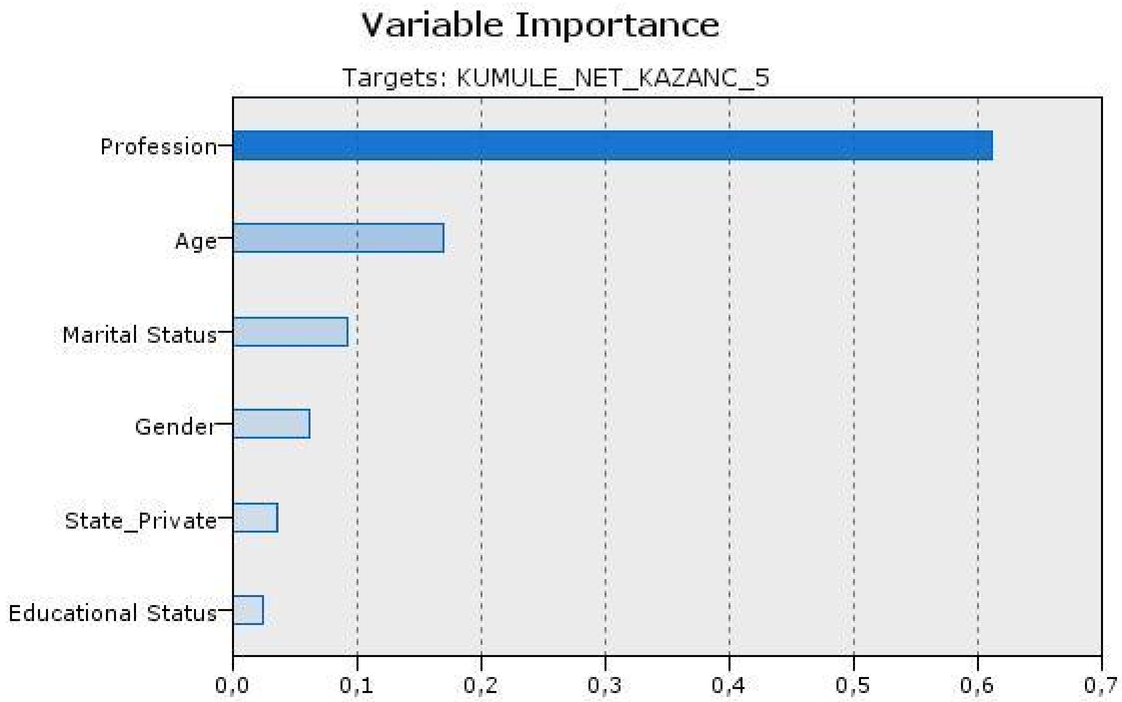


Figure 4.10 : The importance of demographic variables in GenLin model

In the final table, the variable importance table of CHAID model is shown Figureure 4.11. Age is the most important variable on target. The following important variables are Marital Status, State_Private and Gender, respectively.

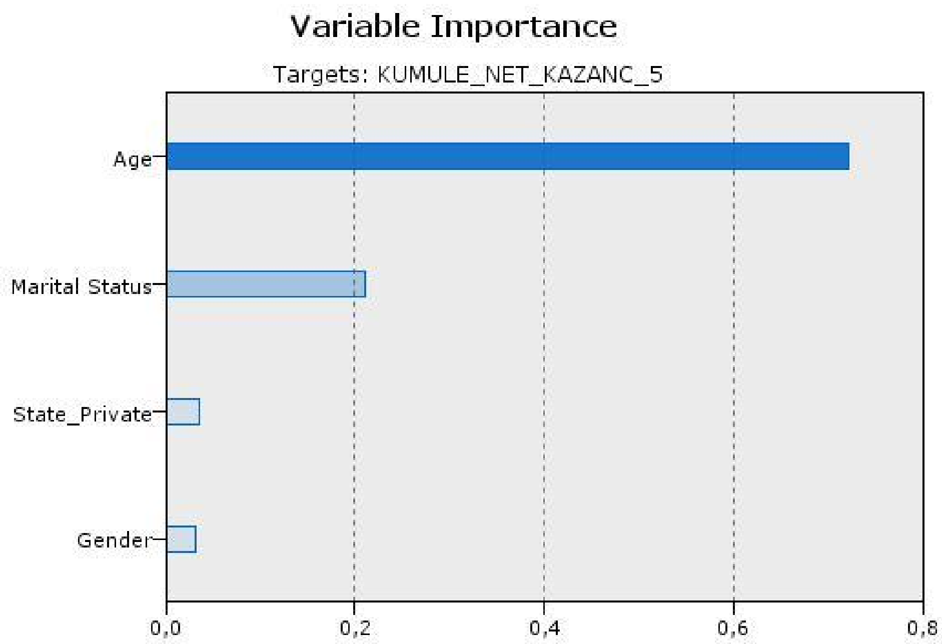


Figure 4.11 : The importance of demographic variables in CHAID model

The C&R Tree, Neural Net, GenLin and CHAID models are applied to data which include only demographic data and Target is cumulative net earning as it is shown in Figure 4.12. A new database which contains the scores of C&R Tree, Neural Net, GenLin and CHAID models is made.

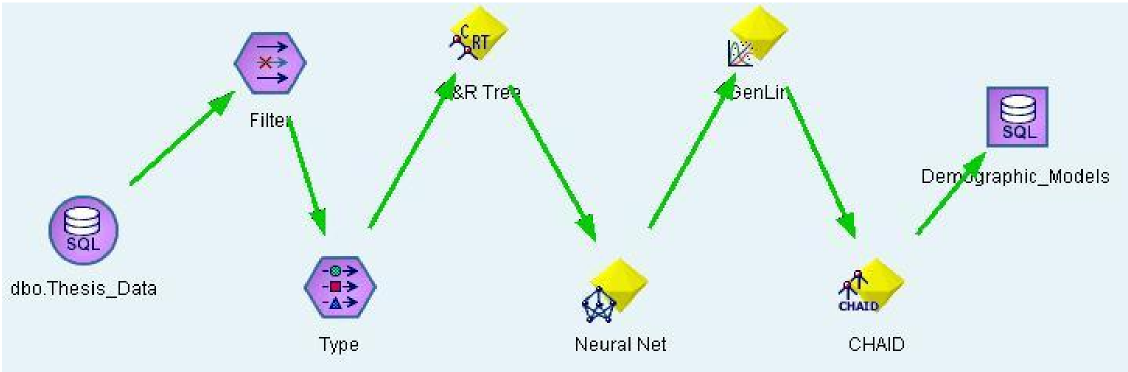


Figure 4.12 : The new database which contain scores of models.

The evaluation, analysis and numeric predictor are applied to new database which contain the scores of C&R Tree, Neural Net, GenLin and CHAID models as it is shown Figure 4.13. The analysis creates a report comparing the accuracy of predictive models. The evaluation creates charts comparing the accuracy of predictive models. The numeric predictor allows creation and comparison of predictive models with numeric outcomes.

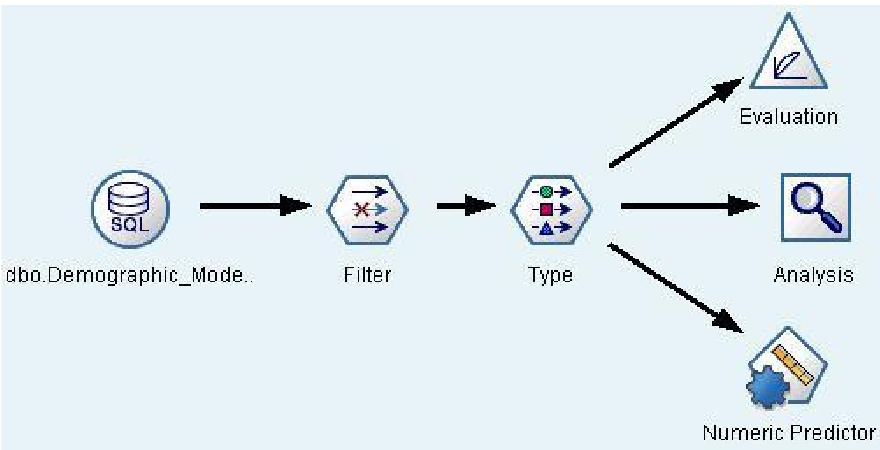


Figure 4.13 : The Evaluation and Analysis of Demographic Models.

The results of the evaluation on four models are shown in Figure 4.14. $R is used instead of C&R Tree model. $N is used instead of Neural Net model. $G is used instead of GenLin model. $R1 is used instead of CHAID model. The red line represents the real cumulative-earning in this Figureure. The C&R Tree and GenLin Models seem better than the others but it is not clear which one is the best according to this Figureure. Therefore, we will examine the analysis and the numeric predictor to find out which one is better than the others.

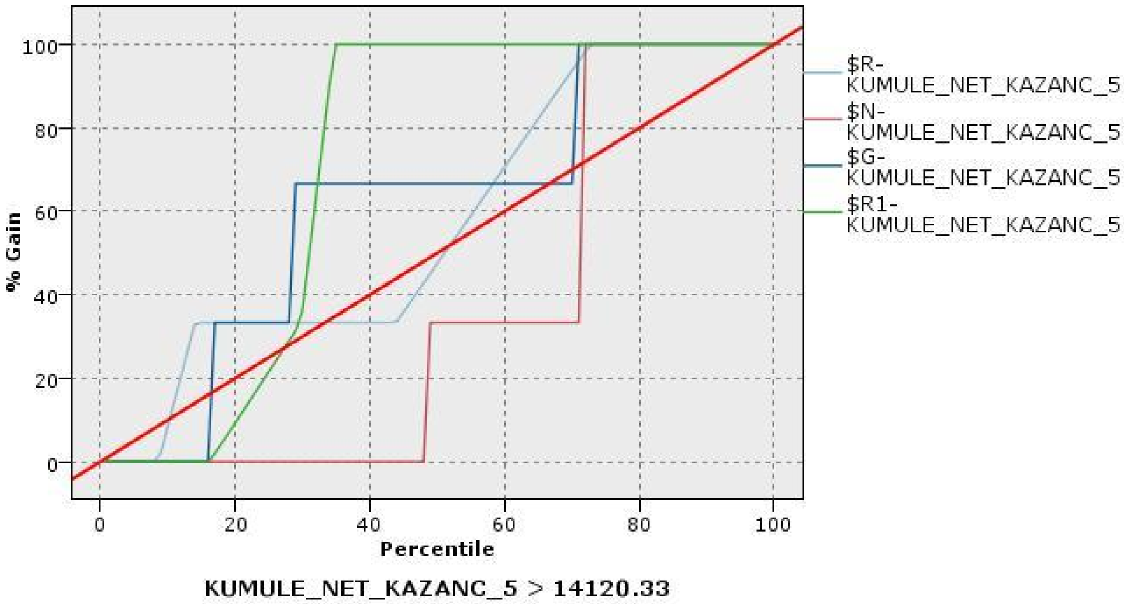


Figure 4.14 : The Evaluation of Demographic Models.

The results of the analysis on four models are shown in Figure 4.15. C&R Tree, GenLin and CHAID models are distinguished when we look at their properties. Especially, their mean errors are equal to zero but the best model does not seem an obvious way. A different model looks good in each step. GenLin is the best model if we look at linear correlation and mean absolute error but its maximum error, minimum error and standard deviation are not good so we do not understand which one is better than the other models. Therefore, we will examine the correlations of models.

Results for output field KUMULE_NET_KAZANC_5

Individual Models

Comparing $R-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-623,724
Maximum Error	28745,236
Mean Error	-0,0
Mean Absolute Error	79,264
Standard Deviation	283,264
Linear Correlation	0,068
Occurrences	50.000

Comparing $N-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-620,701
Maximum Error	28747,095
Mean Error	-0,429
Mean Absolute Error	79,784
Standard Deviation	283,476
Linear Correlation	0,056
Occurrences	50.000

Comparing $G-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-650,61
Maximum Error	28756,556
Mean Error	0,0
Mean Absolute Error	79,153
Standard Deviation	287,393
Linear Correlation	0,08
Occurrences	50.000

Comparing $R1-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-642,406
Maximum Error	28726,554
Mean Error	-0,0
Mean Absolute Error	79,305
Standard Deviation	283,282
Linear Correlation	0,067
Occurrences	50.000

Figure 4.15 : The comparison of C&R Tree, Neural Net, GenLin and CHAID Models with Cumulative Net Earning in Analysis Module.

C&R Tree is the best model because its correlation is the best as it is shown in Figure 4.16.

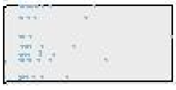
Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
	C&R Tree 1	< 1	0,091	4	0,992
	Generalized Line...	< 1	0,088	4	0,992
	Neural net 1	< 1	0,085	4	0,993
	CHAID 1	< 1	0,079	3	0,994

Figure 4.16 : The comparison of C&R Tree, Neural Net, GenLin and CHAID with Cumulative Net Earning in Numeric Predictor Module.

