T.C.
MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN
PURE AND APPLIED SCIENCES

AN APPLICATION OF DATA MINING STUDY; A WAY TO IMPROVE BUSINESS STRATEGIES OF THE COMPANY

Emre TACENUR

(Industrial Engineering)

THESIS

FOR THE DEGREE OF MASTER OF SCIENCE

IN

ENGINEERING MANAGEMENT

SUPERVISORS
Assist.Prof.Dr. Arzu BALOĞLU
Assist.Prof.Dr. Bahar SENNAROĞLU

İSTANBUL 2006

T.C.

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN
PURE AND APPLIED SCIENCES

AN APPLICATION OF DATA MINING STUDY; A WAY TO IMPROVE BUSINESS STRATEGIES OF THE COMPANY

Emre TACENUR

(Industrial Engineer)

(141102419970086)

THESIS

FOR THE DEGREE OF MASTER OF SCIENCE

IN

ENGINEERING MANAGEMENT

SUPERVISORS
Assist.Prof.Dr. Arzu BALOĞLU
Assist.Prof.Dr. Bahar SENNAROĞLU

İSTANBUL 2006

ACKNOWLEDGMENTS

I would like to thank my supervisors Assist. Prof. Dr. Arzu Baloğlu and Assist. Prof. Dr. Bahar Sennaroğlu who encouraged, guided and motivated me during this study.

I would like to extend my thanks to my colleagues who helped in every stage of thesis guiding me ideas from business side and friends supporting me with sharing their tools and vision.

Special thanks are for my wife and parents for endless patience, love, support and encouragement from the beginning of this long story.

September 2006 **Emre TACENUR**

TABLE OF CONTENTS

ACKNOWLEDGMENTS……I

TABLE OF CONTENTS……II

ÖZET……V

ABSTRACT……VII

ABBREVIATIONS……IX

LIST OF FIGURES……X

LIST OF TABLES……XII

I INTRODUCTION AND OBJECTIVES..... 1

I.1 Why Data Mining?..... 1

II GENERAL BACKGROUND..... 4

II.1 Knowledge Discovery in Databases..... 4
II.2 What is Data Mining?..... 5
II.3 Data Warehousing and Data Mining..... 6
II.4 Data Mining as a Process..... 8
II.4.1 Business Understanding Phase..... 11
II.4.2 Data Understanding Phase..... 12
II.4.3 Data Preparation Phase..... 12
II.4.4 Modeling Phase..... 12
II.4.5 Evaluation Phase..... 13
II.4.6 Deployment Phase..... 13
II.5 Data Mining Methods and Techniques..... 13
II.5.1 Data Mining Methods..... 13
II.5.1.1 Prediction..... 14
II.5.1.2 Profiling-Pattern Description..... 15

II.5.2 Data Mining Techniques ... 17
II.5.2.1 Decision Trees ... 17
II.5.2.2 Neural Networks ... 18
II.5.2.3 Nearest Neighbor Techniques ... 21
II.5.2.4 Regression ... 23
II.5.2.5 Clustering ... 25
II.5.2.6 Association Discovery ... 29
II.5.2.7 Genetic Algorithms ... 31
II.5.2.8 Distance Evaluation ... 32
II.5.3 Visualization & Evaluation ... 33
II.6 Software ... 34
II.7 About Private Health Insurance ... 35
II.8 Insurance and Data Mining ... 38
II.9 Data Mining Techniques in Insurance ... 46
II.10 General Assumptions and Difficulties ... 49

III ANALYSIS ... 52

III.1 Business Understanding ... 52
III.1.1 Business Objective ... 52
III.1.2 Selected Tools ... 54
III.1.3 Problem Definition ... 54
III.1.4 Beginning Strategy ... 56
III.2 Data Understanding ... 57
III.2.1 Selection of Appropriate Data ... 57
III.2.2 Exploring the Data ... 58
III.2.2.1 Policy Number ... 58
III.2.2.2 Beginning Year ... 58
III.2.2.3 Product ... 59
III.2.2.4 Renewal Years ... 61
III.2.2.5 Customer ID ... 61
III.2.2.6 Sales Region ... 61
III.2.2.7 Sales Channel ... 63
III.2.2.8 Age ... 64
III.2.2.9 Gender/Sex ... 65
III.2.2.10 Marital Status ... 66
III.2.2.11 Property/Casualty (P/C) Insurance ... 68
III.2.2.12 Location ... 69
III.2.2.13 Occupation ... 70
III.2.2.14 VIP ... 71
III.2.2.15 Number of Insureds ... 71
III.2.2.16 Premium ... 72
III.2.2.17 Payment term ... 72
III.2.2.18 Payment Method ... 72
III.2.2.19 Number of Claims ... 74
III.2.2.20 Claims ... 74
III.2.2.21 Medical Information ... 75
III.3 Data Preparation ... 75
III.3.1 Model Set ... 75
III.3.2 Data Cleaning ... 76
III.3.3 Data Transformation ... 77
III.4 Modeling ... 80
III.4.1 Technique Selection ... 80
III.4.2 Building Model ... 83

III.4.3 Further Modeling 89
III.4.4 Test and Validation 95

IV RESULTS 96

IV.1 Current Business 96
IV.2 Results and New Strategies 96
IV.2.1 Interpreting Results 97
IV.2.1.1 Model 1 97
IV.2.1.2 Model 3 99
IV.2.2 Strategies 100

V EVALUATION 104

VI REFERENCES 107

VII APPENDICES 111

RESUME 165

ÖZET

İŞ STRATEJİLERİNİ GELİŞTİRMENİN BİR YÖNTEMİ; VERİ MADENCİLİĞİ UYGULAMASI

Pazarda yaşanan yoğun rekabet şirketleri yeni arayışlara iterken sigorta sektörünün de içinde bulunduğu bu pazarlarda ayakta kalabilmek, daha iyi stratejiler üretebilmeyi gerektirmektedir. Hızlı teknolojik gelişim veri saklama maliyetini düşürürken, bu yıllarda şirketler depolanan veriyi pazarda rekabet avantajı sağlayacak bilgiye dönüştürmeye çalışmaktadırlar.

Veri madenciliği büyük veri madenleri içinden bilgiyi yüzeye çıkartmak için hazırlanmış teknikler kümesidir. Kullanılan pek çok istatistik ve yapay zeka araçlarının ötesinde, veri madenciliğinin süreç yaklaşımı sistematik çözümler üretilmesini mümkün kılmaktadır. Bu çalışmada veri madenciliği teknik ve metodolojisi ile bir sigorta şirketine ait gerçek veriden yüksek rekabetin yaşandığı bu sektörde strateji oluşturma ve karar verme süreçlerine girdi sağlayacak bilgi elde edilmeye çalışılmıştır.

Bu çalışmada karlı müşteri profili oluşturmak amacıyla veri madenciliği süreç yaklaşımı ve C5.0 karar ağacı algoritması kullanılmıştır. Dengesiz dağılıma sahip veri kümesi sorunu katsayılarla yeniden örnekleme, K-Means algoritması aracılığıyla ön kümeleme uygulaması ile yeniden örnekleme, birleştirme ve basitleştirme yöntemleri kullanılarak çözülmeye çalışılmıştır. Basitleştirme ve katsayılarla yeniden örnekleme ile daha iyi sonuçlar elde edilmiş, doğru tespit oranı umulan seviyede olmasa da hata oranları tatmin edici seviyede çıkmıştır.

Uygulaması gerçekleştirilen sınıflama çalışması bilgi kaynaklı strateji üretimi için yalnızca bir başlangıçtır. Karlı müşteri profili değerli müşterileri çekmek için oluşturulacak stratejilerde önemli bir girdi olacaktır. Yeni kampanyalar, özel indirimler, kişiye özel hizmetler karlı müşteri bilgisi ile alınabilecek bazı kararlardır. Yüksek cirolu müşterileri tanımlamak, çapraz satış için ilişki analizleri yapmak, tahmin analizleri karar vericilere daha fazla bilgi sağlamak amacıyla daha sonra gerçekleştirilebilecek bazı analizlerdir. Sınıflama tekniği veri değişikliklerine duyarlı olduğundan uygulaması yapılan sınıflama analizinin de dönemsel olarak tekrarlanması gerekmektedir. Daha iyi iş sonuçları elde edebilmek için analiz sonuçları güncel tutulmalıdır.

Anahtar Kelimeler: Veri Madenciliği, Sınıflama, Sigortacılık, İş Stratejileri

Eylül 2006 **Emre TACENUR**

ABSTRACT

AN APPLICATION OF DATA MINING STUDY; A WAY TO IMPROVE BUSINESS STRATEGIES OF THE COMPANY

Hard conditions in the market lead the companies to find new ways to compete better. Better strategies are needed to stand strong in these highly competitive markets one of which is insurance market. The rapid technological development reduced the cost of storing data and in these decades every company is trying to convert this data to a competitive value in the market.

Data mining is a set of techniques to bring information from large data sets to the surface. Beyond many statistical and AI techniques are used in data mining, process approach in method makes DM a systematic solution. In this study, it is intended to use data mining methods to derive conclusions from a large set of real insurance company data to be used strategy setting and decision making process in highly competitive market.

An application of data mining process approach and C5.0 decision tree algorithm is performed in the study to determine the profile of profitable customers. The problem of unbalanced data is tried to be solved by pre-clustering application in re-sampling and combining and simplifying attributes. Better results were obtained by simplifying although the accuracy of detecting is not as expected, the error rate is satisfactory.

The classification technique application in this thesis is only a beginning analysis for information based strategy generation. The profitable customer profile

is an input for the business strategies in attracting valuable customers. New campaigns, special discounts, customer focused services are decisions to be taken with the information of profitable customer. Further analysis of data to describe high revenue customers, association analysis for cross-sales, predictive analysis can be performed to provide more information for decision makers. Moreover, classification analysis made should be performed periodically since method is sensitive to data changes. The results should be kept up to date for better results.

Key Words: Data Mining, Classification, Insurance, Business Strategies

September 2006 **Emre TACENUR**

ABBREVIATIONS

DM **:** Data Mining

KDD **:** Knowledge Discovery in Databases

P/C **:** Property and Casualty Insurance

AI **:** Artificial Intelligence

CID **:** Customer Identification Number

TIN **:** Tax Identification Number

LIST OF FIGURES

Figure II-1 Simple Diagram of Data Mining 5
Figure II-2 Data Warehousing and Data Mining 7
Figure II-3 CRISP-DM Data Mining Process Cycle 11
Figure II-4 Sample Decision Tree 18
Figure II-5 An example of Neural Network Diagram for Multilayer Perceptron .. 20
Figure II-6 Groups of Similar Records 22
Figure II-7 Growth of Health Insurance in US 36
Figure II-8 Lift Curve of Data Mining Project for Mailing Campaign 49
Figure III-1 Phases and Tools Utilized 54
Figure III-2 Distribution of Policies by Beginning Year 59
Figure III-3 Number of Policies Sold by Product 60
Figure III-4 Annual Number of Policies Sold by Products 60
Figure III-5 Number of Policies Sold by Renewal Year 61
Figure III-6 Annual Distribution of Policies Sold by Regions 62
Figure III-7 Annual Distribution of Policies Sold by Regions (Changed) 63
Figure III-8 Distribution of Policies by Sales Channel 64
Figure III-9 Number of Policies by Age 64
Figure III-10 Number of Policies by Age (Product Limited) 65
Figure III-11 Distribution of Policies by Sex 66
Figure III-12 Distribution of Policies by Marital Status 67
Figure III-13 Number of Insureds per Policy by Marital Status 67
Figure III-14 Distribution of Policies by P/C Cross-sales 69

Figure III-15 Distribution of Clients by P/C Cross-sales 69
Figure III-16 Distribution of Policies by City 70
Figure III-17 Distribution of Policies by Occupation 71
Figure III-18 Distribution of Policies by Number of Insureds 71
Figure III-19 Distribution of Number of Installments of Policies 72
Figure III-20 Number of Payment Methods Used per Policy 73
Figure III-21 Distribution of Payment Methods by Policies 74
Figure III-22 Distribution of Number of Claims per Policy 74
Figure III-23 Cumulative Claims vs No of Policies by Claims Amount per Policy 79
Figure III-24 Cumulative Premium vs No of Policies by Premium Amount per Policy 79
Figure III-25 C4.5 vs C5.0 Algorithm 81
Figure III-26 Tree Presentation 85
Figure III-27 Clementine Model Diagram 94

LIST OF TABLES

Table III-1 Customer Life Cycle 53
Table III-2 Model Set Attributes 75
Table III-3 Coincidence Matrix 86
Table III-4 Balancing 86
Table III-5 Coincidance Matrix (Balanced Set) 87
Table III-6 Misclassification Cost of [1->3] and [3->1]=2, Balanced C5.0 Coincidance Matrix 87
Table III-7 Misclassification Cost of [1->3] and [3->1]=1.5, Balanced C5.0 Coincidance Matrix 88
Table III-8 Misclassification Cost of [1->3] and [3->1]=1, Balanced C5.0 Coincidance Matrix 88
Table III-9 Rates for Different Misclassification Costs 88
Table III-10 Subclass Clusters 91
Table III-11 Rates for Different Misclassification Costs (2nd model) 92
Table III-12 Rates for Different Misclassification Costs (2nd model-Merged) 92
Table III-13 Rates for Different Misclassification Costs (3rd model) 94
Table III-14 Performance Rates of Models (Test Set) 95

I INTRODUCTION AND OBJECTIVES

Market forces such as mergers, globalization and intensifying competition are radically altering the financial services landscape. Leading financial services firms are prospering during this turbulent time by taking precautions on expenses. Sales expense is a major cost item of the most companies, however it's a hard decision to cut expenses due to possible unexpected sales drops. Hence, increasing the efficiency of sales channels by applying effective strategies becomes crucial in high competition financial markets. As a consequence, industries such as banking and insurance are involved in customer behavior analysis to reduce costs, enhance research and increase sales. Predictive and descriptive analyses make it possible to leverage the distinct multi-channel nature of the financial services business to gain new insights into customer behavior and preferences and companies can acquire the right customers cost-effectively. As a result, learning from data and set strategies accordingly become crucial for the companies to be successful and data mining techniques are increasingly used for this purpose.

I.1 Why Data Mining?

In order to form a learning relationship with its customers, a company must be able to notice what customers are doing, remember what they and company itself have done over time, learn from this experience and act accordingly for better business decisions and actions.

Data mining is predicted to be "one of the most revolutionary developments of this decade," according to the online technology magazine ZDNET News. [1] In fact, the MIT Technology Review [2] chose data mining as one of ten emerging technologies that will change the world. Implementation of data mining techniques has seen as a revolutionary change in doing business and research.

The data mining has some advantages to be mentioned as revolution and being preferred to learn from data. Data mining, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, represents a difference of kind rather than degree. Many analytical tools utilize a verification-based approach, where a hypothesis developed and then the data tested to prove or disprove the hypothesis. For example, a user might hypothesize that a customer who buys a Motor insurance, will also buy Compulsory Third Party Liability Motor insurance. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In other words, this approach relies on a number of factors, including the ability of the analyst to ask appropriate questions and quickly return results, manage the complexity of the attribute space and think different. In contrast, in addition to verification based statistical methods, data mining can utilize a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, an insurance company may compare their customers' insurance purchases with home ownership, type of automobile driven, age, occupation, income and geographical zone. [3][4] Data mining shelters many methods from statistics, mining and Artificial Intelligence to use the perfect method to the appropriate problem and data.

Advances in technology and business processes have contributed to a growing interest in data mining. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques called Artificial Intelligence such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from distant terminals and an increased ability to combine data from disparate sources into a single searchable source. In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing have also played a role. Over the past several years, there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year. Wal-Mart, the largest retailer in the

world, for example, had a customer database that contains around 43 terra-bytes of data at the end of last decade which is larger than the database used by the U.S. Internal Revenue Services for collecting income taxes. [5] At the same time, the costs of data storage have decreased significantly. Similarly, computing power has continued to almost double in a couple of years, while the relative cost of computing power has continued to decrease.

The availability of various information and huge numbers of data forces the market to get use of the data as much as possible and triggering the need of mining these data by better techniques and better tools. Although data mining has emerged as visible research and development area in the 1990's, rapid growth brought the data mining to be one of the revolutionary changes in the next decade. [6]

The trend closes the gap between analytic software developers like SPSS and SAS and CRM software vendors, and analytic CRM becomes wide spread. The developments in data mining leads better algorithms, standardization and integration of systems. [7]

II GENERAL BACKGROUND

II.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases, in some literature, is separated from data mining. The discovery of knowledge from large databases is a process instead of just an operation. The process, which will be explained in detail in the coming chapters, begins with business understanding and problem setting to data understanding and cleaning then modeling, evaluation and deployment. In some literature data mining is limited to model phase of the process and the whole process is called knowledge discovery in databases process. However, generally data mining and KDD are used as synonyms covering all the process. [8]

The knowledge discovery evolution begins in 1960's by increasing data collection. In time, reaching data becomes easier and data warehousing and decision support systems gains popularity in 1990's leading sophisticated data mining applications in last decade. Below mentioned table by Thearling summarizes the evolution of knowledge discovery. [9]

Table II-1 Knowledge Discovery Evolution

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Warehousing and Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (2000's)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, SAS, SPSS and many others	Prospective, proactive information delivery

II.2 What is Data Mining?

Data Mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. [10] In other words, data mining is a business process for maximizing the value of data collected by the business.



Figure II-1 Simple Diagram of Data Mining

The evaluation of technology and business is expressed generating forests of data until now. The data gathered is so large that digging information from this mine needs sophisticated methods and tools. Data mining is the set of tools to dig the mine of data. In short as shown at the figure above, data that is prepared to analyze is put into the model and prediction or pattern that is used to build strategies is got. Data mining is an iterative learning process and requires long-term hard work and commitment. As everything in life, data mining needs effort but the successful result transforms the company from being reactive to being proactive. More detailed information about the data mining process and techniques will be given in the further titles.

Data Mining provides ways to quantitatively measure what business users know or should know qualitatively. Data mining is a collection of powerful techniques, not a simple to use, one button tool. There is no single data mining

approach, but rather a set of techniques that can be used stand alone or in combination with each other.

Data mining does not replace traditional statistical techniques. Rather than replacing, it is an extension of statistical methods getting help from Artificial Intelligence methods. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed however, the technological evolution emerged the need to analyze enormous data sets with millions of rows and forced the development of new techniques based on exploration of possible solutions. New techniques like neural nets and new approaches to old statistical methods like discriminant analysis changed the efficiency of analysis. [11]

Although data mining is a powerful tool, it requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related, rather than technology-related. Data mining can help to find out patterns and relationships, it does not tell the value or significance of these patterns. The analyst should determine and comment on the results. The results must be combined with the real life observations.

II.3 Data Warehousing and Data Mining

Data to be used in the analysis are generally located on several systems, databases, formats and structures and sometimes on a redundant place. To turn these data to a value, accessing the data is the first phase. The importance of data warehouse comes out to stage at this moment. The data warehouse having combined, organized, cleaned, standardized data is a better source for analysis than operational databases or regular files spread in the systems. Identification of hidden information in data requires; departmental data is incorporated in one store, data is organized to simplify decision making and data analysis and mining for valuable information. [12]

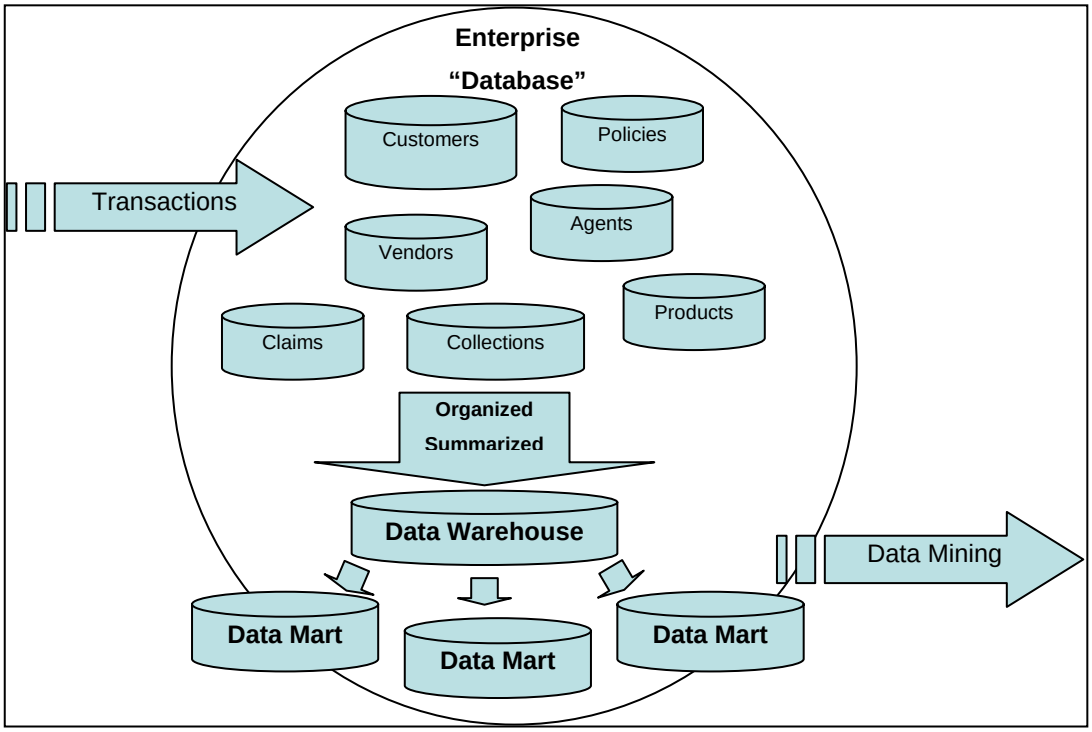

Figure II-2 Data Warehousing and Data Mining

A data warehouse is a copy of transaction data specifically structured for querying, analysis and reporting, hence data mining. Since the data warehouse contains a copy of the transactions which are not updated or changed later by the transaction system, on time data is not available. Time lag between online system and data warehouse changes from the structure of data to decisions of update times. However, since the data is specially structured and may have been cleaned and transformed when it was copied, it is much faster and easier to work on. In the next chapter, process of the data mining is covered and it is going to be mentioned that one of the most important and time consuming phase of data mining analysis is data preparation. The importance of data warehouse and data marts could be better understood on this point of view.

A Data Mart is one step ahead after a data warehouse, a smaller and more focused Data Warehouse, in other words it can be called a mini-warehouse. A Data Mart typically reflects the business rules of a specific business unit within an enterprise. As a result, small, easy-to-manage and subject-focused data marts decrease the data retrieval and analysis time dramatically.

Data Warehouses and Data Marts increase the power and efficiency of the companies' Business Intelligence capabilities by supporting queries and data

mining. Web enabling of these applications makes them more user-friendly. The potential benefits greatly outweigh the costs. Data warehouse/data mart implementation streamlines information delivery for decision support and significantly simplifies development of descriptive and predictive models that have become more popular within actuaries and statisticians in recent years.

A successful implementation of data warehouse system in a major insurance company is analyzed in the CASACT (The Casualty Actuarial Society) paper. [13] According to the results of the implementation, data warehouses, data marts and predictive analytics are essential components of a Business Intelligence system. The data warehouse enables efficient separation of historical data used for analysis from transactional databases. This increases the data access capability significantly and also reduces the load on transactional systems. The data warehouse must be well planned for the organization to realize the expected analytical efficiencies. OLAP is for exploratory data analysis but deeper analysis requires use of specialized tools for data mining. Data mining is different from OLAP because rather than verify hypothetical patterns, it uses the data itself to uncover such patterns. Although OLAP is a deductive process, data mining is an inductive process. To illustrate, when the risk factors for an insured is subject to data mining analysis, data mining tool might discover that people with high claim records and age were bad insurance risks but it might go further and also discover a pattern the analyst did not think to try, such as that geographical territory is also a determinant of risk. [11]

To conclude, although company data warehouses and data marts are not musts for a data mining analysis, they considerably simplify the process. If the company does not have a data warehouse or data mart, the process of data mining includes the preparation of data set which is alike to form a small data mining data mart.