The correlation of C&R Tree is clearly shown in Figure 4.17. $R – KUMULE_NET_KAZANC_5 is instead of C&R Tree model.

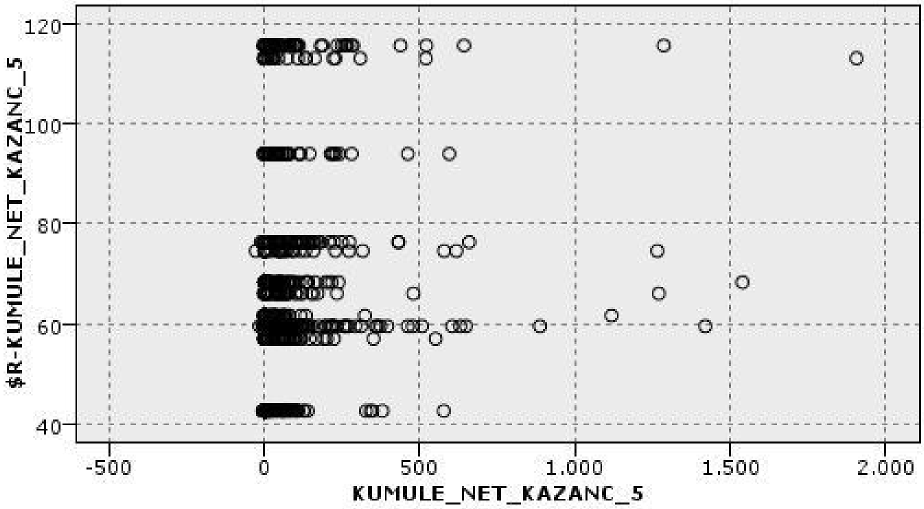


Figure 4.17 : The comparison of C&R Tree, Neural Net, GenLin and CHAID with Cumulative Net Earning in Numeric Predictor Module.

The demographic segmentation is made with C&R Tree model. When the model is applied to data, eleven subgroups are occurred. These subgroups are Node 2, Node 5, Node 6, Node 13, Node 14, Node 15, Node 16, Node 17, Node 18, Node 19 and Node 20.

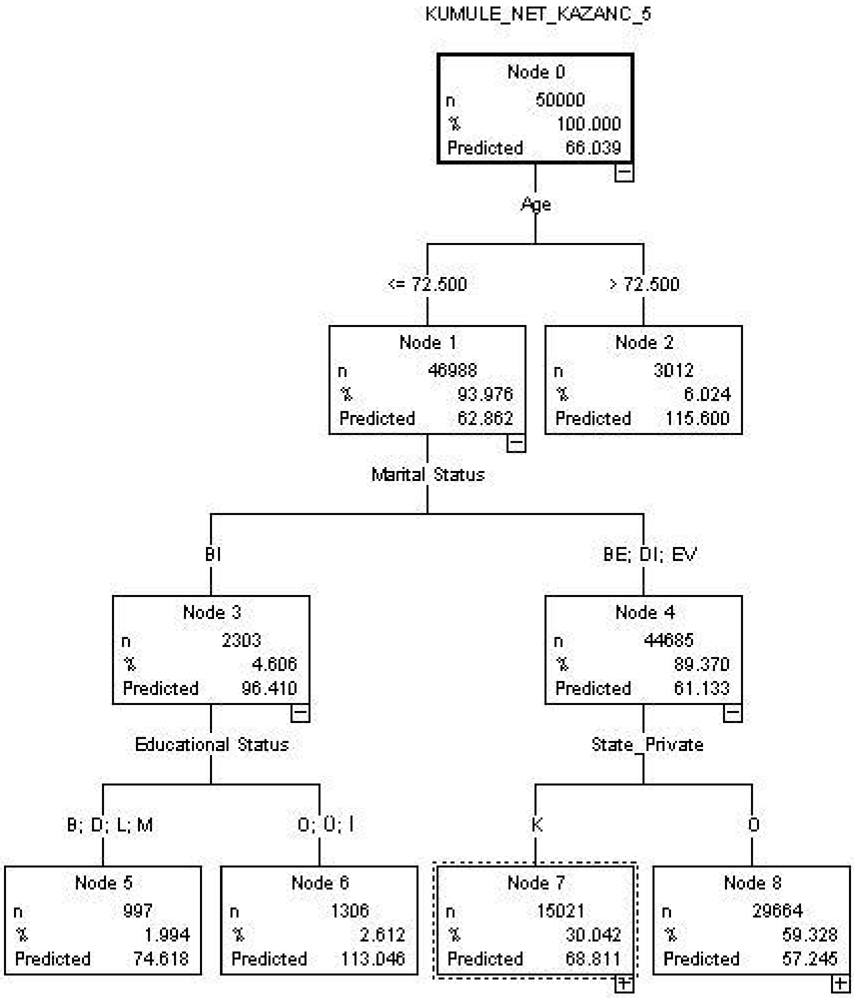


Figure 4.18 : The C&R Tree Model – Part 1

Node2, Node 5 and Node 6 are shown in Figure 4.18. Node 2 is consisted of people who are older than 72,5 year old. Node 5 is consisted of people who are younger than 72.5 year-old, have no information for marital status and are graduated from high school, doctorate, postgraduate or has no information in educational status. Node 6 is consisted of people who are younger than 72.5 year-old, have no information for marital status and are graduated from middle school, primary school or university.

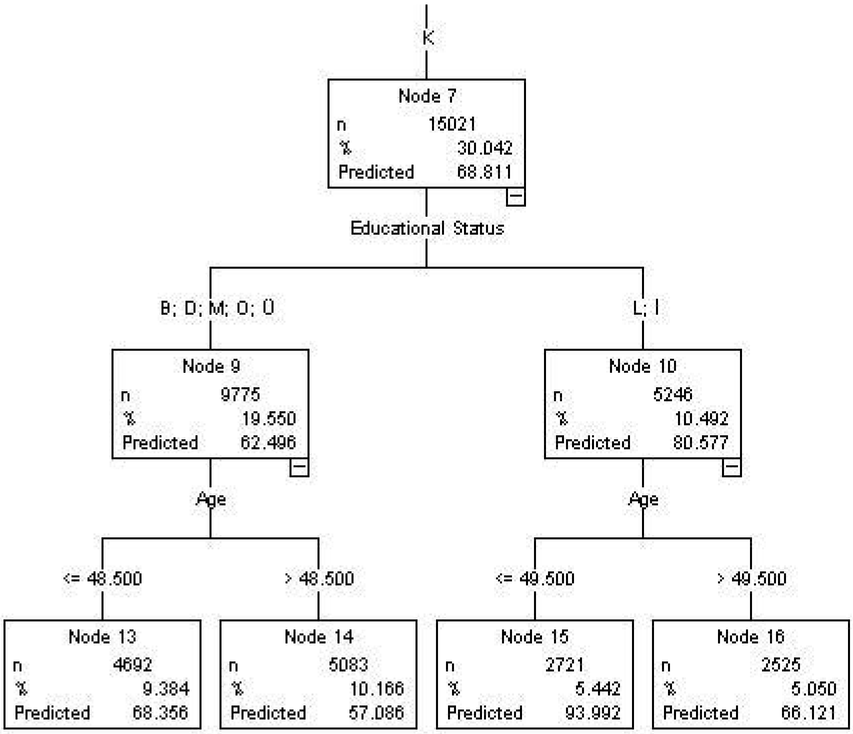


Figure 4.19 : The C&R Tree Model – Part 2

Node 13, Node 14, Node 15 and Node 16 are shown in Figure 4.19. Node 13 is consisted of people who are younger than 48.5 year-old, have information in marital Status, work in state, graduated from doctorate, postgraduate, university, middle school or has no information in educational status. Node 14 is consisted of people who are older than 48.5 year-old, have information in marital Status, work in state, graduated from doctorate, postgraduate, university, middle school or has no information in educational status. Node 15 is consisted of people who are younger than 49.5 year-old, have information in marital Status, work in state, graduated from high school or primary school. Node 16 is consisted of people who are older than 49.5 year-old, have information in marital Status, works in state, graduated from high school or primary school.

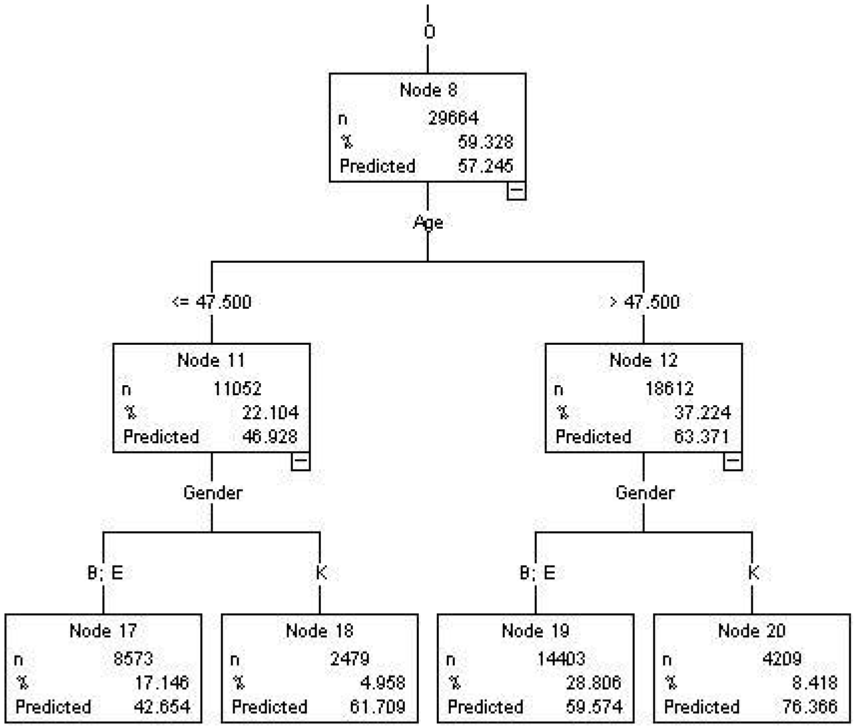


Figure 4.20 : The C&R Tree Model – Part 3

Node 17, Node 18, Node 19 and Node 20 are shown in Figure 4.20. Node 17 is consisted of people who are younger than 47.5 year-old, have information in marital Status, work in private sector and are man or have no information in gender. Node 18 is consisted of people who are younger than 47.5 year-old, have information in marital Status, work in private sector and are woman. Node 19 is consisted of people who are older than 47.5 year-old, have information in marital Status, work in private sector and are man or have no information in gender. Node 20 is consisted of people who are older than 47.5 year-old, have information in marital Status, work in private sector and are woman.

11 clusters have emerged as a result of this segmentation. We see a non-homogeneous distribution when we examine these clusters. For instance, a very large community is in cluster 14. The clusters should be homogenous according to the results to compare. Therefore; clusters which are small and have non-obvious features should be brought together. In this way, the clusters are more easily managed

Taking into account the above-mentioned reasons, 2–5–6 clusters are brought together. 5–6 clusters are divided into a differentiation because marital status is unknown. The lack of information has revealed a cluster of wrong here. This mass is 4 percent of the total mass. Customers over the age of seventy are located in cluster 2. These customers probably have term deposit product or receive their pension from the bank. Here, too, we can say that aggressive marketing is not applicable. Therefore, it will be much more logical and easy to manage by combining these three clusters.

The distinction does not seem logical because of education levels in clusters 9 and 10 because the level of education has been made significantly. For this reason, we remove this distinction. We combine the 13-15 and 14-16 in the following clusters were separated according to age criteria.

We do not change anything in the clusters 17, 18, 19 and 20 because because these clusters have a very sharp and clear distinctions. Age and gender discrimination is clear in these clusters which are under the private sector distinction.

The rules of C&R Tree model are shown below.

The input variables are:

- Age,
- Marital Status,
- Educational Status,
- Gender,
- State_Private.

The values which can be taken by variables are shown at data understanding section. $Score refers the results of the model.