II.4 Data Mining as a Process

Although software tools are the key to data mining, they are only one part of a data mining solution. A benchmark survey for application of data mining realized by Price Waterhouse found that process is far more important than technology. [14] Among the 159 companies that participated in the survey, those with good business processes were more than twice as effective as those with weak processes. Using

more sophisticated tools help companies to be more effective to a limit. Good tools with a weak process do not lead to success.

Some of the details of survey for successful process are listed below:

- Having the best possible data: More than 90% of companies devoting adequate resources to data cleaning were effective.
- Acting quickly on results
- Having a continuous improvement cycle: Companies having were twice as effective.
- Building a team that includes strong knowledge of a variety of disciplines-in particular the product and market, information systems, and statistics or econometrics: Companies with broad and deep skills were four times as effective as those with weak teams.
- Selecting a software tool that facilitates the entire data mining process-including accessing and cleaning data, making predictions, interpreting and acting on results, and continuous improvement.

Data mining is a discipline to work on rather than a product that can be bought. The software affects the success but an effective process and effective man-work brings the success. Beginning by the problem definition to the interpretations of results and turning them to meaningful information for the business and taking right actions generates the value.

The complex capabilities of data mining causes two precursors to be important for a successful data mining exercise; a clear formulation of the problem to be solved and access to the relevant data. Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases. Other steps in the knowledge discovery in databases process, include resource identification, data gathering, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. [4] In this study as generally accepted in literature, it is accepted that data mining covers the whole knowledge discovery process.

Departmental structure and having different focuses in the divisions lead duplicate efforts and loss of experience in data mining projects. Need of a

systematic approach and even more, a cross-industry standard emerged to make large data mining projects, less costly, more efficient, more reliable, more repeatable, more manageable and faster. The Cross-Industry Standard Process for Data Mining (CRISP–DM) was developed in 1996 by analysts representing Daimler Chrysler AG, SPSS and NCR. The methodology reached maturity in 1999 with an insurance sector partner OHRA to run trials and by partial sponsorship of the European Commission under the ESPRIT program, and CRISP-DM version 1.0 was published. CRISP-DM provides an easily available and non-commercial standard process for fitting data mining into the general problem-solving strategy of a business or research unit. Although many different approaches can be used, CRISP-DM became a generally accepted method. In a survey conducted by Kdnuggets in 2004, 42% of the data miners stated that they use CRISP-DM methodology whereas 28% use their own approach and 10% use SEMMA which is a methodology built by SAS. [15]

In general, as it can be seen in the Figure II-3 CRISP-DM Data Mining Process Cycle, CRISP–DM consists of 6 phases. The process is continuous and iterative as symbolized by the outer circle. Each phase depends on the output of its predecessor. The most significant dependencies between phases are indicated by arrows in the figure. Since the process is iterative, the sequence of the phases is not constant, it is always possible to return back to the previous phase or more than one phase for further refinement before moving forward. The process of data mining goes on like a snowball effect that each business or data mining problem analyzed bring new questions to be analyzed. [16] [17]

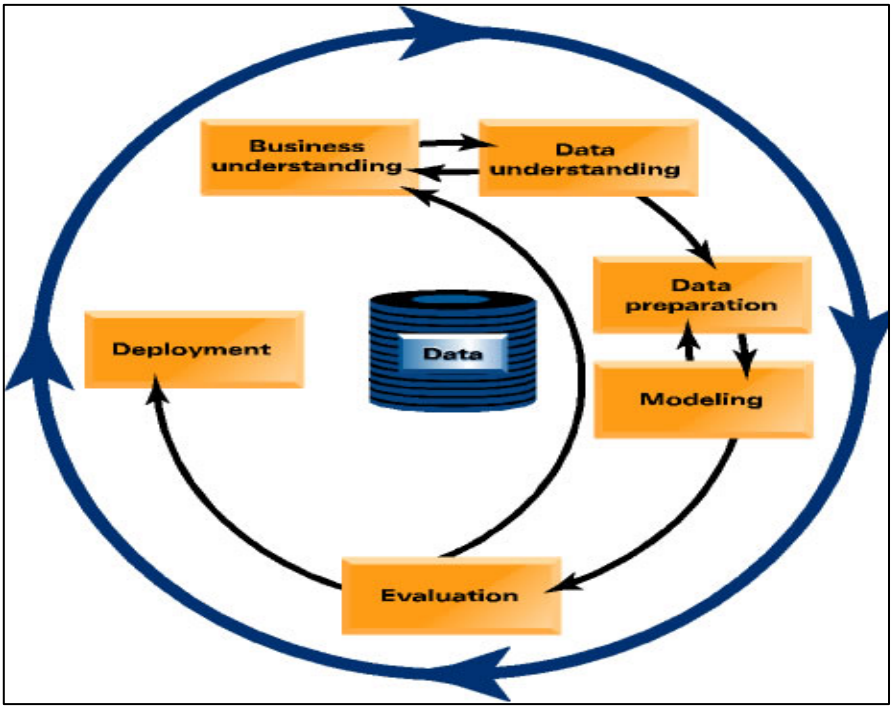

Figure II-3 CRISP-DM Data Mining Process Cycle

The methodology has a hierarchical structure under these 6 main phases. The phases are broken down to general tasks that are supposed to cover all possible situations called Generic tasks. (e.g. Data cleaning under Data preparation phase) The third level named Specialized task level is a more detailed level where the ways how the generic task is carried out are explained. (e.g. Missing value handling under Data cleaning task) The fourth and the last level called the process instance level represents what actually happened in a particular specialize task level. (e.g. The mean value for numeric attributes and the most frequent for categorical attributes were used)

A possible outline drilled to generic tasks hierarchical level can be summarized as follows by the contribution of CRISP-DM [16] [17] [19] and Berry and Linoff's methodology. [10]

II.4.1 Business Understanding Phase

The Business understanding phase covers understanding the project objectives and requirements from the business point of view. The process begins with a well-defined business problem, conversion to data mining problem and preliminary plan.

- Business objective

- Translation of the goal to data mining problem definition
- Beginning strategy to reach objectives

II.4.2 Data Understanding Phase

The data understanding phase is the phase to begin working on the data. Finding and selection of data, activities to get familiar to data, discovering first insights of the data are some activities in this phase.

- Select Appropriate Data
- Describe data
- Explore the Data (Extraction)

Data understanding and preparation phases are the most laborious part of a data mining study. Building the data-mining database, exploring the data and preparing it for modeling are the most time-consuming. For example, in a web click stream analysis data, these phases constitute approximately 80% to 95% of a project time and resource allocation. [18]

II.4.3 Data Preparation Phase

The data preparation phase covers tasks including table, record and attribute selection as well as transformation and cleaning of data for modeling tools. These tasks are carried out to prepare the data to be used in models.

- Create a Model Set
- Data Cleaning
- Data Transformation (selection, integration, formatting)

II.4.4 Modeling Phase

Modeling phase is the phase of selection and application of data mining techniques. If the data set does not cover the requirements of the selected techniques going back to preparation phase may be needed.

- Select techniques
- Build Models
- Test models
- Assess models

II.4.5 Evaluation Phase

It is important to more thoroughly evaluate the model and review the steps executed to construct the model to be certain that it properly achieves the business objectives. The objective is to check if there is some important business issue that has not been sufficiently considered. A decision on the use of the data mining results should be reached at the end.

- Review the process
- Determine whether objective is achieved

II.4.6 Deployment Phase

This phase is the conversion of the model to real life to get the advantage. Depending on the requirements, the deployment phase can be as simple as a final report or as complex as implementing a repeatable data mining process across the enterprise. Generally analyzer gives the results and functional departments carry out the deployment. However, even if the analyst will not carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

- Final report
- Plan deployment
- Plan monitoring

II.5 Data Mining Methods and Techniques

II.5.1 Data Mining Methods

In general it is possible to categorize the techniques into two groups; Directed and Undirected. Directed data mining tries to explain or categorize some particular target field such as income or response. On the other hand, undirected data mining attempts to find patterns or similarities among groups of records without the use of a particular target field or collection of predefined classes.

In other words as expressed in many literatures, Profiling-Description describes what is in the data using demographic variables and tries to explain past.

Cause and effect relations are not concerned, only patterns in the past are determined. However, Prediction tries to find patterns in data from prior periods that are capable of explaining outcomes in a later period. Predictive models require separation in time between the model inputs and output.

Under these two major categories, the types of data mining tasks can be grouped as follows:

Predictive models:

- Classification: segmentation with predefined classes e.g. 1st year students, 2nd year students and 3rd year students
- Numeric Prediction/Regression: e.g. predict claims amount

Descriptive-Pattern description models:

- Link Analysis: e.g. people who buy hamburger, often buy fries with a probability of 90%
- Clustering: segmentation without predefined classes
- Deviation Detection: e.g. detection of a fraud in medical claims in insurance company

A data mining application typically requires the combination of two or more tasks. On the following topics more detail will be given about these data mining tasks. [10][11][20][21]

II.5.1.1 Prediction

II.5.1.1.1 Classification

The goal of classification is to develop a model that maps a data item into one of several predefined classes. Once developed, the model is used to classify a new instance into one of the classes. Examples include the classification of bankruptcy patterns based on the financial ratios of a firm and of customer buying patterns based on demographic information to target the advertising and sales of a firm effectively toward the appropriate customer base.

Decision tree is a popular and useful technique for classification. This powerful technique is useful to explore data to gain insight in relationships of a large number of candidate input variables to reach an output variable, since it

divides the complex data sets into smaller subsets by applying simple decision rules. The subsets become more homogeneous to be analyzed easily.

Some techniques used in predictive analysis are memory-based, instance-based methods (K-Nearest Neighbor), Naïve-Bayes, Neural Networks, Genetic Algorithms, Factor and Discriminant analysis. [10][21][22]

II.5.1.1.2 Regression

Prediction of numeric continuous variables is generally used with regression name due to the widely usage of the technique. The most commonly used predictive method for building models is regression. The model built by regression uses input to predict a variable. Models have traditionally been developed using statistical methods such as linear, polynomial and logistic regression. Prediction of claims using input variables like age, sex, location, marital status, medical condition etc. can be an example of regression.

Although both classification and regression are used for prediction, the difference between is while the output of classification is categorical, that of regression is numeric and continuous. Neural Networks and Induction algorithms also serve for that purpose. [10][20][21]

II.5.1.2 Profiling-Pattern Description

II.5.1.2.1 Link Analysis

Link analysis, also known as affinity grouping, is used to establish relationships between database records. The link analysis can be divided into two by the time bar concept, association and sequence discovery. Association discovery finds rules about items that appear together in an event such as a purchase transaction. The typical application, also its name sometimes overrides the name of link analysis, is market-basket analysis. This technique is applied to transaction data of sales at a time to identify product affinities. All retail stores are interested in which products are sold together to arrange shelf and location designs. (e.g. customer who buys cigarette also buys matches) An application in insurance is to find relationships among medical procedures by analyzing claim forms that are submitted. The other link analysis type is sequence discovery which defines an

association related over time. As an example, sale of a barbeque triggers sales of meat in the next transactions in a retail store.

Link analysis is often applied in conjunction with clustering. The method is based on the graph theory. While nodes are representing the things that have relationships, edges are nodes connected by a relationship. A good example is search engines where nodes are web pages while edges are hyperlinks.

Some examples of approaches to Link analysis are Association and sequence discovery algorithms, genetic algorithms and neural networks. [10][11][20][21]

II.5.1.2.2 Clustering

The goal is to identify clusters of records that exhibit similar/different behaviors or characteristics in the data. Unlike classification, predefined clusters does not exist, similar records are grouped by the algorithm. Moreover, at the beginning, it is not known by which attributes the data will be clustered. Consequently, experienced businessman interpretation of the clusters is needed.

The clusters may be mutually exclusive and exhaustive or may consist of a richer representation such as hierarchical or overlapping categories. Examples include discovering homogenous groups of consumers in marketing databases and segmenting the records that describe sales during "Mother's Day" and "Father's Day." Once the database is segmented, link analysis is often performed on each segment to identify the association among the records in each cluster. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means. K-Means, is a technique similar to nearest neighbor techniques (memory-based-reasoning and collaborative filtering) depending on a geometric interpretation of the data. Some other algorithms include Gaussian mixture models, Agglomerative clustering, Divisive clustering, genetic algorithms and Neural networks.

Clustering analysis make sense of large amounts of data. Large and noisy data is often hard to interpret. It may help to break data set into more homogeneous sets and ease analysis, moreover it may naturally find customer segments but these results need human interpretation. The decomposition of the data into smaller pieces enables seeing the forest without getting lost in the trees. Possible business

targets that clustering can be applied are customer segmentation, product grouping and cross-sales and guide marketing for campaigns. [10][11][20][21]

II.5.1.2.3 Deviation Detection

Discovering abnormal deviations is the aim. The detected deviation may be patterns that do not fit to previously measured classes, significant changes in data in time scale, outliers in the dataset that do not belong to any particular cluster and discrepancies between an observation and a reference.

Deviation Detection is usually performed after a database is clustered to determine whether the deviations represent noisy data. Deviation detection is often the source of true discovery since deviations represent anomaly from some known expectation or norm. [10] [21]

II.5.2 Data Mining Techniques

Building a model to represent the data set is at the heart of data mining process. Implementing the previously mentioned methods to the models, there exists lots of algorithms stemming from statistics, data mining and artificial intelligence. While some techniques belong to completely different approaches, also techniques may vary with analyzer approach differences and combinatory usage in models. Because of the wide spectrum of techniques, only some popular algorithms will be mentioned in this section.

II.5.2.1 Decision Trees

Decision Tree technique develops a classification model from a set of records. Each record in the training set is assigned to one of the many predefined classes which represent the records best as a general concept description. After the model is set, the model can be used for automatic prediction of the class of unclassified records. If each node of decision tree has two branches at most the tree is called as Binary Tree, if node can have more than two branches the tree is called n-way (Multi-way) tree. The representations of this technique are easier to understand, and their implementation is more efficient than those of neural network or genetic algorithms.

This technique can be used in data exploration and modeling phases and especially useful when there are many ways to reach the target. For example a profitable customer can be a high premium paying customer from the commission point of view or low loss ratio customers from the profit sharing point of view.

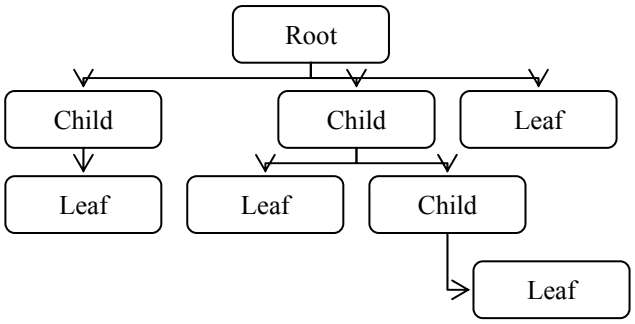

Figure II-4 Sample Decision Tree

Decision tree consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The best split is defined as one that does the best job of separating the data into groups where a single class predominates in each group. Each node has more homogeneous data set and similar size for best split. To test the data sets are homogeneous or not, some tests like Gini, Entropy, Chi-Square are used. These tests are also used for another primary consideration when developing a tree, deciding on how large to grow the tree or what nodes to prune off the tree, in other words limit the tree.

Specific decision tree methods include Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID) algorithm, C5.0 and Quest.

Although decision trees are popular, easy, visualization is powerful, they are limited to one output variable and output must be categorical, and processing numeric values can be complex. Moreover, separation of categories with strict rules like customer having less than 60% loss ratio is profitable but 60.01% loss ratio is unprofitable, may lead biases in decision. [10][11][20][21][23]

II.5.2.2 Neural Networks

Neural Networks is a class of powerful, general-purpose tools applicable in prediction, classification, link analysis and clustering models. Neural networks are

popular since they enable efficient modeling of large and complex problems in which there may be hundreds of input attributes having many interactions like biological neural networks. This widely used technique imitates the way the human brain learns and uses rules coming from data patterns to construct hidden layers of logic for analysis. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. The technique can be divided into two as directed and undirected.

Directed neural net algorithms such as Back Propagation and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular directed neural net algorithm which is used since 1980's. Back propagation can be used to develop not only a classification model, but also a regression model.

Undirected neural net algorithms such as ART do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target data set. Such self-organizing networks divide data set into clusters depending on similarity and each cluster represents an unlabeled category. Kohonen's Feature Map is a well-known method in undirected neural networks.

In the process of a neural network, inputs with weights go through the activation (squashing) function that consists of the combination function that combines the inputs into a value (Generally simply multiplication of weights and input values adding up to generate next node's value) and the transfer function that calculates the output using combination function output. The cycle is repeated iteratively to optimize the target value. Each iteration passing trough all nodes is called an epoch.

The neural networks can be simple or very complex in nature. A simple network may consist of a couple of inputs and one output equals to linear regression statistical technique when combination is weighted sum and transfer function is linear. Another network may have a middle layer between inputs and output which empowers the model, this sub-product layer is generally called hidden layer. The size of this layer increases the efficiency of the network but also raises the risk of over fitting. These networks can produce more than one output. If the hidden layer

has certain non-linear activation functions, more specifically, the combination function is weighted sum and transfer function is logistic, these nets are called logistic regression. The popular type of neural nets, multilayer perceptron is a bit more complicated, that is, the inputs are combined in each hidden node by weighted sum and a transfer function (hyperbolic tangent). The outputs of the hidden nodes are combined again inside the output node.

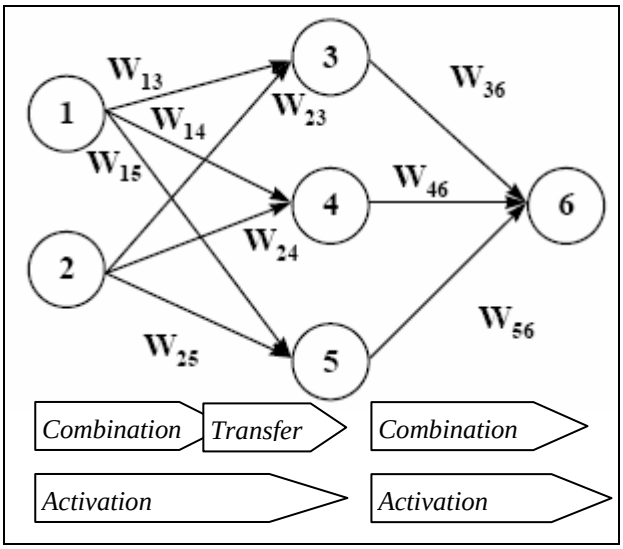

Figure II-5 An example of Neural Network Diagram for Multilayer Perceptron

The weights of node connections are unknown parameters which are estimated by a training method. However, finding the best set of weights for the network in many alternatives under the limitation of time is a difficult problem that affects the success of the model to a great extend. The most common method is back propagation which was one the first ones, however better methods are presented. Some of other methods are conjugate gradient, quickprop, quasi-Newton, Levenberg-Marquardt and genetic algorithms. Each training method has a set of parameters that control various aspects of training such as avoiding local optima or adjusting the speed of conversion. As an example, back propagation begins a first training then calculates the error by taking the difference between the calculated result and the expected actual result. The error feedback is used to make adjustments to minimize the error. Also the name back propagation comes from the iterative method of sending the errors back through the network. However, quickprop uses some partial derivatives to fit a multidimensional parabola and converges minimum error in short time.

Due to the iterative process in neural nets, the solution extremely fits to training set, but the approach using a validation set generalizes the over fitting by the possibility of selecting one of the possible solutions generated in iterations of training set. Some do not use validation set and test training results with test set periodically and observe if the error also decreases in the test set. However, in this approach using the same data of test set during training and test may decrease the effectiveness of the test data to test the final model.

The greatest strength of neural net models is their ability to approximate any continuous function without making any assumptions about the underlying form of the function to be approximated. The linear combinations of sigmoidal surfaces generated increases the capability of estimation but also complexity.

For organizations with a great depth of statistical information, neural networks are ideal because they can identify and analyze changes in patterns, situations, or tactics far more quickly than any human mind. One important characteristic of neural networks is that they are not clear and transparent to view inside, which means there is not much explanation of how the results come about and what rules are used. Therefore, these learning algorithms imitating biological human brain are useful when there exists well understanding of input and output and experience available to train the model. Although the neural net technique has strong representational power, interpreting the information encapsulated in the weighted links can be very difficult. A Neural Network is like a black box that knows how to process inputs to create a useful output but the calculations in the box are quite complex and difficult to understand. The method tends to over fit to the training set due to the wide range of parameters used and iterative approach leading almost every set to a solution, hence the understanding of the data and taking measures are important. The iterative approach also needs extensive time and data volume to reach a solution. Furthermore, neural networks are very sensitive to data cleanness and more successful when less input variable exists although they can process more parameters than statistical models. [10][11][21][23][47]

II.5.2.3 Nearest Neighbor Techniques

General tendency in finding solutions to new problems is often by the review of the similar situations faced before, not to invent electricity once more. Memory

Based Reasoning or in other words, Nearest Neighbor classification techniques operate on the same principle that the results are based on analogous situations in the past. Nearest neighbor algorithm predicts unknown values for records in a data set based on a combination of values for the records most similar to it in an historical dataset. These most similar data are the neighbors giving the name nearest neighbor. For example, if the patient has a circular rash and has recently been bitten by a tick, the patient possibly has Lyme disease because circular rash is the first symptom many patients notice. If the patient later develops fever and joint pain, the diagnosis becomes more certain because these are symptoms that often follow the initial rash in Lyme disease.

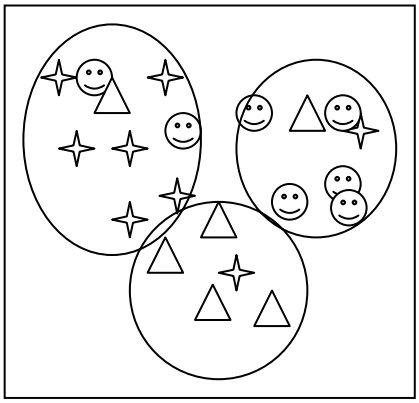

Figure II-6 Groups of Similar Records

The flow of the technique is finding the neighbors of data and combining the records to form a new class. Calculating and deciding on the distance between the records is the first step. The classification goes on using distances to desired level of neighborhood level. Some distant neighborhoods can be weighted lower in the model with respect to the closely related classes. The results of the neighbors are combined to a result set. (e.g. linear regression)

The decisions on the distance are difficult when multi dimensions and non-numeric fields exist and in many sets this is the case. Multi dimensional sets are treated attribute by attribute and the categorical attributes are turned computable numeric values. Memory based reasoning also requires long computations while operating on each record.

Some applications of memory based reasoning are fraud detection, customer response prediction and classification, medical diagnosis and treatments and some interesting applications like face recognition. e.g. the technique can be used to

recommend movies based on the votes and neighborhood of the other clients in a movie store in some literature called collaborative filtering method. [10][11]

II.5.2.4 Regression

Regression algorithms predict the value of a continuous attribute of an instance using the given values for other attributes. In the process, a function is found to return the value of the continuous attribute given the values of other attributes. The function is used for prediction of missing continuous value of an incoming instance. As an example a regression algorithm can be used to predict a safe credit card limit value for a customer. Regression is a common technique easily used in combination with other data mining techniques like decision trees that is for example, each split in a decision tree is chosen to minimize the error of a simple regression model on the data at that node.

II.5.2.4.1 Linear Regression

Regression is based on correlated attributes used to predict from one to the other. The most common and easiest type of regression is linear regression. Linear regression tries to fit a straight line to explain behavior and this line is used to estimate a value for one variable given that the other.

Simply the process of the linear regression is to fit the data to the equation $Y=\beta_0 + \beta_1 X_1$. The fitting aims to minimize the distance between the observed data points and the equation line.