Rule 1: If Age > 72,5 then $Score: 115,6 (Node 2)

Rule 2: If Age < 72,5 and Marital Status = BI and (Educational Status = B or Educational Status = D or Educational Status = L or Educational Status = M) then $Score: 74,618 (Node 5)

Rule 3: If Age < 72,5 and Marital Status = BI and (Educational Status = O or Educational Status = Ü or Educational Status = İ) then $Score: 113,046 (Node 6)

Rule 4: If Age <= 48,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = K and (Educational Status = B or Educational Status = D or Educational Status = M or Educational Status = O or Educational Status = Ü) then $Score: 68,356 (Node 13)

Rule 5: If Age > 48,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = K and (Educational Status = B or Educational Status = D or Educational Status = M or Educational Status = O or Educational Status = Ü) then $Score: 57,086 (Node 14)

Rule 6: If Age <= 49,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = K and (Educational Status = L or Educational Status = İ) then $Score: 93,992 (Node 15)

Rule 7: If Age > 49,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = K and (Educational Status = L or Educational Status = İ) then $Score: 66,121 (Node 16)

Rule 8: If Age <= 47,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = O and (Gender = B or Gender = E) then $Score: 42,654 (Node 17)

Rule 9: If Age <= 47,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = O and Gender = K then $Score: 61,709 (Node 18)

Rule 10: If Age > 47,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = O and (Gender = B or Gender = E) then $Score: 59,574 (Node 19)

Rule 11: If Age > 47,5 and (Marital Status = BE or Marital Status = DI or Marital Status = EV) and State_Private = O and Gender = K then $Score: 76,366 (Node 20)

4.2.2 Value Segmentation

The value-based segmentation distributes customers to the segments according to according to the net earning obtained from the customers. As a result of this, the profitable customers are analyzed to find out what makes them profitable. This analysis is needed to increase the profitable of the less profitable customers.

The value based segmentation is created by combining two different approaches. The first approach is to determine which products are the hook products for customers. Another approach is to determine the impact of the product on the profitability. After obtaining these two approaches the value-based segmentation will be achieved.

In the first approach, a new data set which includes all of products of customer and the product ownership of customer is created. Then, we applied three models to data set to determine the effect of products on the product ownership.

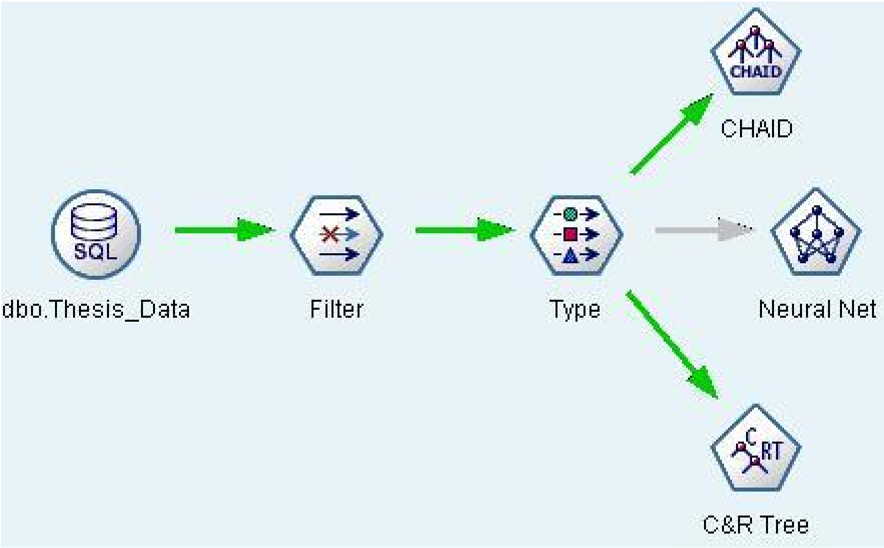


Figure 4.21 : CHAID, Neural Net and C&R Tree models.

A new database which contains the scores of Neural Net, CHAID and C&R Tree models is made.

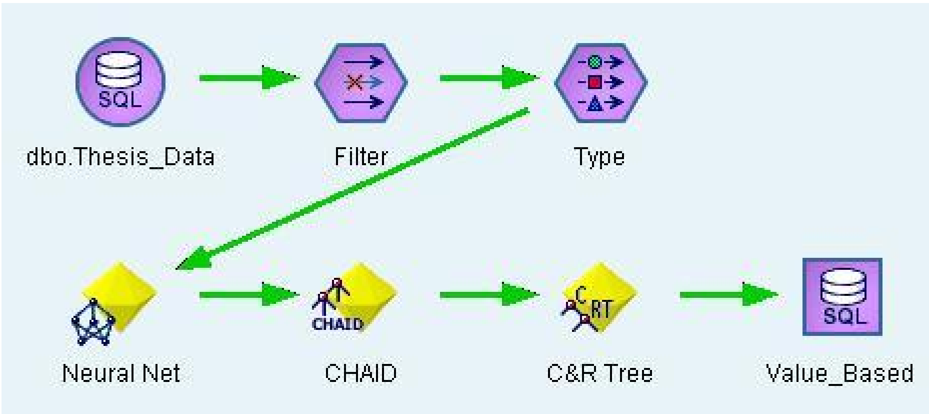


Figure 4.22 : The new database which contain scores of models.

The evaluation and analysis are applied to new database which contain the scores of Neural Net, CHAID and C&R Tree models as it is shown Figure 4.23. The analysis creates a report comparing the accuracy of predictive models. The evaluation creates charts comparing the accuracy of predictive models.

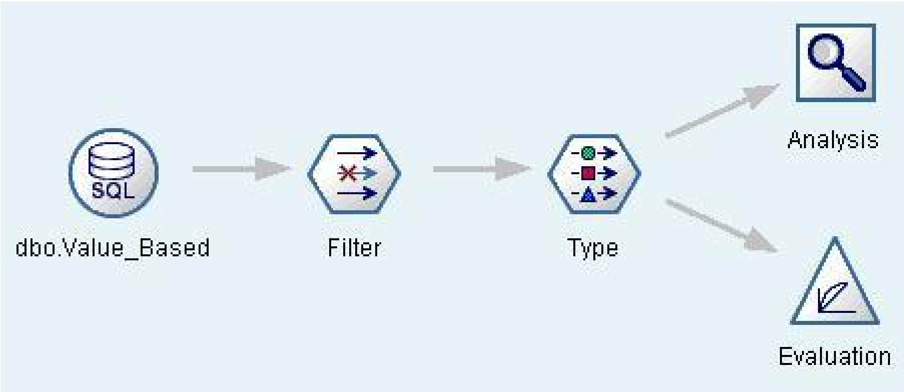


Figure 4.23 : The Analysis and Evaluation of Models.

The results of the evaluation on three models are shown in Figure 4.24. $N is used instead of Neural Net model. $R is used instead of CHAID model. $R1 is used instead of C&R Tree model. The red line represents the product ownership in this Figure. All of the models seem very close to each other so it is not clear which one is the best according to this Figure. Therefore, we will examine the analysis to find out which one is better than the others.

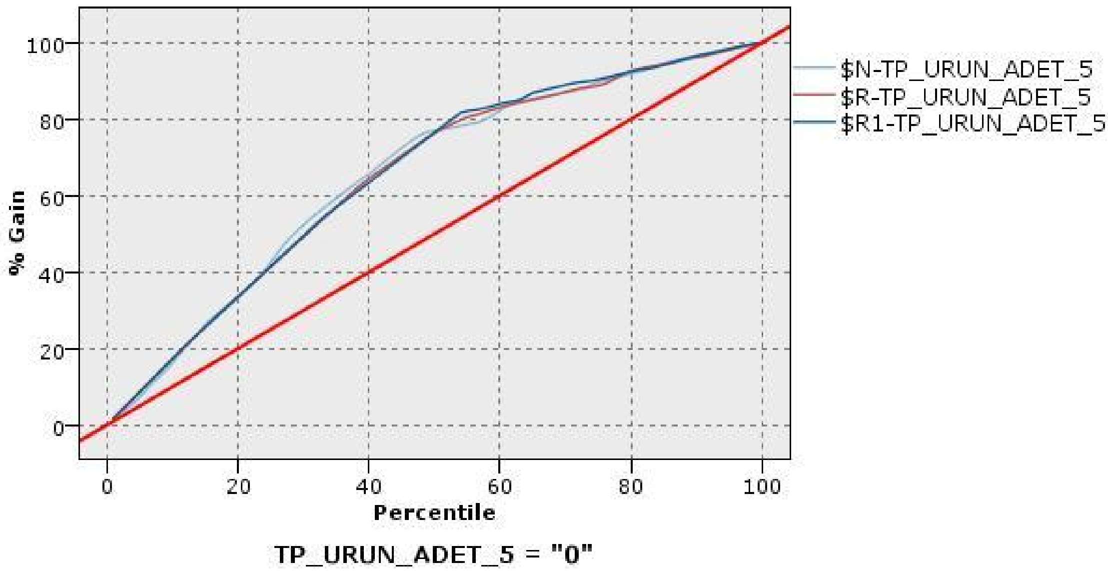


Figure 4.24 : The Evaluation Graph of Value-Based Models.

The results of the analysis on three models are shown in Figure 4.25. The best model can be easily seen in this analysis. The Neural Net model made the best estimate in the models so we choose it.

Results for output field TP_URUN_ADET_5

Individual Models

Comparing $N-TP_URUN_ADET_5 with TP_URUN_ADET_5

Correct	34.747	69,49%
Wrong	15.253	30,51%
Total	50.000	

Comparing $R-TP_URUN_ADET_5 with TP_URUN_ADET_5

Correct	28.173	56,35%
Wrong	21.827	43,65%
Total	50.000	

Comparing $R1-TP_URUN_ADET_5 with TP_URUN_ADET_5

Correct	29.762	59,52%
Wrong	20.238	40,48%
Total	50.000	

Figure 4.25 : The comparison of Neural Net, CHAID and C&R Tree Models with The Product Ownership in Analysis Module.

The variable importance table of Neural Net model is shown Figure 4.26. Deposit account is the most important variable on target. The following important variables are personal loan, foreign exchange deposit account, direct debit, investment account, credit card, insurance, overdraft account, checking account, nar account and gold account, respectively. Deposit account and personal loan are the most important factors to increase the product ownership of the customer.

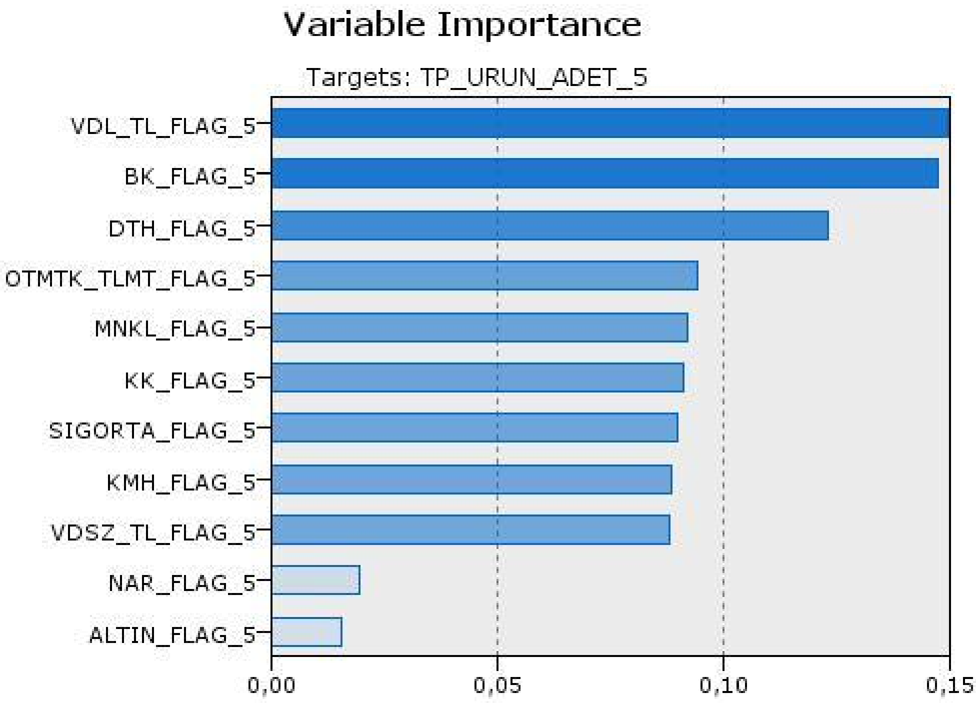


Figure 4.26 : The importance of variables in Neural Net model

In the second approach, a new data set which includes all of products of customer and the cumulative net earning of customer is created. Then, we applied three models to data set to determine the effect of products on the profitability of customer.

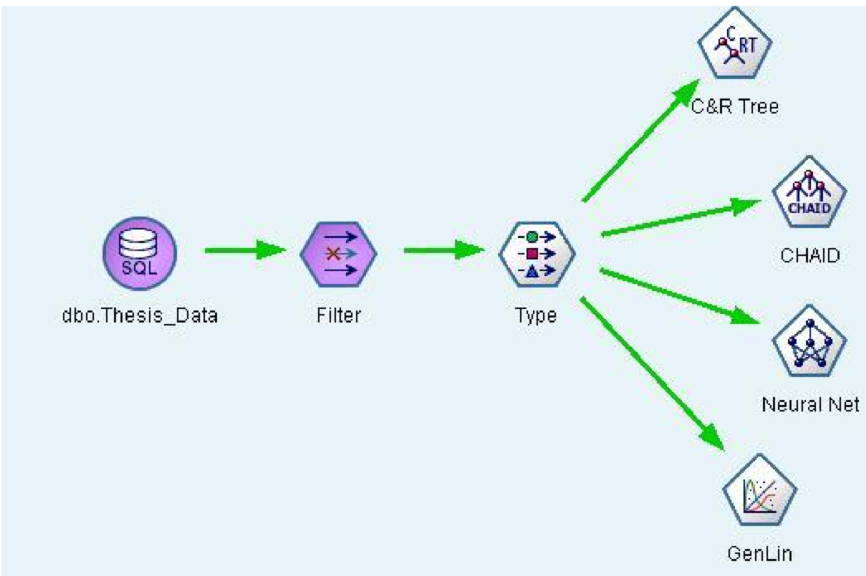


Figure 4.27 : C&R Tree, Neural Net, GenLin and CHAID models

A new database which contains the scores of Neural Net, GenLin, CHAID and C&R Tree models is made.

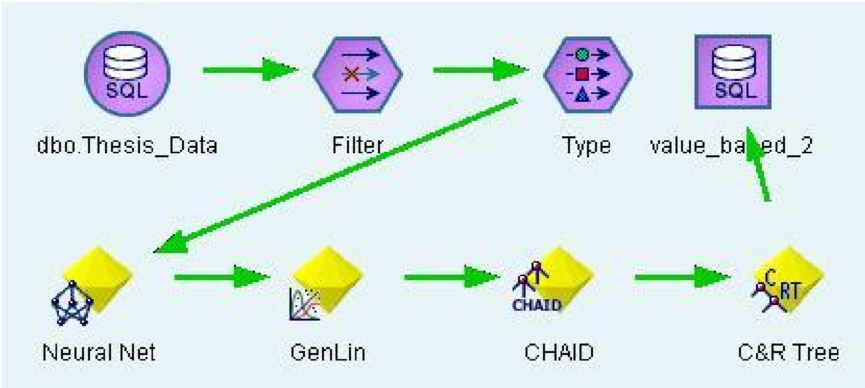


Figure 4.28 : The new database which contain scores of models.

The analysis, numeric predictor and evaluation are applied to new database which contain the scores of Neural Net, GenLin, CHAID and C&R Tree models as it is shown Figure 4.29. The analysis creates a report comparing the accuracy of predictive models. The numeric predictor allows creation and comparison of predictive models with numeric outcomes. The evaluation creates charts comparing the accuracy of predictive models.

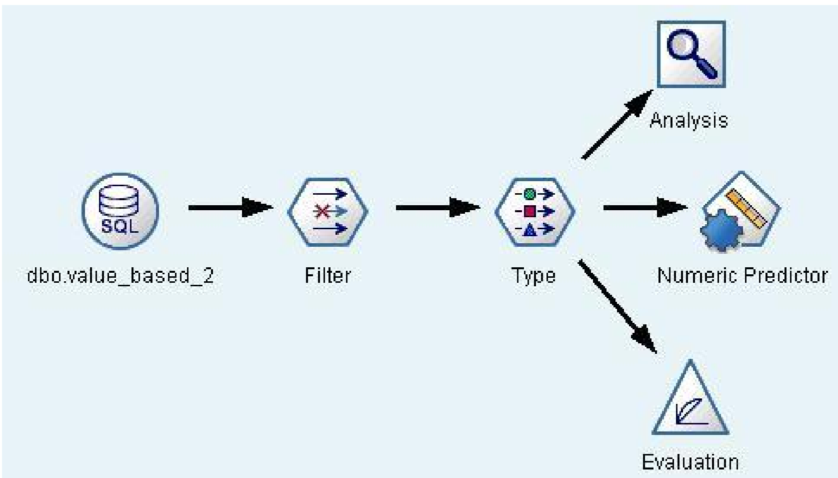


Figure 4.29 : The Evaluation and Analysis of Models.

The results of the evaluation on four models are shown in Figure 4.30. $N is used instead of Neural Net model. $G is used instead of GenLin model. $R is used instead of

CHAID model. $R1 is used instead of C&R Tree model. The red line represents the real cumulative-earning in this Figure. The Neural Net and GenLin models seem better than the others but it is not clear which one is the best according to this Figure. Therefore, we will examine the analysis and the numeric predictor to find out which one is better than the others.

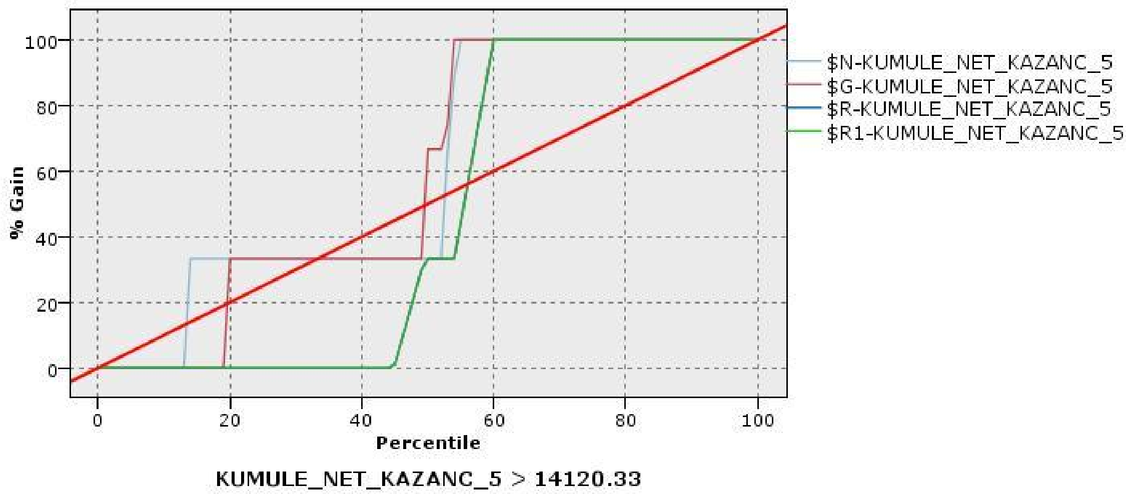


Figure 4.30 : The Evaluation of Models.

The results of the analysis on four models are shown in Figure 4.31. GenLin and CHAID models are distinguished when we look at their properties. Especially, their mean errors are equal to zero but the best model does not seem an obvious way. A different model looks good in each step. GenLin is the best model if we look at linear correlation and maximum error but its mean absolute error, minimum error and standard deviation are not good so we do not understand which one is better than the other models. Therefore, we will examine the correlations of models.

Results for output field KUMULE_NET_KAZANC_5

Individual Models

Comparing $N-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-672,535
Maximum Error	28752,737
Mean Error	1,281
Mean Absolute Error	65,31
Standard Deviation	278,38
Linear Correlation	0,197
Occurrences	50.000

Comparing $G-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-686,149
Maximum Error	28748,813
Mean Error	0,0
Mean Absolute Error	65,676
Standard Deviation	278,378
Linear Correlation	0,197
Occurrences	50.000

Comparing $R-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-669,386
Maximum Error	28770,837
Mean Error	-0,0
Mean Absolute Error	63,268
Standard Deviation	277,931
Linear Correlation	0,204
Occurrences	50.000

Comparing $R1-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-673,286
Maximum Error	28770,837
Mean Error	-0,0
Mean Absolute Error	63,422
Standard Deviation	278,014
Linear Correlation	0,203
Occurrences	50.000

Figure 4.31 : The comparison of C&R Tree, Neural Net, GenLin and CHAID Models with Cumulative Net Earning in Analysis Module.

C&R Tree is the best model because its correlation is the best as it is shown in Figure 4.32.

Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
	Generalized Line...	< 1	0,205	4	0,958
	CHAID 1	< 1	0,205	4	0,958
	Neural net 1	< 1	0,204	4	0,958
	C&R Tree 1	< 1	0,199	4	0,96

Figure 4.32 : The comparison of C&R Tree, Neural Net, GenLin and CHAID with Cumulative Net Earning in Numeric Predictor Module.

The variable importance table of GenLin model is shown Figure 4.33. Deposit account is the most important variable on target. The following important variables are personal loan, insurance, foreign exchange deposit account, investment account, checking account, credit card, direct debit, nar account, gold account and overdraft account, respectively. Deposit account is the most important factor to increase the profitability of the customer.

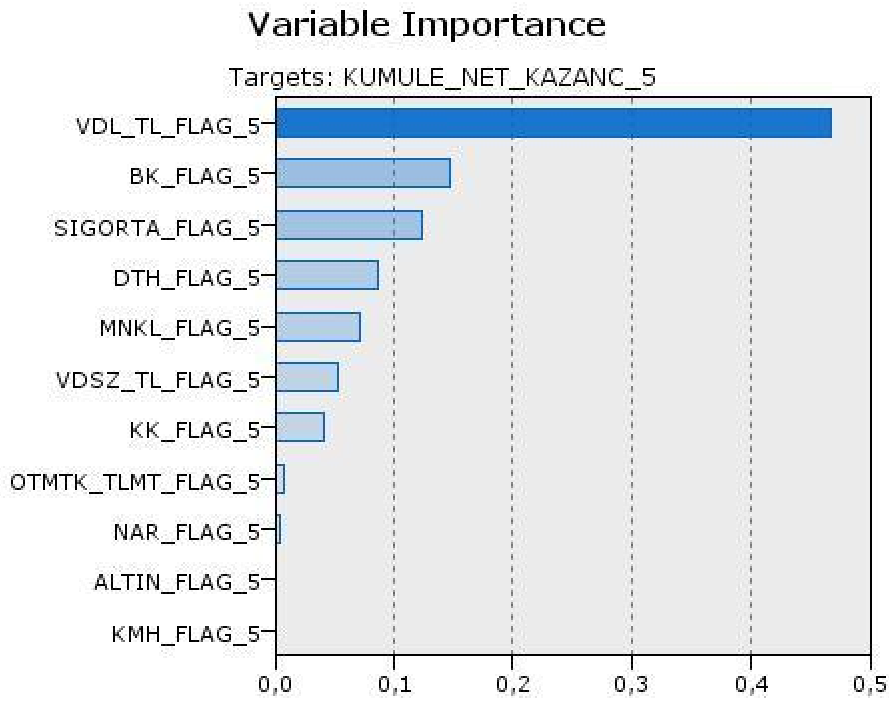


Figure 4.33 : The importance of variables in GenLin model.

In these two approaches, we have learned that deposit account and personal loan are important products that increase the product ownership and the profitability of customer. We know that credit card is the most profitable product when it is used. Also, it provides a better understanding of the customer. In addition to personal loan, deposit account and credit card, the clusters which are created in FCM and Kmeans are also used as an input in here. Now, we will use these variables with classification algorithms to achieve the value based segmentation.

Firstly, we use clusters which are determined with FCM algorithm as an input in modeling phase. The other inputs are: personal loan, deposit account and credit card. The target is cumulative net earning for modeling phase.

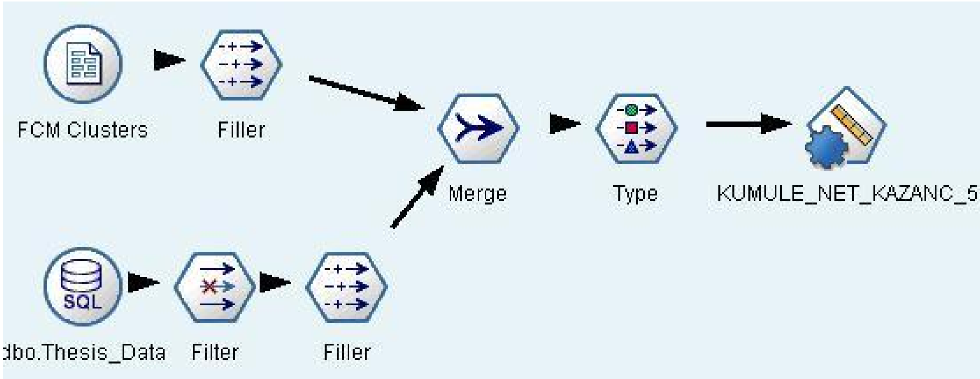


Figure 4.34 : The Selection of Best Model with Numeric Predictor.

C&R Tree is the best model because its correlation is the best as it is shown in Figure 4.35.

Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
	C&R Tree 1	< 1	0,205	4	0,958
	CHAID 1	< 1	0,204	4	0,959
	Generalized Line...	< 1	0,196	4	0,961
	Neural net 1	< 1	0,196	4	0,962

Figure 4.35 : The comparison of C&R Tree, CHAID, GenLin and Neural Net with Cumulative Net Earning in Numeric Predictor Module.