The linear regression can be easily applied and explained. However, linear regressions are not generally sufficient enough for real world cases. In real life, interactions are so complex that multiple variable techniques have to be utilized. Hence, other techniques like multiple regression, logistic regression, decision trees and neural nets are necessary for these cases. [10][11]

II.5.2.4.2 Multivariate Adaptive Regression Splines (MARS)

The Multivariate Adaptive Regression Splines (MARS) is developed by the inventors of classification and regression tree (CART). One of the developments in MARS compared to CART is replacing the hard splits to a continuous transition. This is modeled by a pair of straight lines in each node and leading to a smooth

function (Spline) at the end. Furthermore, dependence to the predecessors in a tree can be eliminated with MARS but the tree structure of CART is not used in MARS to produce rules. Instead, MARS algorithm can determine most important variables for prediction, associations between these variables and dependence to these variables. MARS is an automatic non-linear step-wise regression tool.

Over fitting is also a problem of MARS like neural nets and decision trees. Cross validations with validation set or test set are useful for this problem. [11]

II.5.2.4.3 Logistic Regression

Logistic regression tries to fit a curve to observed data instead of a line like in linear regression. The technique is a special case of generalized linear modeling using odds ratios. The algorithm compares the odds of the event of one category to the odds of the event in another category. Odds ratio is simply the ratio of the probability of being in that class to the probability of not being in that class. For example, if the probability of having exam is 20%, then odds ratio is 20% / (1-20%), 25%. Odds function is a non-symmetric function going infinity. However, using logistic function adds some advantages like a symmetric function having negative values to positive values. As a result, log odds are the basis of logistic regressions.

Logistic regression formulation becomes as $\ln (Y/(1-Y)) = \beta_0 + \beta_1 X_1$. The logistic function itself has a characteristic S shape. The parameters on the model shift the curve left or right and stretch or compress the curve. This function has useful properties like being around 0, having the slope about 45% and moving about the region of -1 to 1. Beyond this range, it gradually flattens out, saturating at 100% or 0%. These properties make the logistic a natural curve for expressing probabilities.

Ease of interpretation is one advantage of modeling with logistic regression. It is useful in binary and discrete variables.

However, in large data sets, high dimensionality makes the detection of nonlinearities and interactions difficult. In addition, it is very likely that some segments of the data space have more records than other segments. When the data is not evenly distributed, the model that fits the whole data space might not be the best

choice depending on the intended application. Although there are many existing methods such as backward elimination and forward selection that can help data analyst to build logistic regression model, judgment should be exercised regardless of the method selected. [10][20][21]

II.5.2.5 Clustering

Clustering techniques are employed to segment a database into groups, each of which shares common properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of data mining operations, generally cluster analysis are followed by other data mining techniques. Clustering is a tool used primarily for undirected data mining with no pre-classified training data set and no distinction between independent and dependent variables, but can be used for directed data mining for forming marketing segments which is a popular application of clustering.

A database can be clustered by traditional methods of Gaussian mixture models, Agglomerative clustering, Divisive clustering, undirected neural nets such as ART and Kohonen's Feature Map, conceptual clustering techniques such as COBWEB and UNIMEM, or Bayesian approach like AutoClass.

Conceptual clustering algorithms consider all the attributes that characterize each record and identify the subset of the attributes that will describe each created cluster to form concepts. The concepts in a conceptual clustering algorithm can be represented as relationships of attributes and their values. Bayesian clustering algorithms automatically discover a clustering that is maximally probable with respect to the data by a Bayesian approach. The various clustering algorithms can be characterized by the type of acceptable attribute values such as continuous, discrete or qualitative; by the presentation methods of each cluster; and by the methods of organizing the set of clusters, either hierarchically or into flat files. K-means algorithm, which is the most popular method as mentioned before, and Gaussian mixture models, Agglomerative clustering, Divisive clustering will be explained in the next topics.

Since the method needs some evaluation like deciding on K in K-Means or ending level of Agglomerative clustering, best tool for this decision is variance because of the similarity is the concern. Best cluster is with the lowest variance. If the cluster size is big, average variance is an evaluation technique. However, since agglomerative and divisive clustering begins or ends with zero variance, the time that elapses between when the cluster is formed and when it is merged into another, larger cluster is a more suitable way to evaluate. Another measure that works for all clustering techniques is to compare the average distance between cluster members and the cluster centroid (seed) with the average distance between cluster centroids (using the distance metric that was used to create the clusters in the first place).

Clustering can be utilized to understand large amounts of data, customer segmentation, reducing records and to break up large data sets into smaller homogeneous pieces. [10][21][23]

II.5.2.5.1 K-means Clustering

K-means clustering is an undirected method that looks for a number of clusters which are defined in terms of proximity of data points to each other. In other words, the method depends on a measure of distance or similarity between points. Different distance metrics used in k-means clustering can result in different clusters.

K-means clustering assumes a geometric interpretation of the data that is, the records are points in an n-dimensional data space. Algorithm begins with determination of number of clusters, which is presented with letter K. Selection of K data points (seeds) randomly is the second step. Remaining data are assigned to the closest cluster, bind to one of the K seeds. The mean of the each cluster is calculated and the seed of each cluster becomes this calculated mean. Since the place of each cluster's seed has changed, reassignment is needed to find new closest records. The cycle goes on until no change occurs by the recalculations and seeds and clusters are fixed. Due to the iterative process beginning seeds are important for the performance time. The method is efficient provided the initial cluster seeds are intelligently placed. The distance is Euclidean, so the K-means algorithm attempts to minimize the sum-of-squares.

The beginning assumption of the number of clusters may change the results if there isn't any reliable reason of this selection. In this case, it is possible to run algorithm for different K values and determining the result by providing lowest variance etc. [10][21]

An example to K-Means clustering is sizing military clothes. [23] The standard sizing of women's clothes where all dimensions increase together and further diversified by body measures causes many different sizes which is difficult to manage and causes high costs. The analyst came with a radical approach and a clustering algorithm, K-Means, is applied to solve the problem. The detailed body measures are analyzed and the clusters by just a few variables were formed. The study results in less variety of sizes and better fitting uniforms.

II.5.2.5.2 Gaussian Mixture

Gaussian mixture is very similar to K-Means clustering adding model a probabilistic approach. Gaussian distribution, a probability distribution often assumed for high-dimensional problems has given method the name Gauss. As before, the algorithm starts by choosing K seeds with a difference however, the seeds are considered to be the means of Gaussian distributions. Remaining work is to optimize the parameters of each Gaussian and the weights used to combine them to maximize the likelihood of the observed points.

The algorithm iterates over two steps called the estimation step and the maximization step. The estimation step calculates the responsibility that each Gaussian has for each data point. Each Gaussian has strong responsibility (weight) for points that are close to it and weak responsibility (weight) for points that are distant. In the maximization step, the mean of each Gaussian is moved towards the centroid (seed) of the entire data set, weighted by the responsibilities. These steps are repeated until the Gaussians are no longer moving.

The reason this is called a "mixture model" is that the overall probability distribution is the sum of a mixture of several distributions. [10]

II.5.2.5.3 Agglomerative Clustering

Agglomerative clustering starts with each record forming a cluster. A similarity matrix is created and the closest clusters are merged to reduce number of

clusters and increase the members of clusters. The similarity matrix is renewed with the new clusters and this process continues until all records are in one big cluster.

The distance calculations among clusters in multidimensional cases three different ways can be utilized. In the single linkage method, the distance between two clusters is given by the distance between the closest members. This method produces clusters with the property that every member of a cluster is more closely related to at least one member of its cluster than to any point outside it. In the complete linkage method, the distance between two clusters is given by the distance between their most distant members. This method produces clusters with the property that all members lay within some known maximum distance of one another. In the third method, the distance between two clusters is measured between the centroids of each. The centroid (seed) of a cluster is its average element.

II.5.2.5.4 Divisive Clustering

Opposite approach to Agglomerative clustering is utilized that all data set is one cluster at the beginning to be splitted to smaller parts. The set is divided into two clusters minimizing the variance and the process goes further by dividing the new clusters to converge zero variance in clusters. As it can be understood from the process, the method is similar to decision trees aiming pure clusters. [10]

II.5.2.5.5 Kohonen maps

Kohonen map, in other words self-organizing map, is a neural network-based approach to clustering. The basic map has an input layer which is connected to the inputs like neural networks and an output layer. As in other neural networks, each unit in the Kohonen map has an independent weight associated with each incoming connection. In contrast to the multilayer perceptrons, the output layer consists of many units instead of just a handful. Each of the units in the output layer is connected to all of the units in the input layer, but not to each other. The output layer is laid out like a grid.

Competitive learning is an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples. A division of labor occurs in the network when different neurons form specialization on different types of inputs. The specialization is enforced by competition among

the neurons, that is when an input is feed to the network, the neuron that is best able to represent the input wins the competition and is allowed to learn it even better.

When a record of the training set is presented to the network, the values of the record flow forward through the network to the units in the output layer. The units in the output layer compete with each other to be output of the network and the one with the highest value wins. The reward is not only to adjust the weights leading up to the winning unit to strengthen its response to the input pattern but also the paths to its neighbors in the grid are strengthened as well. This way, the number of clusters found is not determined by the number of output units because several output units may together represent the cluster. Clusters similar to each other should be placed closer than more dissimilar clusters. [10][25]

II.5.2.6 Association Discovery

Association discovery techniques discover the rules to identify affinities among the collection of items. In other words, association rule extraction algorithms try to find some relationships or hidden patterns in data. Given data, these algorithms try to find rules in if-then form. [27] For example, given an insurer's sales database, such an algorithm can produce a rule like if a customer buys a travel insurance, he also buys a health insurance". As a data mining technique, association rules focus almost exclusively on categorical data rather than on numerical data. The algorithms discover the affinity rules by sorting the data while counting occurrences to calculate confidence. There are a variety of algorithms to identify association rules such as Apriori algorithm and using random sampling. Bayesian Net can also be used to identify distinctions and relationships between variables.

Association rules may be categorized as actionable rules that contain high-quality usable information, trivial rules that tell what everyone already knows and inexplicable rules which can not be explained and not actionable. Unfortunately, the results are generally trivial and inexplicable rules.

Association rules are a good technique for exploring item-based data to determine which things co-occur. The probabilities and joint probabilities in the co-occurrence matrix, are calculated and processed to determine rules.

An association rule uses measures of support, confidence and lift to represent the strength of association in between. The support is the proportion of market baskets where the rule is true. (e.g. If it is thought that who buys rubber also buys pencil, the support is the proportion of baskets containing rubber and pencil together) The confidence is the probability of record occurrences given the triggered part, (e.g. the support ratio divided by the proportion of baskets having pencil). The term Lift (improvement) is how much better a rule is at predicting the result than just assuming the result in the first place. Lift is the ratio of the records that support the entire rule to the number that would be expected, assuming there was no relationship between the products. (e.g. ratio of baskets having rubber and pencil divided by ratio of baskets having rubber and the result divided by probability of pencil in the basket) If lift ratio is greater than 1, it means that the rule is better at predicting the result than just guessing. If lift ratio is less than 1, the rule is doing worse. However, this time it is a sign of negative rule that is if the lift is low rubber may not be associated with pencil.

The sets having many unique items are challenging for the method. For example, a Supermarket having 10 thousand unique items is leading to more than 50 million 2-item combinations and 100 billion 3-item combinations. In order to solve this size problem, usually groups of items are utilized. To conclude, number of groups is important for an efficient process. [10][21][26]

II.5.2.6.1 Market-Basket Analysis

Market-Basket analysis is a popular technique of association rules. In retail market, each customer purchases different set of products in different quantities in different times. The relations among the products going together to the shopping basket are crucial information for a retailer, which is the name origin of the method.

The technique uses the sales information to identify who customers are, to understand why they make certain purchases, to understand products like together purchased products, promotionable products, fast-slow sold products, to take action like promotion decisions, store layouts, package products etc.

Actually market-basket analysis is combinational technique mainly based on association techniques. The data consists of customer data, transactions and product data. Generally there exists a relational database containing these tables. [10]

II.5.2.6.2 Sequence Discovery

Sequence discovery is very similar to association discovery except that the collection of items occurs over a period of time. A sequence is treated as an association in which the items are linked by time. When customer names are available, their purchase patterns over time can be analyzed. For example, it could be found that, if a customer buys a rubber, he will buy pencil within one month 30% of the time. [21]

II.5.2.7 Genetic Algorithms

Genetic algorithms are based on processes in biological evolution like neural networks and nearest neighbor algorithms. The basic idea is that over time, evolution has selected the "fittest species." The name of genetic algorithms comes from the imitation of biological evolution process in which the members of one generation compete to pass on their characteristics to the next generation, until the best generation (model) is found. For a genetic algorithm, one can start with a random group of data and maximize (optimize) the fitness function. For example, in clustering analysis, a fitness function could be a function to determine the level of similarity between data sets within a group and optimization of the goodness of fit.

The process includes selection, leaving the weak out, crossover, split and change parts, and mutation, random change of a part.

Genetic algorithms have often been used with other data mining techniques like neural networks to model data. For example genetic algorithms are utilized in training neural networks to produce weights and in memory based reasoning to find distance values. They have been used to solve complex problems that other technologies have a difficult time with. However, the software support for these algorithms is few. The genetic algorithms are generally used for complex scheduling, assignment, resource optimization and classification. [10][11][21]

II.5.2.8 Distance Evaluation

Although distance measuring is not directly a technique, it occupies a great place in the analysis of techniques like memory based reasoning, clustering etc. In order to calculate the similarities or differences, distances among the records should be well-defined. Even when the variables are more than one, generally this is the case, or variables are categorical, it is difficult to decide on the distance. [10]

II.5.2.8.1 Distance Measuring

If there exists a multi variable problem that variables are numeric, the distance function is calculated separately for each attribute. The most popular distance functions for numeric fields are the absolute value of difference, square of the difference, normalized absolute value of difference (absolute value of difference divided by maximum difference) and absolute value of difference of standardized values (absolute value of difference divided by standard deviation).

A problem that has categorical fields with two possible values is easier to interpret. If the records match distance is 0 whereas don't match value is 1. However, possible values are much more than two in life. In this case, hierarchy leveling which is a tree with many branches, can be utilized. The categories that are closely related placed near each other in the tree and that are not related at all are connected only through the root. The distances can be transferred to numeric values by the number hierarchy levels reaching to common ancestor as an example.

II.5.2.8.2 Distance Combination

The distance between the fields are determined in the above topic, however in the case of a multidimensional problem, these distances should be combined to form record to record distance. Some of the ways of combination are Euclidean approach that is square root of the sum of the squares of each field distance, Manhattan approach that is sum of the field distance, weighted approach that is sum of the field distances times field-specific weights.

When there are many categorical fields, the ratio of matching to non-matching fields is a useful measure. When relationships within a record are more important than differences between records, the similarity is not based on size, the angle

between vectors is a better similarity metric than straight distance. As an example, sine of the angle can be used as distance.

II.5.3 Visualization & Evaluation

A picture or a graph is sometimes worth hundreds of numbers. Data mining techniques are closely related with visual techniques which are increasing the understandability and interpretability of the data. However, visualizations have limitations that they are useful for detecting patterns hidden in small subsets. These techniques are often used in conjunction with data mining techniques: features that are difficult to detect by scanning numbers may become obvious when the summary of data is graphically presented. Visualization techniques can also guide users when they do not know what to look for to discover the feature. Moreover, techniques help end-users understand information extracted by data mining techniques.

Data visualization is important in analysis to see unpredictable situations, increase understanding and discover anomalies. Since data mining generally concerns with discrete data, histograms showing occurring times and time series analysis to see change over time are frequently used. Many visualization tools such as scatter plots, bar graphs and maps are applied to validate assumptions about the data and to get insight of the data.

In order to evaluate the work done, comparisons between predicted results and actual results in pre-classified data that was not used as part of the model building process. The misclassification rate which is an example of assessment methods in classifying algorithms, is a simple measure of how well the classifier is performing. It measures the number of times that the classifier is correct on each of the partitions of the model set. Confusion matrix which shows the error types; falsely classifieds and falsely unclassifieds is also used for assessment since sometimes error of not detecting although should be selected is more fatal error. For example, not to be classified for check-up in life insurance although having a serious death risk (false negative) is a more important error than sending to check-up although there is no risk (false positive). Variance is a common way to evaluate prediction algorithms. Another way is Lift curves that compare the probability of

finding a member of the desired class in a group selected by the model to the prevalence of that class in the population. [10][21]

II.6 Software

There are numerous software in the market that can be utilized in data mining analysis. Some are specialized in one or few techniques, some are gathering some techniques for some functions and some are more combined suits having many techniques for many functions. On regard of the detailed analysis of the software for data mining in Dunham M.H., a list of software including the functions and techniques they cover, is summarized in the table in the Appendix A. [26]

Among so many possibilities the selection of a software becomes difficult even when the cost of these tools are thought to be can not disregarded. This evolving structure of data mining discipline causes neither of the tools to can be the best at all. Moreover, requirements differ from problem to problem. While evaluating software, performance, functionality, usability and process support of the software should be taken into account.

Performance is the ability to handle different data sources efficiently. Although hardware configuration has a major impact on software performance when it is looked on computer-network capability perspective, some data mining algorithms are inherently more efficient than others on analyst-data miner point of view. This criterion generally is an evaluation of computing time concern. Functionality is the variety of methodologies and techniques included in software for data mining. Adaptation of software to different data mining problems coming in time is an important issue as data mining is an iterative process. Moreover, after investment to the software no one knows what the time or in other words, future data brings. Usability is the matter of users. The software should be able to be operated by anyone belonging to different levels ability and without loss of functionality or usefulness. Software not only easy to learn but also guiding the user is preferable. A good tool will provide meaningful diagnostics to help debug problems and improve the output. Process Support comes from the data mining process. Data mining does not only consist of analysis, the process carries

considerable amount of burden like data cleaning and preparation, visualization etc. The software should support the tasks data mining process. [26]

II.7 About Private Health Insurance

Since almost nothing can be done for many patients due to medical evaluation is not adequate, the cost of treatment was low, hospitals were inadequate leading home treatment of patients before 1920's. The low demand for health insurance at the time was matched by the unwillingness of commercial insurance companies to offer private health insurance policies. Commercial insurance companies did not believe that health was an insurable commodity because of the high potential for adverse selection and moral hazard. They felt that they lacked the information to accurately calculate risks and write premiums accordingly. For example, people in poor health may claim to be healthy and then sign up for health insurance. "The opportunities for fraud in health insurance upset all statistical calculations" The Insurance Monitor mentioned in 1919.

However, identification of various infectious agents in early 20th century, production of insulin, sulfa and penicillin in the mid century empowered physicians and hospitals and accordingly rapid change in costs of medication occurred. In the late 1920's the pre-paid hospital services rooted private health insurance. The success of the system attracted the insurance companies. The supply of health insurance increased once commercial insurance companies decided to enter the market for health coverage. Demand for health insurance increased as medical technology further advanced, and as government policies encouraged the popularity of health insurance as a form of employee compensation. Although some state coverage were established like NHS (National Health Service) in UK, the growing expectations and demand keep the private health insurance also growing. The growth of US market which is one the biggest markets can be seen as an example in the below mentioned graph. [29][30]

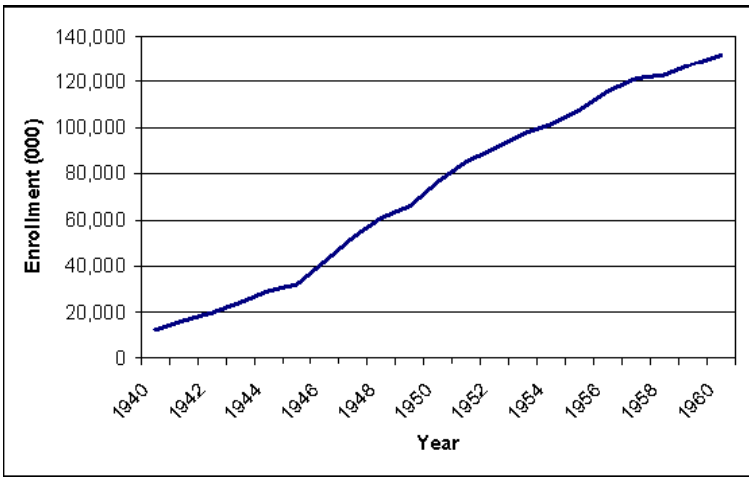

Figure II-7 Growth of Health Insurance in US

When Turkish industry is focused, Health Insurance is defined to be a separate branch by 1990 under Life branch and began to develop by offering limited benefits. Due to the lack of confidence to the social security system, multinational companies require Group Health Covers for their employees and when the more multinational companies started investing in Turkiye, the more share in the health market has been taken by private insurance companies. As a consequence of trend, growing individual demand to health insurance forced private insurance companies to develop policies for individuals in addition to policies providing group health coverage and by the help of sales teams individual health insurance spread out in Turkiye. Although the most policyholders are contributors to the social security system at the same time, they almost stopped using the services provided by the social security institutions.

In first years, health insurance policies began mostly out-patient oriented and in-patient benefits were richened with time. The introduction of private health insurance boomed the private health institutions. With these modern hospitals, a new approach came to scene and Bayındır Hospital which was owned by a group who has also a life insurance company, presented a product which gives extra advantages to treatments in Bayındır Hospital. Contracts with some selected Medical Institutions were done by Halk Yaşam (former Yapı Kredi Emeklilik) and Koç-Allianz Hayat which were the first companies to initiate the system in 1994. The leading companies in the market aim to expand this system gradually to the whole country and thus create a private health service system. These products offer unlimited benefits with some exclusions and cost measures. Even though the policy

is unlimited, when the insured wants to get service from a non-contracted medical institution then, a maximum limit, deductibles or coinsurance are applied.

Check-up benefits and ambulance system benefits are started to be offered in 1994. Since 1995, many different kinds of products have been introduced to the market, some of which offer benefit limited products both per prescription, per visit limit and annual limit per benefit whereas new models with annual maximum limits per person are being developed as package options. In order to increase market penetration, more attractive policy options are being developed in the market for the middle class people. As far as product designs are concerned, the market has been moving from a fee for service model to annual limits and even to unlimited benefit models. Halk Yaşam and Koç-Allianz Hayat have started a family physician service including also free laboratory tests and keeping health condition records of the insureds.

In late 1994 and 1995, guaranteed renewability rights were added to some of the products of the leading companies in the market. In 1997, companies which were not satisfied with the service quality and continuous increase in prices started to quit from the market.

In 2001 Private pension law put the condition for life insurance companies who would like to serve pension contracts, segregation of the health business is obliged. As a result health portfolios in Life companies transferred to insurance companies of the insurance groups. Health insurances are served by non-life insurance organizations since then.

As end of 2005, there are approximately 1.13 million policyholders in the health branch who are either individually insured or members of group plans. According to the trend mentioned in Sigma (Periodical published by SwissRe, one of the leading reinsurers in the world) in emerging countries, non-life insurance including health is expected to grow rapidly. [31] The rate of insureds in Turkiye is still very low compared to developed countries. On the health focus, there are many reasons to expect growth in demand. One of the main reasons is the low quality social security system which forces higher income people to private coverages. Moreover, government afford on the social security system reform aims to transform social security system from publicly owned to privately owned system.

While the wealth of the nation increases, the number of people who can afford private health plans increase and expectations from health coverages increase. Increase in understanding the notion of insurance and importance of health leverages the private health insurance. The laws and regulations increase the quality of insurance system and companies. The reputation of sector increases and financially stronger companies exist in the market. Rather than personally owned and managed companies, institutional organizations own the majority of market share. Moreover, foreign investors are taking place in the market by acquisitions and mergers. Market share of foreign investment reached about 40% in recent years and is expected to reach the half by TSRŞB (Türkiye Sigorta ve Reasürans Şirketleri Birliği) in the near future. [32] The product differentiation and improving services increases the demand.

Almost 80% of the private insureds are from the upper income class and the upper middle class while others are from group health insurance plans which are offered as a fringe benefit by the employers in the private sector. The development of the market with increasing number of insureds will reduce the risk of anti-selection which shall lead to more affordable premiums for middle class people. [33]

The insurance market in Turkiye suffers from tough price competition, exceeding rational, calculated limits. The need to move competition to other fields such as product, service, market enlargement etc. is evident. The differentiation and increase in efficiency are two of key success factors.