Secondly, we use clusters which are determined in K-means algorithm as an input in modeling phase. The other inputs are: personal loan, deposit account and credit card. The target is cumulative net earning for modeling phase.

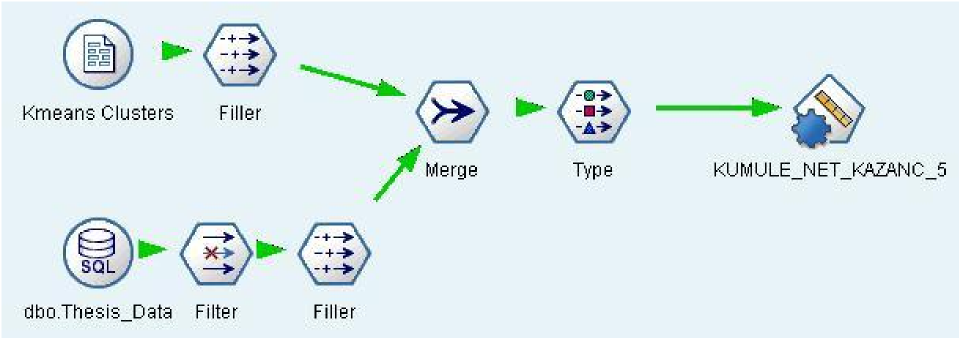


Figure 4.36 : The Selection of Best Model with Numeric Predictor.

C&R Tree is the best model because its correlation is the best as it is shown in Figure 4.37.

Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
	C&R Tree 1	< 1	0,186	4	0,965
	CHAID 1	< 1	0,186	4	0,966
	Generalized Line...	< 1	0,18	4	0,968
	Neural net 1	< 1	0,178	4	0,968

Figure 4.37 : The comparison of C&R Tree, CHAID, GenLin and Neural Net with Cumulative Net Earning in Numeric Predictor Module.

The different models are created with FCM and K-means algorithms. C&R Tree is the best model in two ways. Now , we compare these algorithms to determine which one is better.

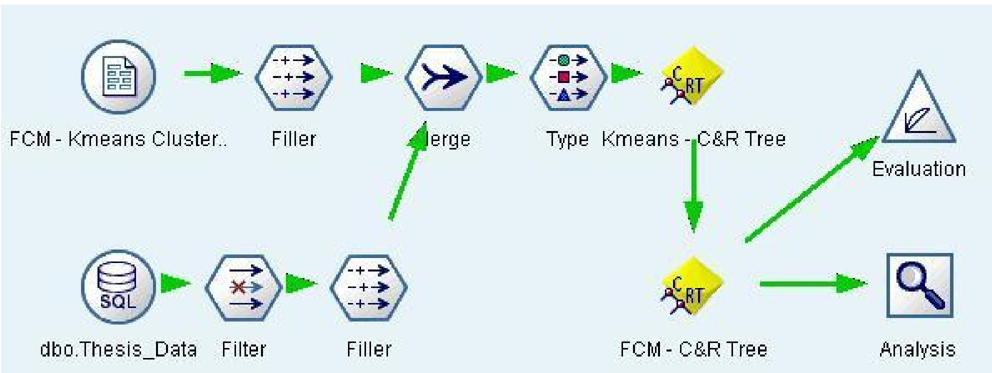


Figure 4.38 : The Evaluation and Analysis of Models.

The results of the evaluation on two models are shown in Figure 4.39. $R is used instead of C&R Tree model which is created with K-means clusters. $R1 is used instead of C&R Tree model which is created with FCM clusters. The red line represents the real cumulative-earning in this Figure. The $R1 seem better than the $R but it is not clear which one is the best according to this Figure. Therefore, we will examine the analysis to find out which one is better than the other.

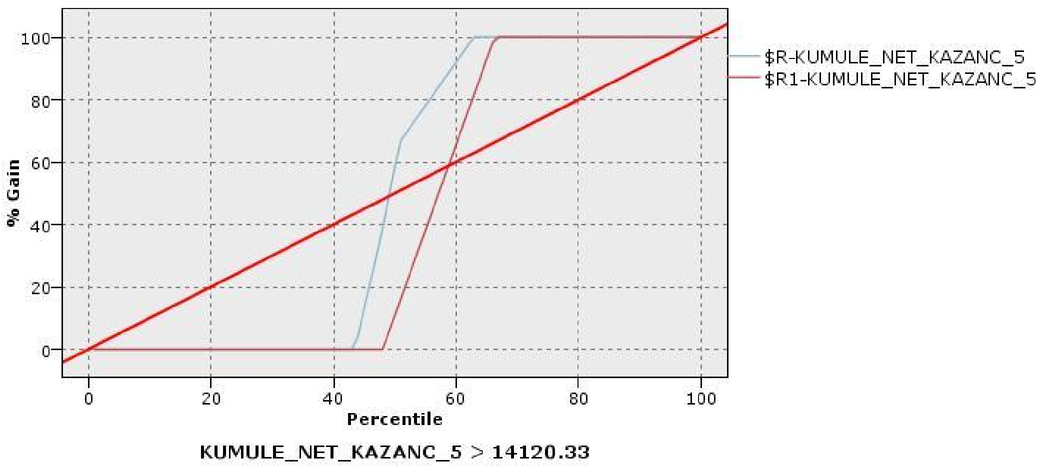


Figure 4.39 : The Evaluation of Models.

The results of the analysis on two models are shown in Figure 4.40. $R1 model is distinguished when we look at their properties. Especially, its linear correlation is better than linear correlation of $R. Therefore, we decided to use Fuzzy C-means in value based segmentation.

Results for output field KUMULE_NET_KAZANC_5

Individual Models

Comparing $R-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-661,638
Maximum Error	28755,158
Mean Error	0,0
Mean Absolute Error	65,68
Standard Deviation	278,956
Linear Correlation	0,186
Occurrences	50.000

Comparing $R1-KUMULE_NET_KAZANC_5 with KUMULE_NET_KAZANC_5

Minimum Error	-668,463
Maximum Error	28769,897
Mean Error	0,0
Mean Absolute Error	63,218
Standard Deviation	277,918
Linear Correlation	0,205
Occurrences	50.000

Agreement between $R-KUMULE_NET_KAZANC_5 $R1-KUMULE_NET_KAZANC_5

Comparing Agreement with KUMULE_NET_KAZANC_5

Minimum Error	-665,051
Maximum Error	28762,527
Mean Error	0,0
Mean Absolute Error	63,968
Standard Deviation	278,145
Linear Correlation	0,201
Occurrences	50.000

Figure 4.40 : The Analysis of Models.

The value-based segmentation is made with Fuzzy C-means and C&R Tree model. When the model is applied to data, nineteen subgroups are occurred. These subgroups are Node 8, Node 10, Node 15, Node 16, Node 18, Node 19, Node 23, Node 25, Node 26, Node 27, Node 28, Node 29, Node 30, Node 31, Node 32, Node 33, Node 34, Node 35 and Node 36.

The Figure of decision tree is very large. Therefore, we divided to decision tree to parts. The first part of decision tree is shown in Figure 4.41.

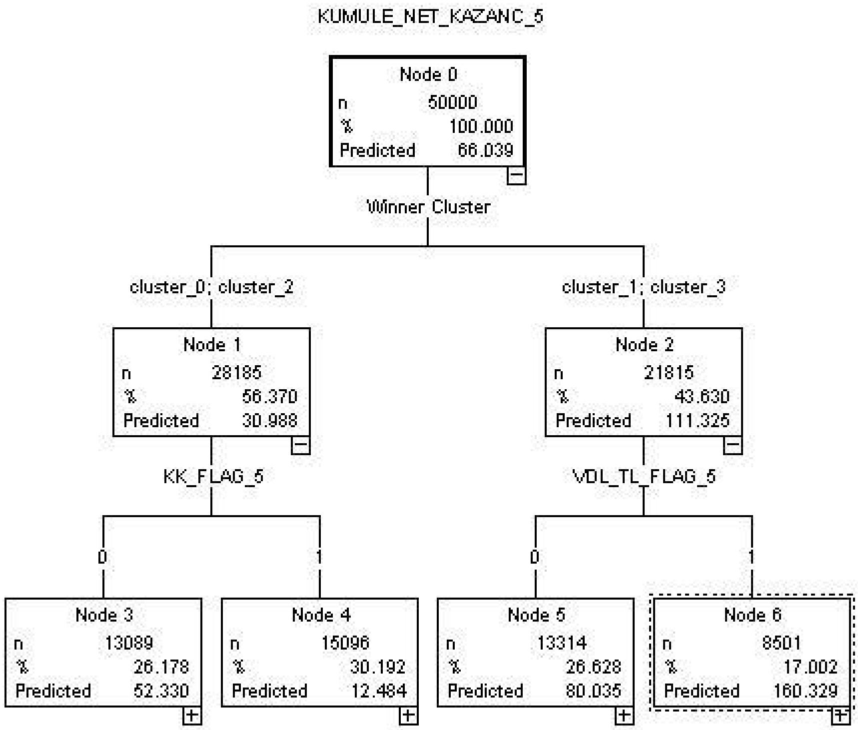


Figure 4.41 : The FCM - C&R Tree Model – Part 1

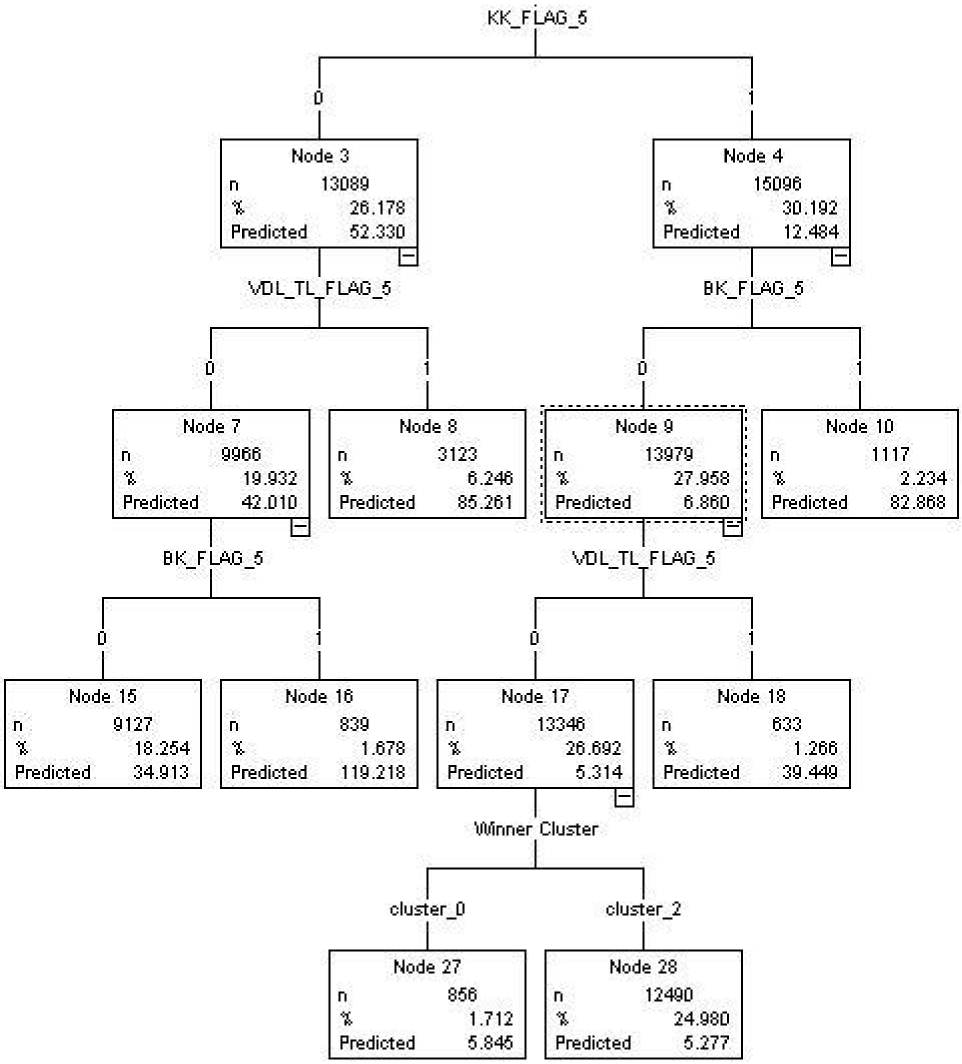


Figure 4.42 : The FCM - C&R Tree Model – Part 2

Node 8, Node 10, Node 15, Node 16, Node 18, Node 27 and Node 28 are shown in Figure 4.42. Node 8 is consisted of people who are in “cluster 0” or “cluster 2”, have deposit account and have not credit card. Node 10 is consisted of people who are in “cluster 0” or “cluster 2”, have credit card and personal loan. Node 15 is consisted of people who are in “cluster 0” or “cluster 2”, have not credit card, deposit account and personal loan. Node 16 is consisted of people who are in “cluster 0” or “cluster 2”, have not credit card and deposit account, have personal loan. Node 18 is consisted of people who are in “cluster 0” or “cluster 2”, have credit card and deposit account, have not personal loan. Node 27 is consisted of people who are in “cluster 0”, have credit card, have not personal loan and deposit account. Node 28 is consisted of people who are in “cluster 2”, have credit card, have not personal loan and deposit account.