II.8 Insurance and Data Mining

Effectively analyzing information from customers, partners, and suppliers has become important to more companies. Many companies abroad have implemented a data warehouse strategy and are now starting to look at what they can do with all that data.

The questions like “How widespread is data mining?” and “Which industries are using?” are common. In this decade it is easy to say, the usage of data mining is

widespread and even extending into some surprising areas like the example below. [19]

STATISTICS INTERN: SEPTEMBER–DECEMBER 2003
Work with Basketball Operations *Responsibilities include:* _ Compiling and converting data into format for use in statistical models _ Developing statistical forecasting models using regression, logistic regression, **data mining**, etc. _ Using statistical packages such as Minitab, SPSS, XLMiner Experience in developing statistical models a differentiator, but not required. Candidates who have completed advanced statistics coursework with a strong knowledge of basketball and the love of the game should forward your résumé and cover letter to: Boston Celtics Director of Human Resources 151 Merrimac Street Boston, MA 02114

Data mining with its many various, simple to complex algorithms and its nature trying to cope with large amounts of data has been usually looked with doubt to cover practical and commercial business needs. However, the data mining techniques are used in many industries from production to service sectors to public sector, from biology to risk management and marketing. [7]

Financial institutions are increasingly mining their data to gain a competitive advantage. Banks, in particular, are using corporate data to tailor their interaction with individual customers. For example, some banks add ATM or POS slip or transcript marketing messages special to the individual customer and transaction, and some banks manage queue waiting time in branches targeting their most profitable customers for better service, making less profitable customers wait longer.

Insurance sector is also one of the largest data gathering points by nature. To evaluate the risk many information should be got and analyzed. From this aspect, insurance is one of the main areas of data analysis and data mining. Widely spread examples of utilization of data mining techniques are encountered during literature survey.

Current emphasis on customer relationship management makes the marketing function an ideal application area to greatly benefit from the use of data mining

tools for decision support. Many insurers are using data mining to help cross-selling other products to existing customers. Another leading aim is to identify possible new customers. Using demographic information and purchasing data from data in the company and/or from a combination of in-company and out-company sources, companies are identifying customer profiles. Attributes of a good customer steers the sales channels like agents, telemarketers or mailers directly to success. In addition to generating sales revenue in the short term, the results of the analysis whether it is successful or not, serve as input for the next data mining project so that the tools make even better predictions. Using data mining techniques, insurance firms can more accurately select which policies and services to offer to which customers also for retaining customers. Database segmentation and more advanced modeling techniques enable analysts to more accurately choose whom to target for retention campaigns. Current policyholders that are likely to switch can be identified through predictive modeling. A logistic regression model is a traditional approach to predict those policyholders who have larger probabilities of switching. Identifying the target group for retention campaigns may be improved by modeling the behavior of policyholders. [23]

Insurance firms can increase profitability by identifying the most profitable customer segments and then prioritize marketing campaigns accordingly. Problems with profitability can occur if firms do not offer the right policy or the right rate to the right customer segment at the right time. For example, for an insurer or reinsurer to use the log normal distribution for rating when the Pareto distribution is the true distribution would likely prove to be an expensive mistake, which illustrates the importance of having the right tool to identify and estimate the underlying loss distribution. With data mining operations such as segmentation or association analysis, insurance firms can now utilize all of their available information to better develop new products and marketing campaigns.

Some insurers are applying data mining techniques to reduce portfolio risk. Using data on policy attributes and historical losses, companies analyze loss patterns to isolate groups of policies that represent a significant loss risk or to define risky customer profile to reduce future risk. The world is too complex to run only on linear techniques. Data mining methodology can often improve existing linear and generalized actuarial models by finding additional important variables, by

identifying interactions, and by detecting nonlinear relationships. In turn, data mining can result in more accurate pricing and hence a better competitive position. An important general difference in the focus between existing actuarial techniques and data mining is that data mining is more oriented towards applications than towards describing the basic nature of the underlying phenomena. For example, uncovering the nature of the underlying individual claim distribution or the specific relation between drivers' age and auto type are not the main goal of data mining. Instead, the focus is on producing a solution that can improve the predictions for future premiums. The data mining techniques are very effective in determining how the premiums related to multidimensional risk factors such as drivers' age and type of automobile. [34]

As another example, insurers are using data mining to identify policies that were issued based on incorrect or fraudulent information. The characteristics of possible fraudulent information reduces the cost of following up suspicious data, a pre-alert system can be launched. The U.S. Department of Agriculture's Risk Management Agency has saved $71-$110 million a year in fraudulent crop insurance claims since 2001 by utilizing data mining, techniques to watch about suspicious 1,500 farms. Data mining, processing millions of claims that large health insurers process annually, proved to be effective in fighting fraud. A study by Conning Research & Consulting in 2000, using the ratio of claims exposure reduction to the expense of investigation units equipped by data mining tools found ratios ranging from a low of 3 to 1 to a high of 27 to 1, depending on the year and line of insurance. This means that every $1 spend returns back $3 to $27. [35]

Data mining can be used to structure reinsurance more effectively than the using traditional methods. In the case of reinsurance, a group of paid claims would be used to model the expected claims experience of another group of policies. With many segments created by data mining, analysts can expect higher levels of confidence in the outcome. The selection of policies for reinsurance can be based upon the model of experienced risk and not just the generalization that it is generally the case in actuarial work.

The settlement of claims is often subject to delay, so an estimate of the claim severity is often used until the actual value of the settled claim is available. The

claims provision estimate is an important value affecting company's profit and loss statement and the reserve amount that should be carried legally subject to the estimation. The estimate can depend on the claim amount, approximate payment date, inflation, currency rates, political situation etc. Data mining operations such as Link Analysis and Deviation Detection can be used to improve the claim estimation. The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated, then over time, the assumption becomes that the future will be much like the distant past. However, as more data become available, the predictive model can be updated, and the assumption becomes that the future will be much like the recent past. Data mining technology enables insurance analysts to compare old and new models and to assess them based on their performance. When the newly updated model outperforms the old model, it is time to switch to the new model. Given the new technologies, analysts can now monitor predictive models and update as needed.

While the mathematical algorithms underlying most data mining tools have been around for more than 20 years, it is only in the last decade that these algorithms have been packaged in user-friendly technology tools. In addition, it is only recently that companies are gaining easy access to the data they need to mine. That's because insurance companies (like most other companies) typically have dozens of different information systems serving different purposes. They generally still have this kind of a system in Turkish insurance industry. While very efficient at processing the transactions needed to run the business (claims, policies, loss reserves, general ledger, etc.), these disparate systems make it difficult for companies to know what information is available or reliable and how to access it.

In insurance world, using data warehousing technology, companies are now consolidating the information from various operational systems and combining it with external data to create a comprehensive, consistent and reliable record of the company's transactions, customers and products. Once the data warehouse is available, companies can then more easily turn to the task of mining their data. However, it should be confessed that for Turkish insurance sector usually this is not the case.

In the future, integration of data mining to the business will change the way of business. Proactive evaluation of customer will alert before losing a profitable customer because of a small claim disagreement. Furthermore, a possible cross-selling chance after a nice gesture about the claim payment will be alerted by the system. [21]

In the Turkish insurance industry, most observers see opportunities and threats occurring as a result of the existing and expected conditions. Since financial legislation is not compatible with banking sector, financial strength is low and some companies are sensitive to market conditions. With the change by increasing foreign investments in finance sector (including insurance industry), further consolidation is expected parallel to increasing competition. Market share of the first 10 company is 93.4% whereas 34 companies exist in the health insurance market. [36] The industry expecting legislations for a long time and sectoral legislations are in the agenda of this government. These developments may accelerate consolidation in industry. With interruption of global players in the industry, classical distribution methods over agencies are also subject to change. Although some insurance companies are older than a half century, conservative characteristics of the sector lead the companies left far away behind today. With the changes and introduction of new BES system, classical health and property insurance is on a junction to go on or leave. This condition is also an opportunity for remaining companies, with a higher share, better productivity, higher profits and in a fast growing country increasing new customers.

In spite of centuries of sales experience, most insurers rely heavily on price for acquiring and retaining customers. The reality is that customers base insurance purchasing decisions on many factors. While this does not mean that we can ignore pricing, it means that alternatives exist for insurers to remain competitive. By relying on the power of data mining, insurance companies can maintain the consistency and accuracy of their underwriting decisions; they can significantly reduce the impact of fraudulent claims; and they can have a better understanding of their customers' wants and needs.

Within the field of data mining, supervised and unsupervised learning is possible. While with the database of policyholders' loss histories whether

policyholders have had a loss during a specific time period or the magnitude of a loss can be analyzed; with sales data exploratory analysis can be done to identify policy endorsements most often purchased together so new product bundles can be created to better suit the customer needs. The technology exists for building reliable predictive models rooted in statistics that consider many more factors than an underwriter can. Insurers can also use data mining to improve detecting and predicting claims fraud. Most insurers also use a rules-based approach to alert claims. As is the case for underwriting, there is generally no statistical basis for the rules. Fraud rules are based on limited past experience and are ineffective in detecting emerging fraud techniques. Data mining improves fraud detection. For example, to uncover anomalous claims behavior that may indicate fraud, unsupervised learning is an excellent approach. [37]

Underwriters, marketers and actuaries can build predictive models that identify policyholders most likely to depart, new customers most likely to respond to product offers, and existing customers most inclined to purchase additional products. Data mining enables insurers to help their agents and sales channels, both captive and independent, simply by giving them pre-qualified leads. By making the right offer to the right customer at the right time, the company adds value to the relationship.

Responding to competition with lower prices alone has never been a long-term solution. With data mining, insurers have many ways to improve their competitive position: better underwriting, improved fraud detection and more effective business strategies.

To survive in the new world of financial services, insurance firms must be able to match their product and service offerings with the sophistication and variations in their marketplace. That kind of business intelligence requires a data mining system that can shift enterprise to be able to analyze the profiles and preferences of existing customers, to predict customer buying habits and to focus sales and marketing campaigns on prospects who have a high probability of becoming customers.

Data mining methodology often can improve upon traditional statistical approaches to solving business solutions. Data mining often can improve existing

models by finding additional, important variables, by identifying interaction terms and by detecting non-linear relationships. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs. Some possible data mining application areas in insurance firms can be summarized such as:

Establishing rates, pricing: Modern data mining models can more accurately predict risk, therefore insurance companies can set rates more accurately

Acquiring new customer: Sales efforts that are guided by more quantitative data mining approaches can lead to more focused and more successful results, marketing strategy to the target.

Retaining customers: Analysis at customer segment level to promote bundled insurance policies or retention campaigns to the most likely to leave or pass to competitor.

Developing new products: Offering policies so that insurance agents can sell them, identifying profitable customers, segmentation of products and being aware of the market changes.

Creating geographic exposure reports: Risk changes by geographical or socio-economic conditions. Combining demographic data calculating risk is better.

Detecting fraudulent claims: For example, the United States spends more than $1 trillion each year on health care. Of that amount, over $100 billion is estimated to be lost annually to fraud. Add in $20 billion in property casualty fraud, and an "insurance fraud industry" would be in the top 25 of the Fortune 500. [38]

Reinsurance: Not only reinsurance reduces the risk of large claim amounts but also reduces the revenue because of the premium transferred to the reinsurer. Optimizing the reinsurance is the target.

Performing campaign management: Increase in return of investment in campaigns. (See Figure II-8 Lift Curve of Data Mining Project for Mailing Campaign)

Estimating outstanding claim provisions: The prediction of the item that directly affects profit and loss.

Coordinating actuarial and marketing departments: Traditionally independent working styles of departments, turns to coordination by interdependent analysis of data.

Data mining has a wide application area from marketing side to claims side in insurance sector. For database marketing, segmentation, promoting, campaigns data mining can be used to increase efficiency of sales. On the other side in claims process, fraud detection is a critical issue in insurance sector.

Due to poor understanding of clients' behavior, misleading sales strategies cause inefficient occupation of sales force and increase companies' costs. In this study, it's aimed to collect, consolidate, structure irregular company data and analyze the data using data mining tools to generate meaningful customer behavior information to support the sales strategy decision in an insurance company. Mainly two questions need to be answered by the company, how to select people and which product to promote to uncover patterns hidden in the data that can be used to predict the behavior of customers and products.