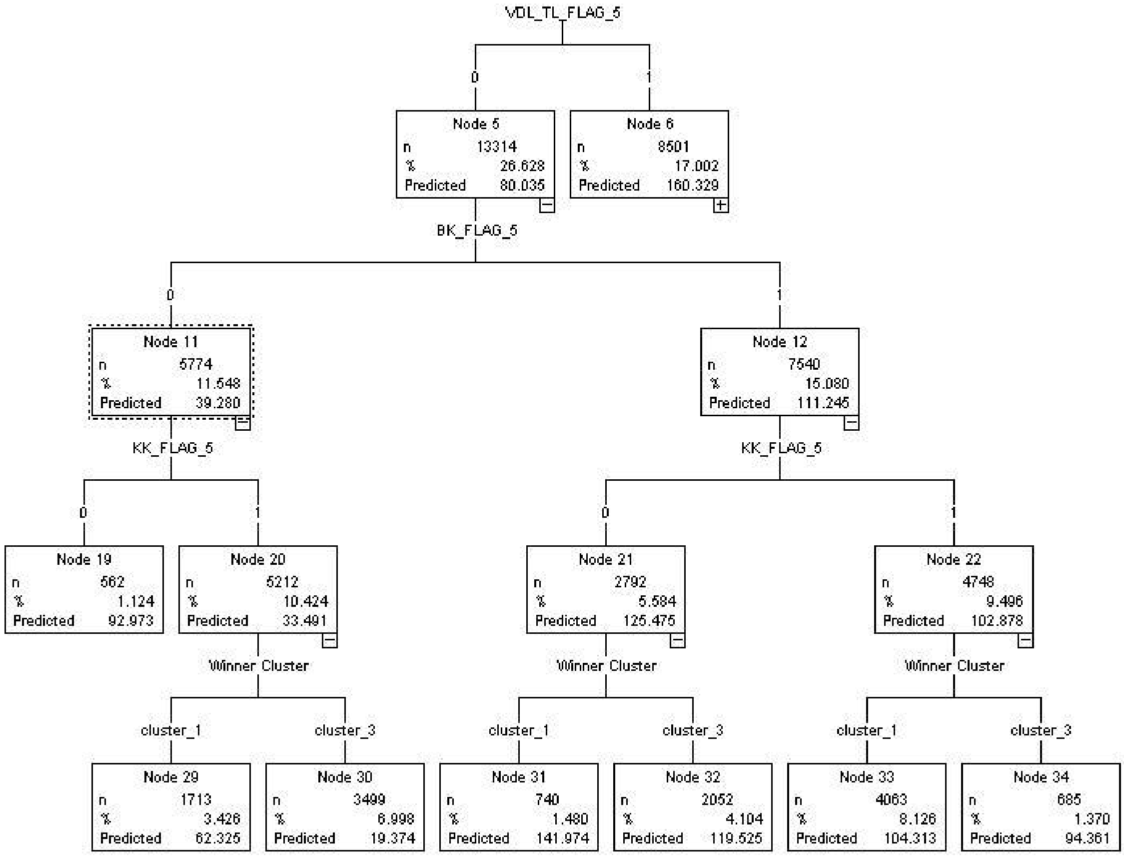


Figure 4.43 : The FCM - C&R Tree Model – Part 3

Node 19, Node 29, Node 30, Node 31, Node 32, Node 33 and Node 34 are shown in Figure 4.43. Node 19 is consisted of people who are in "cluster 1" or "cluster 3", have not deposit account, credit card and personal loan. Node 29 is consisted of people who are in "cluster 1", have not deposit account and personal loan, have credit card. Node 30 is consisted of people who are in "cluster 3", have not deposit account and personal loan, have credit card. Node 31 is consisted of people who are in "cluster 1", have not deposit account and credit card, have personal loan. Node 32 is consisted of people who are in "cluster 3", have not deposit account and credit card, have personal loan. Node 33 is consisted of people who are in "cluster 1", have not deposit account, and have personal loan and credit card. Node 34 is consisted of people who are in "cluster 3", have not deposit account, and have personal loan and credit card.

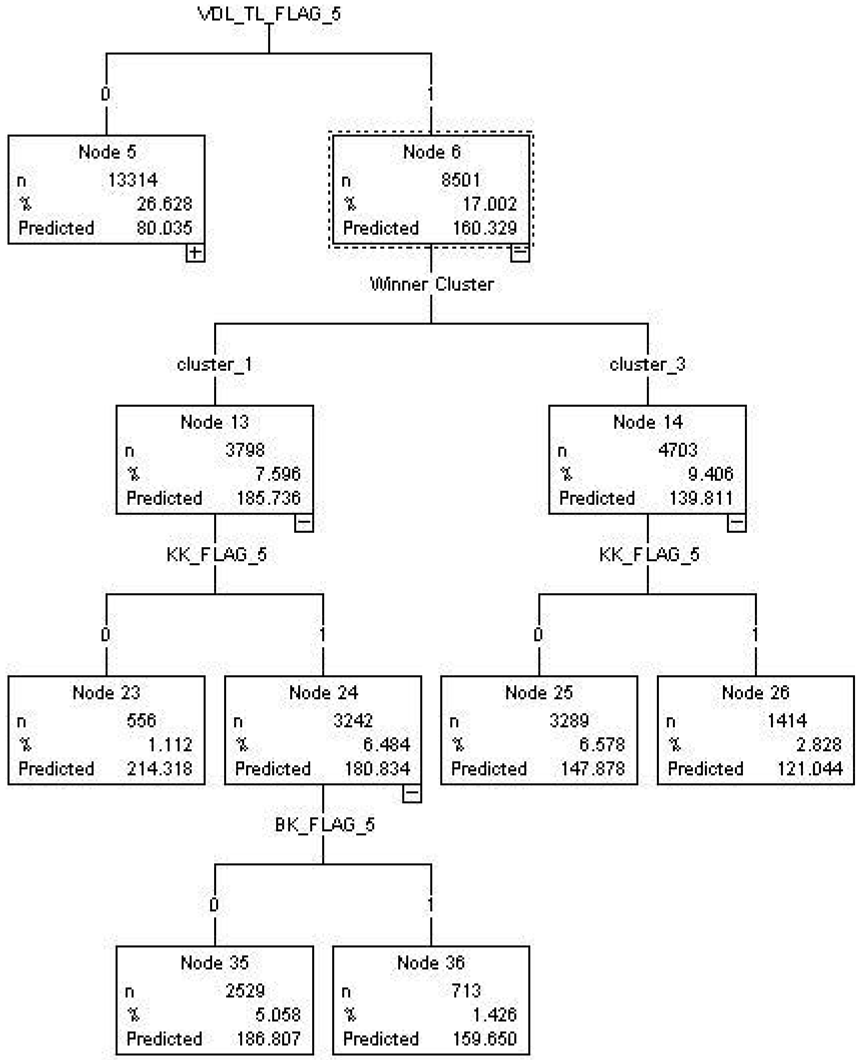


Figure 4.44 : The FCM - C&R Tree Model – Part 4

Node 23, Node 25, Node 26, Node 35 and Node 36 are shown in Figure 4.44. Node 23 is consisted of people who are in "cluster 1", have deposit account, and have not credit card. Node 25 is consisted of people who are in "cluster 3", have deposit account, and have not credit card. Node 26 is consisted of people who are in "cluster 3", have deposit account and credit card. Node 35 is consisted of people who are in "cluster 1", have deposit account and credit card, and have not personal loan. Node 36 is consisted of people who are in "cluster 1", have deposit account, credit card and personal loan.

The rules of C&R Tree model are shown below.

The input variables are:

- VDL_TL_FLAG_5 refers deposit account,
- BK_FLAG_5 refers personal loan,
- KK_FLAG_5 refers credit card,
- Winner Cluster refers clusters which are created in Fuzzy C-Means.

If value of the product is 1 then the customer has the product. If value of the product is 0 then the customer doesn't have the product. There are four clusters: Cluster_0, Cluster_1, Cluster_2 and Cluster_3. $Score refers the results of the model.

Rule 1: If VDL_TL_FLAG_5 = 1 and KK_FLAG_5 = 0 and (Winner Cluster = Cluster_0 or Winner Cluster = Cluster_2) then $Score = 85,261 (Node 8)

Rule 2: If BK_FLAG_5 = 1 and KK_FLAG_5 = 1 and (Winner Cluster = Cluster_0 or Winner Cluster = Cluster_2) then $Score = 82,868 (Node 10)

Rule 3: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 0 and (Winner Cluster = Cluster_0 or Winner Cluster = Cluster_2) then $Score = 34,913 (Node 15)

Rule 4: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 1 and KK_FLAG_5 = 0 and (Winner Cluster = Cluster_0 or Winner Cluster = Cluster_2) then $Score = 119,218 (Node 16)

Rule 5: If VDL_TL_FLAG_5 = 1 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and (Winner Cluster = Cluster_0 or Winner Cluster = Cluster_2) then $Score = 39,449 (Node 18)

Rule 6: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_0 then $Score = 5,845 (Node 27)

Rule 7: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_2 then $Score = 5,277 (Node 28)

Rule 8: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 0 and (Winner Cluster = Cluster_1 or Winner Cluster = Cluster_3) then $Score = 92,973 (Node 19)

Rule 9: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_1 then $Score = 62,325 (Node 29)

Rule 10: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_3 then $Score = 19,374 (Node 30)

Rule 11: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 1 and KK_FLAG_5 = 0 and Winner Cluster = Cluster_1 then $Score = 141,974 (Node 31)

Rule 12: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 1 and KK_FLAG_5 = 0 and Winner Cluster = Cluster_3 then $Score = 119,525 (Node 32)

Rule 13: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 1 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_1 then $Score = 104,313 (Node 33)

Rule 14: If VDL_TL_FLAG_5 = 0 and BK_FLAG_5 = 1 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_3 then $Score = 94,361 (Node 34)

Rule 15: If VDL_TL_FLAG_5 = 1 and KK_FLAG_5 = 0 and Winner Cluster = Cluster_1 then $Score = 214,318 (Node 23)

Rule 16: If VDL_TL_FLAG_5 = 1 and KK_FLAG_5 = 0 and Winner Cluster = Cluster_3 then $Score = 147,878 (Node 25)

Rule 17: If VDL_TL_FLAG_5 = 1 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_3 then $Score = 121,044 (Node 26)

Rule 18: If VDL_TL_FLAG_5 = 1 and BK_FLAG_5 = 0 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_1 then $Score = 186,807 (Node 35)

Rule 19: If VDL_TL_FLAG_5 = 1 and BK_FLAG_5 = 1 and KK_FLAG_5 = 1 and Winner Cluster = Cluster_1 then $Score = 159,65 (Node 36)

5. DISCUSSION AND CONCLUSIONS

In this study, we have a sample data set which contains the banking information. After the data set is prepared, we compared four different predictive algorithms to achieve the best models for customers in this data. Then, we made two different segmentations: demographic segmentation and value segmentation with the best models. As a result of these segmentations, we divided our customers into segments according to their yields. The characteristics of good customers were found so we can improve the efficiency and performance of bad customers to increase their yields while examining the high-yield ones. On the other hand, campaign for each segment can be easily performed because we have similar customers in the segments and we know their products, values and product diversity.

Finally, we made two different customer segmentations: demographic and value segmentations for campaign management with four predictive algorithms (Neural Net, C&R Tree, CHAID and GenLin), k-means and fuzzy c-means in this study. As a result of these studies, we clearly understand that customer segmentation sine qua non for understanding and managing of the customer. In other words, it is the first step of customer relation management and it shows the paths to be followed in the management of customer. Campaign management is the last step of customer relation management. The knowledge and know-how from customer segmentation is applied and tested in this step. According to the results that are obtained here, this cycle will continue or be re-designed for improvement.

In this paper, we compared two structures to reach the best result for the value segmentation. These structures are developed with fuzzy c-means, k-means and predictive models. The results of fuzzy c-means and k-means are used as an input for predictive models. C&R Tree appeared as the best model for these two approaches. It is clear that the model which is developed on fuzzy c-means is better than the model which is developed on k-means when we compared the results of these two approaches. Therefore, we can easily say that fuzzy c-means is better than k-means and fuzzy c-means & C&R Tree is the best approach in modeling phase for this thesis. We clearly see the effects and distribution of input variables on target, when the approach which is fuzzy c-means & C&R Tree is applied to data.

The fuzzy c-means and C&R Tree approach is developed on the products which are deposit account, personal loan and deposit account and proper groups which are created in fuzzy c-means. This approach has nineteen rules to explain the model which is created at the end of this approach. Each rule defines a segment or sub segment. These rules can be explained in general as follows. Firstly, the segment which has the highest profitability consists of only term deposits customers. In other words, this segment consists of affluent customers. Secondly, the segment which has the least profitability consists of only credit card customers. Thirdly, although there is no product, it has not the least profitability. This information indicates that customer use the products which are not included in model. Fourthly, although there are all products, it has not the highest profitability. This information indicates that having all products does not mean the highest profitability. Lastly, the remaining rules are not as specific as rules which are in the above, despite of their distinctive features.

In conclusion, the observation of all the details of customers who have almost the same features like as a group are provided by customer segmentation which is formed by fuzzy c-means. This structure provides the all information to perform a successful campaign management. We tried to show how this process that covering customer segmentation is applied in this paper so this study will be a source in this area.

REFERENCES

Anderson ET. *Sharing the wealth: when should firms treat customers as partners?*. Management Science 2002;48(8): 955–71.

Anderson K, Kerr C. *Customer relationship management*. New York: McGraw-Hill Trade; 2001.

Bergeron B. *CRM: the customer isn't always right*. Journal of Corporate Accounting & Finance 2002;**14**(1):53–7.

G. Mihelis a,*, E. Grigoroudis b, Y. Siskos b, Y. Politis b, Y. Malandrakis b : *Customer satisfaction measurement in the private bank sector* ; European Journal of Operational Research 130 (2001) 347±360

James Devlina,_, Philip Gerrardb : *Astudy of customer choice criteria for multiple bank users* ; Journal of Retailing and Consumer Services 12 (2005) 297–306

J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

J.T. Tou, R.C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, London, 1974.

L.Kaufman,P.J.Rousseeuw,*Finding Groupsin Data:An Introduction to Cluster Analysis*, Wiley,NewYork,1990.

Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan,Michael E. Welge, *Knowledge management and data mining for marketing*, Decision Support Systems 31 (2001) 127–137.

Michael Haenlein, Andreas M. Kaplan, Anemone J. Beeser: *A Model to Determine Customer Lifetime Value in a Retail Banking Context*; European Management Journal Vol. 25, No. 3, pp. 221–234, 2007

Nan-Chen Hsieh: *An integrated data mining and behavioral scoring model for analyzing bank customers* ; Expert Systems with Applications 27 (2004) 623–633

Pete Chapman, Randy Kerber, Julian Clinton, Thomas Khabaza,Thomas Reinartz, Rüdiger Wirth, *The CRISP-DM Process Model*, CRISP-DM Discuusion Paper, 1999.

Sushmita Mitra a, WitoldPedrycz , BishalBarman : *Shadowed c-means: Integrating fuzzy and rough clustering* ; Pattern Recognition 43 (2010) 1282–1291

Su-Yeon Kim a, Tae-Soo Jung b, Eui-Ho Suh c, Hyun-Seok Hwang d,* : *Customer segmentation and strategy development based on customer lifetime value: A case study*; Expert Systems with Applications 31 (2006) 101–107

S.Mitra, H.Banka, W.Pedrycz, *Rough–fuzzy collaborative clustering*, IEEE Transactionson Systems, Man,and Cybernetics—Part B36(2006) 795–805.

Y.Y. Yao : *A comparative study of fuzzy sets and rough sets* ; Journal of Information Sciences **109** (1998) 227-242

Z.Huang, *Extensions to the k-means algorithm for clustering large data sets with categorical values*, Data Mining and Knowledge Discovery 2 (1998) 283–304.

Z. Pawlak, *Rough sets*, International Journal of Computer and Information Sciences **11** (5) (1982) 341–356.

Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Boston, 1991.

Z. Pawlak, *Rough set approach to knowledge-based decision support*, European Journal of Operational Research **99** (1) (1997) 48–57.

APPENDICES

APPENDIX A

Table A.1 : The variables of data.

Data Name	Value Type	Comment
Must_No	Int	Customer ID
Müşteri Tipi	Nvarchar(1)	Customer Type
Doğum Tarihi	DateTime	Birth Date
Muta Şb	Varchar(4)	Branch Offiec of Bank Customer
Meslek Kod	Small Int	Job Code
meslek	Nvarchar(255)	Job
KamuOzel	Varchar(1)	Work Place Information - (Is it state or private?)
Medeni Durum	Varchar(2)	Marital Status
Eğitim Durum	Nvarchar(1)	Educational Background
Cinsiyet	Nvarchar(1)	Gender
İlk tanımlama tar	DateTime	Initial Date (for customer)
İLİ	Nvarchar(50)	City
MİY_SICIL_NO	Varchar(5)	Customer Relations Officer ID
ABONE24_TUT_1	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_1	Float	The Amount of Gold Account
ALTIN_FLAG_1	Bit	The Flag of Gold Account
BK_FLAG_1	Bit	The Falg of Personal Loan
BK_ORT_TUT_1	Float	The Average Monthly Amount of Personal Loans
BORDRO24_FLAG_1	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_1	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_1	Bit	The Flag of Debit Card
DIALOG_ISL_TUT_1	Float	The Amount of Transactions on The Call Center
DTH_FLAG_1	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_1	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_1	Int	The Number of Loan Delay
INTRNT_ISL_TUT_1	Float	The Amount of Transactions on The Internet
KK_FLAG_1	Bit	The Flag of Credit Card
KK_LIMIT_1	Float	The Limit of Credit Card
KKB_BNK_LMT_1	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_1	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_1	Int	The Credit Score of Customer

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
KMH_FLAG_1	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_1	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_1	Float	Cumulated Net Profit
MENKUL_ORT_TUT_1	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_1	Bit	The Flag of Investment Account
NAR_BKY_TUT_1	Float	The Amount of Nar
NAR_FLAG_1	Bit	The Flag of Nar which is a kind of Product for Retail Marketing
OTMTK_TLMT_FLAG_1	Bit	The Flag of Direct Debit
SGK_FLAG_1	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_1	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_1	Float	The Amount of Insurance Commission
SIGORTA_FLAG_1	Bit	The Flag of Insurance
TAKIP_FLAG_1	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_1	Float	The Amount of Loan Installment
TP_CALIS_TUT_1	Float	The Total Amount of Work
TP_URUN_ADET_1	Int	The Number of Total Product
TP_VARLIK_1	Float	Total Assets
TP_VDL_TUT_1	Float	Total amount of Term Deposits
VDL_ORT_TUT_1	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_1	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_1	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_1	Bit	The Flag of Demand Deposit Account
ABONE24_TUT_2	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_2	Float	The Amount of Gold Account
ALTIN_FLAG_2	Bit	The Flag of Gold Account
BK_FLAG_2	Bit	The Falg of Personal Loan
BK_ORT_TUT_2	Float	The Average Monthly Amount of Personal Loans
BORDRO24_FLAG_2	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_2	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_2	Bit	The Flag of Debit Card

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
DIALOG_ISL_TUT_2	Float	The Amount of Transactions on The Call Center
DTH_FLAG_2	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_2	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_2	Int	The Number of Loan Delay
INTRNT_ISL_TUT_2	Float	The Amount of Transactions on The Internet
KK_FLAG_2	Bit	The Flag of Credit Card
KK_LIMIT_2	Float	The Limit of Credit Card
KKB_BNK_LMT_2	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_2	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_2	Int	The Credit Score of Customer
KMH_FLAG_2	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_2	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_2	Float	Cumulated Net Profit
MENKUL_ORT_TUT_2	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_2	Bit	The Flag of Investment Account
NAR_BKY_TUT_2	Float	The Amount of Nar
NAR_FLAG_2	Bit	The Flag of Nar which is a kind of Product for Retail Marketing
OTMTK_TLMT_FLAG_2	Bit	The Flag of Direct Debit
SGK_FLAG_2	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_2	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_2	Float	The Amount of Insurance Commission
SIGORTA_FLAG_2	Bit	The Flag of Insurance
TAKIP_FLAG_2	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_2	Float	The Amount of Loan Installment
TP_CALIS_TUT_2	Float	The Total Amount of Work
TP_URUN_ADET_2	Int	The Number of Total Product
TP_VARLIK_2	Float	Total Assets
TP_VDL_TUT_2	Float	Total amount of Term Deposits

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
VDL_ORT_TUT_2	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_2	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_2	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_2	Bit	The Flag of Demand Deposit Account
ABONE24_TUT_3	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_3	Float	The Amount of Gold Account
ALTIN_FLAG_3	Bit	The Flag of Gold Account
BK_FLAG_3	Bit	The Falg of Personal Loan
BK_ORT_TUT_3	Float	The Average Monthly Amount of Personal Loans
BORDRO24_FLAG_3	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_3	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_3	Bit	The Flag of Debit Card
DIALOG_ISL_TUT_3	Float	The Amount of Transactions on The Call Center
DTH_FLAG_3	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_3	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_3	Int	The Number of Loan Delay
INTRNT_ISL_TUT_3	Float	The Amount of Transactions on The Internet
KK_FLAG_3	Bit	The Flag of Credit Card
KK_LIMIT_3	Float	The Limit of Credit Card
KKB_BNK_LMT_3	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_3	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_3	Int	The Credit Score of Customer
KMH_FLAG_3	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_3	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_3	Float	Cumulated Net Profit
MENKUL_ORT_TUT_3	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_3	Bit	The Flag of Investment Account
NAR_BKY_TUT_3	Float	The Amount of Nar
NAR_FLAG_3	Bit	The Flag of Nar which is a kind of Product for Retail Marketing

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
OTMTK_TLMT_FLAG_3	Bit	The Flag of Direct Debit
SGK_FLAG_3	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_3	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_3	Float	The Amount of Insurance Commission
SIGORTA_FLAG_3	Bit	The Flag of Insurance
TAKIP_FLAG_3	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_3	Float	The Amount of Loan Installment
TP_CALIS_TUT_3	Float	The Total Amount of Work
TP_URUN_ADET_3	Int	The Number of Total Product
TP_VARLIK_3	Float	Total Assets
TP_VDL_TUT_3	Float	Total amount of Term Deposits
VDL_ORT_TUT_3	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_3	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_3	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_3	Bit	The Flag of Demand Deposit Account
ABONE24_TUT_4	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_4	Float	The Amount of Gold Account
ALTIN_FLAG_4	Bit	The Flag of Gold Account
BK_FLAG_4	Bit	The Falg of Personal Loan
BK_ORT_TUT_4	Float	The Average Monthly Amount of Personal Loans
BORDRO24_FLAG_4	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_4	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_4	Bit	The Flag of Debit Card
DIALOG_ISL_TUT_4	Float	The Amount of Transactions on The Call Center
DTH_FLAG_4	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_4	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_4	Int	The Number of Loan Delay
INTRNT_ISL_TUT_4	Float	The Amount of Transactions on The Internet
KK_FLAG_4	Bit	The Flag of Credit Card
KK_LIMIT_4	Float	The Limit of Credit Card

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
KKB_BNK_LMT_4	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_4	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_4	Int	The Credit Score of Customer
KMH_FLAG_4	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_4	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_4	Float	Cumulated Net Profit
MENKUL_ORT_TUT_4	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_4	Bit	The Flag of Investment Account
NAR_BKY_TUT_4	Float	The Amount of Nar
NAR_FLAG_4	Bit	The Flag of Nar which is a kind of Product for Retail Marketing
OTMTK_TLMT_FLAG_4	Bit	The Flag of Direct Debit
SGK_FLAG_4	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_4	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_4	Float	The Amount of Insurance Commission
SIGORTA_FLAG_4	Bit	The Flag of Insurance
TAKIP_FLAG_4	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_4	Float	The Amount of Loan Installment
TP_CALIS_TUT_4	Float	The Total Amount of Work
TP_URUN_ADET_4	Int	The Number of Total Product
TP_VARLIK_4	Float	Total Assets
TP_VDL_TUT_4	Float	Total amount of Term Deposits
VDL_ORT_TUT_4	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_4	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_4	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_4	Bit	The Flag of Demand Deposit Account
ABONE24_TUT_5	Float	The Amount of Direct Debit Account
ALTIN_BKY_TUT_5	Float	The Amount of Gold Account
ALTIN_FLAG_5	Bit	The Flag of Gold Account
BK_FLAG_5	Bit	The Falg of Personal Loan
BK_ORT_TUT_5	Float	The Average Monthly Amount of Personal Loans

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
BORDRO24_FLAG_5	Bit	The Flag of Customer who take salary through bank
BORDRO24_MAAS_TUT_5	Float	The Amount of Salary Paid Through Bank
DBT_KART_FLAG_5	Bit	The Flag of Debit Card
DIALOG_ISL_TUT_5	Float	The Amount of Transactions on The Call Center
DTH_FLAG_5	Bit	The Flag of Foreign Exchange Deposit Account
DTH_ORT_TUT_5	Float	The Avarage Amount of Monthly Foreign Exchange Deposit Account
GCKM_TAKSIT_ADET_5	Int	The Number of Loan Delay
INTRNT_ISL_TUT_5	Float	The Amount of Transactions on The Internet
KK_FLAG_5	Bit	The Flag of Credit Card
KK_LIMIT_5	Float	The Limit of Credit Card
KKB_BNK_LMT_5	Float	The Limit of Customer in All Banks
KKB_BNK_RISK_5	Float	The Total Risk of Customer in All Banks
KKB_KRD_NOT_5	Int	The Credit Score of Customer
KMH_FLAG_5	Bit	The Flag of Credit Deposit Account
KMH_ORT_TUT_5	Float	The Average Amount of Monthly Credit Deposit Account
KUMULE_NET_KAZANC_5	Float	Cumulated Net Profit
MENKUL_ORT_TUT_5	Float	The Average Amount of Monthly Investment Account
MNKL_FLAG_5	Bit	The Flag of Investment Account
NAR_BKY_TUT_5	Float	The Amount of Nar
NAR_FLAG_5	Bit	The Flag of Nar which is a kind of Product for Retail Marketing
OTMTK_TLMT_FLAG_5	Bit	The Flag of Direct Debit
SGK_FLAG_5	Bit	The Flag Customer who is retired
SGK_MAAS_TUT_5	Float	The Amount of Salary Paid Through Bank - Retired Customer
SGRT_KOM_TUT_5	Float	The Amount of Insurance Commission
SIGORTA_FLAG_5	Bit	The Flag of Insurance
TAKIP_FLAG_5	Bit	The Flag of Customer who is insolvent
TAKSIT_TUT_5	Float	The Amount of Loan Installment
TP_CALIS_TUT_5	Float	The Total Amount of Work
TP_URUN_ADET_5	Int	The Number of Total Product
TP_VARLIK_5	Float	Total Assets
TP_VDL_TUT_5	Float	Total amount of Term Deposits

Table A.1 : The variables of data (continued).

Data Name	Value Type	Comment
VDL_ORT_TUT_5	Float	The Average Amount of Monthly Term Deposit Account
VDL_TL_FLAG_5	Bit	The Flag of Term Deposit Account
VDSZ_ORT_TUT_5	Float	The Average Amount of Monthly Demand Deposit Account
VDSZ_TL_FLAG_5	Bit	The Flag of Demand Deposit Account

Table A.2 : The analysis of data.

Parameters	Min	Max	Mean	Std. Dev
Birth Date	03.02.1900 00:00	17.11.2004 00:00	--	--
Initial Date	01.01.1900 00:00	27.11.2007 00:00	--	--
Age	6	110	52.410	11.312
Year of The Customer	3	110	14.268	6.963
sicil no	0	61286	28.481.536	21.106.714
TAKSIT_TUT_1	0.000	122.275.600	181.016	826.138
TP_VARLIK_1	0.000	164.362.892.520	46.966.493	832.864.584
TP_VDL_TUT_1	0.000	140.708.427.490	43.842.704	731.462.102
VDSZ_ORT_TUT_1	-62.660	352.121.260	542.559	4.485.665
KMH_ORT_TUT_1	0.000	304.224.060	67.369	1.524.553
VDL_ORT_TUT_1	0.000	95.752.938.910	32.520.722	525.323.038
MENKUL_ORT_TUT_1	0.000	22.778.185.460	2.760.503	114.338.392
BK_ORT_TUT_1	0.000	765.267.890	3.949.017	16.428.668
KUMULE_NET_KAZANC_1	0.000	0.000	0.000	0.000
DTH_ORT_TUT_1	0.000	36.548.332.550	12.004.939	220.720.214
SGRT_KOM_TUT_1	0.000	0.000	0.000	0.000
INTRNT_ISL_TUT_1	0.000	11.501.971.220	1.630.442	65.958.035
DIALOG_ISL_TUT_1	0.000	119.396.480	63.707	1.494.635
BORDRO24_MAAS_TUT_1	0.000	0.000	0.000	0.000
SGK_MAAS_TUT_1	0.000	0.000	0.000	0.000
ABONE24_TUT_1	0.000	0.000	0.000	0.000
TP_CALIS_TUT_1	0.000	0.000	0.000	0.000
ALTIN_BKY_TUT_1	0.000	0.000	0.000	0.000
NAR_BKY_TUT_1	0.000	0.000	0.000	0.000
TAKSIT_TUT_2	0.000	226.306.610	183.420	1.191.411
TP_VARLIK_2	0.000	167.812.870.610	46.431.008	840.663.722
TP_VDL_TUT_2	0.000	143.564.939.870	43.302.733	737.302.454
VDSZ_ORT_TUT_2	-79.670	342.982.720	556.439	4.638.657
KMH_ORT_TUT_2	0.000	139.642.830	61.422	797.146
VDL_ORT_TUT_2	0.000	96.054.268.940	32.008.130	523.446.486
MENKUL_ORT_TUT_2	0.000	23.233.737.130	2.731.675	115.353.243
BK_ORT_TUT_2	0.000	755.396.890	4.059.061	16.632.616

Table A.2 : The analysis of data (continued).

Parameters	Min	Max	Mean	Std. Dev
KUMULE_NET_KAZANC_2	-3.640.620	1.095.725.740	842.708	5.946.078
DTH_ORT_TUT_2	0.000	45.700.761.350	12.046.663	252.294.792
SGRT_KOM_TUT_2	0.000	2.276.070	12.080	43.232
INTRNT_ISL_TUT_2	0.000	11.501.971.220	1.608.403	65.436.017
DIALOG_ISL_TUT_2	0.000	119.396.480	62.921	1.482.488
BORDRO24_MAAS_TUT_2	0.000	202.521.320	164.914	1.220.768
SGK_MAAS_TUT_2	0.000	73.170.050	255.122	2.249.653
ABONE24_TUT_2	0.000	12.404.080	36.336	172.812
TP_CALIS_TUT_2	-5.529.170	164.994.893.720	49.424.856	825.809.167
ALTIN_BKY_TUT_2	0.000	383.087.200	152.698	3.329.021
NAR_BKY_TUT_2	0.000	242.269.490	39.586	1.492.537
TAKSIT_TUT_3	0.000	36.162.160	184.431	631.126
TP_VARLIK_3	0.000	169.171.718.540	45.898.736	841.280.360
TP_VDL_TUT_3	0.000	144.144.459.720	42.674.906	735.458.452
VDSZ_ORT_TUT_3	-101.540	356.603.830	580.657	4.747.499
KMH_ORT_TUT_3	0.000	76.139.650	62.115	580.906
VDL_ORT_TUT_3	0.000	98.205.666.340	31.541.611	527.828.286
MENKUL_ORT_TUT_3	0.000	23.338.576.970	2.701.970	114.902.097
BK_ORT_TUT_3	0.000	870.000.000	4.253.693	17.735.482
KUMULE_NET_KAZANC_3	0.000	0.000	0.000	0.000
DTH_ORT_TUT_3	0.000	45.814.401.560	11.840.316	250.448.742
SGRT_KOM_TUT_3	0.000	0.000	0.000	0.000
INTRNT_ISL_TUT_3	0.000	5.030.908.570	1.357.127	40.133.263
DIALOG_ISL_TUT_3	0.000	133.748.390	53.039	1.395.914
BORDRO24_MAAS_TUT_3	0.000	0.000	0.000	0.000
SGK_MAAS_TUT_3	0.000	0.000	0.000	0.000
ABONE24_TUT_3	0.000	0.000	0.000	0.000
TP_CALIS_TUT_3	0.000	0.000	0.000	0.000
ALTIN_BKY_TUT_3	0.000	0.000	0.000	0.000
NAR_BKY_TUT_3	0.000	0.000	0.000	0.000
TAKSIT_TUT_4	0.000	108.577.890	188.679	818.922
TP_VARLIK_4	0.000	191.269.655.640	45.317.275	918.091.657

Table A.2 : The analysis of data (continued).

Parameters	Min	Max	Mean	Std. Dev
TP_VDL_TUT_4	0.000	166.799.011.990	42.243.808	812.659.249
VDSZ_ORT_TUT_4	-73.880	384.947.770	618.083	4.966.320
KMH_ORT_TUT_4	0.000	106.107.550	68.070	819.692
VDL_ORT_TUT_4	0.000	99.653.522.470	31.309.483	530.119.003
MENKUL_ORT_TUT_4	0.000	22.993.767.620	2.646.183	112.716.679
BK_ORT_TUT_4	0.000	895.796.640	4.442.097	18.681.239
KUMULE_NET_KAZANC_4	-3.640.600	1.243.433.100	961.949	6.621.894
DTH_ORT_TUT_4	0.000	48.672.084.400	11.502.431	258.573.351
SGRT_KOM_TUT_4	0.000	2.583.330	13.321	47.220
INTRNT_ISL_TUT_4	0.000	12.270.809.700	1.778.479	70.768.815
DIALOG_ISL_TUT_4	0.000	472.002.150	65.056	2.554.633
BORDRO24_MAAS_TUT_4	0.000	398.864.760	195.346	2.022.881
SGK_MAAS_TUT_4	0.000	66.883.960	130.603	1.420.830
ABONE24_TUT_4	0.000	12.173.720	43.814	191.502
TP_CALIS_TUT_4	-1.499.500	171.322.063.110	48.486.058	838.520.356
ALTIN_BKY_TUT_4	0.000	420.732.000	231.462	4.571.281
NAR_BKY_TUT_4	0.000	296.447.760	53.274	1.910.089
TAKSIT_TUT_5	0.000	108.577.890	188.994	796.915
TP_VARLIK_5	0.000	191.569.794.180	44.314.144	908.813.941
TP_VDL_TUT_5	0.000	167.644.576.830	41.347.289	806.155.750
VDSZ_ORT_TUT_5	-586.390	410.666.410	605.418	4.798.216
KMH_ORT_TUT_5	0.000	190.563.470	67.850	1.101.719
VDL_ORT_TUT_5	0.000	100.521.671.100	30.664.054	519.667.665
MENKUL_ORT_TUT_5	0.000	22.538.401.410	2.559.275	109.403.719
BK_ORT_TUT_5	0.000	891.700.040	4.541.385	18.687.654
KUMULE_NET_KAZANC_5	-564.150	28.804.810	66.039	283.923
DTH_ORT_TUT_5	0.000	66.792.062.000	11.367.794	320.514.383
SGRT_KOM_TUT_5	0.000	2.583.330	13.741	47.819
INTRNT_ISL_TUT_5	0.000	29.100.622.500	1.968.100	132.453.211
DIALOG_ISL_TUT_5	0.000	204.400.100	49.054	1.386.483
BORDRO24_MAAS_TUT_5	0.000	201.355.980	200.278	1.254.543
SGK_MAAS_TUT_5	0.000	77.698.960	200.519	1.291.657
ABONE24_TUT_5	0.000	9.468.200	44.428	190.282

Table A.2 : The analysis of data (continued).

Parameters	Min	Max	Mean	Std. Dev
TP_URUN_ADET_5	--	--	--	--
TP_CALIS_TUT_5	-3.277.010	189.854.857.420	47.932.465	898.619.890
ALTIN_BKY_TUT_5	0.000	412.883.730	226.535	4.443.716
NAR_FLAG_5	--	--	--	--
NAR_BKY_TUT_5	0.000	260.563.260	54.155	1.718.943

C.V.

Name Surname	: Ümit Şahin
Address	: Talatpaşa cd. Ziya Paşa sk.. Kaya Apt. No:13 D:7 Bahçelievler / İstanbul
Place of Birth, Year of Birth	: Eminönü, 1985
Foreign Language	: English
Primary School	: KADIRGA İLKÖĞRETİM OKULU, 1996
Middle School	: ÇAPA ATATÜRK İLKÖĞRETİM OKULU, 1999
Highschool	: VATAN ANADOLU LİSESİ, 2003
University	: BAHÇEŞEHİR ÜNİVERSİTESİ, 2008
Graduate School	: BAHÇEŞEHİR ÜNİVERSİTESİ, 2011
Institute	: Graduate School of Natural and Applied Sciences
Graduate Program	: M.S. Program Computer Engineering
Experience	: T. HALKBANK A.Ş., 2008 – (continued)