Using data mining tools; previously unknown, valid patterns and relationships in large customer data sets will be discovered. The results will:

- minimize acquisition costs by understanding which marketing programs will generate the highest responses before investing limited marketing budgets in the wrong campaigns or inaccurate targets
- enable company to pursue aggressive customer growth strategies through cross-selling

As a result;

- increase in sales
- enhanced sales force productivity, reduced costs
- increase in customer satisfaction are expected.

II.9 Data Mining Techniques in Insurance

Since data mining is not an industry limited method, all of the techniques are also subject to utilization in insurance industry for different problems. Hence, some

examples of techniques used and successful applications are tried to be mentioned in this section.

As a leading example, CoIL challenge carried out in the year 2000 can be given. The CoIL Challenge 2000 offered the chance to work on an insurance real-world data set. The tasks were to predict which customers are potentially interested in a caravan insurance policy and to describe the actual or potential customers; and possibly explain why these customers buy a caravan policy. In this challenge many methods are used by competitors including, Naïve Bayesian, back propagation neural networks, self-organizing maps (SOMs), evolutionary algorithms, Kohonen nets clustering algorithms, C4.5 (former version of C5), CART, and other decision tree induction algorithms, fuzzy clustering and rule discovery, support vector machines (SVMs), logistic regression, boosting and bagging, and more. As it can be seen in this example, the method and techniques do not only change by data structure and aim of the analysis but also the point of view of the analyzer. Although the aim and data set is exactly the same, competitors used various techniques and combinations of them to reach the best solution. [39]

The analysis of claim patterns in health insurance to detect fraud and inappropriate, excessive diagnostic or testing methods used is carried by HIC, The Australian Health Insurance Commission. In addition, HIC has developed and implemented a pattern-recognition technique for classifying the degree of appropriate practice of service providers. By the method, a score or class is assigned to each practitioner; those practitioners who have a high likelihood of an inappropriate practice pattern are identified. In the study, claims database and a general practitioners database were used to apply association rules and neural nets. On a large, about 550 GB, data, the analysis performed in a reasonable time. An inappropriate payment of over $550,000 per year which couldn't had been detected for over 5 years by conventional techniques, was uncovered by the detection of analysis. The classification of practitioners enabled better monitoring. [40]

In the Proceedings of the 23rd VLDB Conference, Swiss Life's data mining efforts to utilize the data containing 650,000 contracts was one of the subjects. In 1990's, the company established the system by the first aim of marketing that is identifying potential customers, determination of leaving customers and customer

segmentation etc with association rules, decision trees and classification techniques. To support marketing initiative to preserve and extend the market share of the insurance company data mining served a good role. [41]

Behavioral segmentation of retail customers for characterizing the purchasing profiles are carried out in Toro Insurance Group. Factor analysis and clustering algorithms are used for more efficient marketing, product development and better customer knowledge.

In Farmers Insurance Group, the data mining is used to increase profitability and for better risk analysis. A 4 year data containing 35 million records is analyzed using decision trees by the tool Underwriting Profitability Analysis (UPA) which is a software developed for the insurance market by IBM. Forty rules are generated to lower cost for claims of several million dollars.[42]

In an example, data mining is used to increase success of mailing campaign of a French insurance company and the results are impressive, represented below. The aim is to select potential customers and decide which product to promote. Clustering and neural network techniques are applied in parallel and intersection set of results are processed for segmentation. The decision trees are used to identify the variables that affect the ownership of each policy. Association rules are applied for cross-selling opportunities. Efficiency is leveraged greatly; getting 4 units success by spending 1 unit is possible instead of spending 10 to get 10 units success. Cost and benefit analysis of the campaign can be affected dramatically by application of techniques. The lift curve of the project can be seen below. [43]

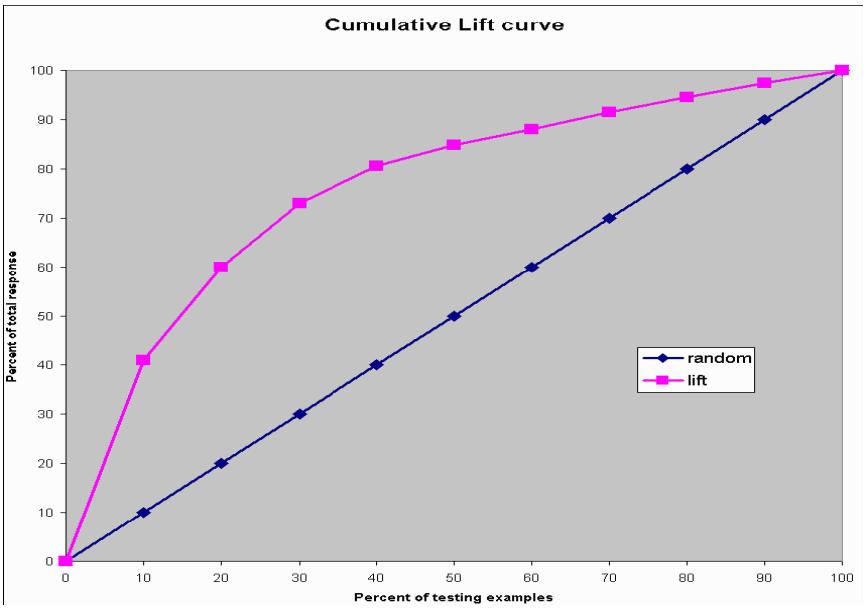

Figure II-8 Lift Curve of Data Mining Project for Mailing Campaign

A similar work is done by Two Crows in a bank to improve efficiency of direct mailing campaign for credit sales by the use of decision trees to predict high potential responders and neural nets for credit scoring model. The net profit after the cost of modeling project is deducted increased 35% with respect to the traditional campaigns. Return on investment is (ROI) 213%. [23]

The paper by Guo [21] also introduces the application of clustering and decision tree data mining methods to insurance sector.

II.10 General Assumptions and Difficulties

In this section, some general challenges of a data mining process are also mentioned accompanying difficulties faced on the analysis stage and forthcoming assumptions during the analysis stage.

When Statistical and/or Data mining techniques are used, it means that some assumptions are made by default. In this study, past data is used in order to make better decisions about future actions. Hence, the most important assumption is that the future will be enough like the past that results from past data will remain applicable in the future.

Data Mining is an iterative learning process and requires long-term hard work and commitment but the reward is company transformation from being reactive to

being proactive. In this study with a limited time to pass all the data mining process without a project team automatically narrowed the coverable analysis. Actual data of a time period is studied in this thesis but data mining is a continuous process due to continuing creation of new records. Hence, data mining approach is dynamic and cyclic. Only one cycle in this process is presented. New business problems or changing behaviors in time will emerge the need of data mining analysis.

One of the challenges in data mining is to figure out which patterns are predictive and which are not. Correlation does not always imply causation. Some patterns reflect some underlying reality (i.e. The party that holds the Government tends to lose seats in Parliament at the next elections.) but some patterns don't (i.e. If a fair coin comes up heads 5 times in a row, there is still a 50% chance that it will come up tails on the sixth toss.) A complete analysis is needed not to learn things that are not true. If sales appear to be down in November compared to October and previous months, it may not mean a decline in sales performance. Because of the Ramadan Bayram working days are less in November and when sales per working days are compared, it may not indicate a drop. All the mining outcomes should be evaluated by the analyzer to be meaningful or not.

The model set is on the collection of historical data. However, the data set must reflect the population that the model will describe, in order results to be valid. A sample that does not properly reflect its parent population is biased and using a biased sample as a model set causes learning things that are not true. (i.e. customers who started 3 years ago may differ from customers started today) It is assumed that the data set reflects the behavior of customers today and in the future without a bias.

Moreover, general socio-economic conditions are assumed not to be changed since the data set period until now. In other words, the socio-economic effects on data are neglected in the study.

Coping with different types of data is another challenge for data mining. There are many kinds of data and databases used in different applications. Since most available databases are relational, it is crucial that a data mining system performs efficient and effective process on relational data. Moreover, many applicable databases contain complex data types, such as structured data and complex data objects, hypertext and multimedia data, spatial and temporal data,

transaction data, legacy data, etc. A powerful system should be able to perform effective data mining on such complex types of data as well. However, the diversity of data types and different goals of data mining make it unrealistic to expect one data mining system to handle all kinds of data. Specific data mining systems should be constructed for knowledge mining on specific kinds of data, such as systems dedicated to knowledge mining in relational databases, transaction databases, multimedia databases, etc.

Since data sets are large effectively extracting information is an important concept. The data mining algorithms must be efficient and scalable to large databases. In other words, the processing time of an algorithm must be predictable and acceptable in large databases. Algorithms with exponential or even medium-order polynomial complexity may be out of practical usage.

The data mining results should accurately present the contents of the database and be useful for certain applications. The accuracy should be expressed by measures of uncertainty, in the form of approximate rules or quantitative rules. Noise and exceptional data should be handled elegantly in data mining systems. This also motivates a systematic study of measuring the quality of the discovered knowledge, including reliability, by construction of statistical, analytical and simulative models and tools.

Different kinds of knowledge can be discovered from a large amount of data. Moreover, it should be possible to examine the results in various forms. This requires high level graphical user interfaces and representational techniques so that the data mining task can be understood by non experts and the discovered knowledge can be understandable and directly usable by users.

With time, the availability of the data from different sources increases. The networks and internet are opening a wide range of source bur these sources are distributed, heterogeneous or unformatted. Although data mining finds ways to cope with these irregularities and parallel and distributed data mining algorithms are developed to find a solution to distributed data, this complexity brings other difficulties and generates new problems to be solved. [44]

III ANALYSIS

III.1 Business Understanding

Marketing strategy evaluates in time and demand-supply relation changes. The market begins with product focused strategies then reached sales focused approaches. In time, the market begins to drive the strategies and finally its time of customers, the requests of customers waiting more than a good product now leads market. The hard competition wears the insurance companies.

In this study, a leading insurance company is under focus. The past time real records are utilized to improve business strategies.

III.1.1 Business Objective

The company has an irregular customer database having generic problems. No detailed customer analysis have been done accept some non-systematic querying efforts, moreover no data warehouse has been constructed before which hardens to work on the data. Poor understanding of client behavior misleads sales strategies and inefficient occupation of sales channels increases company's costs, both real and opportunity costs. The strategies and campaigns are generally built on demand of sales channels. Untargeted sales efforts leads inefficiency in sales, high cost, much time, less success. Moreover, the profitability of the sales is not concerned in these efforts although aim of business is basically the profit for shareholders. This may cause high sales but high loss situations.

In this study, it's aimed to collect, consolidate, structure irregular company data and analyze the data using data mining tools to generate meaningful customer/product behavior information to support the marketing strategy decision in the insurance company to increase efficiency and profitability. Main business

objective is to find answers to these two main questions; how to select people and which product to promote to build CRM strategies on.

A general understanding of customer life cycle is needed to decipher the problems. A customer life cycle begins with identification of potential customer, proposal and response. The identification of potential customer is important to select right/high potential customer for effective selling activity. Furthermore, who are the possible high profit potential customers, who are risky, who tends to be loyal customer are the questions to support effective selling. The knowledge support of descriptions for sales comes from the customer stage database. Moreover, relations of products are used for cross-selling and anomalies in transactions are used for fraud detection in customer stage. Leaving customers are also identifiers for new sales and moreover for retention campaigns. A short summary and examples are given in the below table. [42]

Table III-1 Customer Life Cycle

	Acquisition	**Activation**	**Add value**	**Retention**
State	Potential	Responder	Customer	Ex-Customer
Customer Potential	Target	New customer	Customer	High-Potential *Voluntary leaving *Forced leaving
Action	*Acquisition Campaign	*Response Acquisition Campaign *Information gathering	*Cross-selling *Up-selling	*Retention campaign *Force to leave
Data Mining Models	*Predictive model for selling *Predictive model for Risk Analysis *Descriptive model on relevant variables		*Descriptive model for Customer Behavior *Predictive model for cross-selling *Predictive model for fraud detection	*Predictive model for churn

The marketing strategies penetrate all over the customer life cycle. In order to carry out more focused study in the limited time, the objective of the study is limited to the first stage of the customer life cycle that is customer acquisition. The results are expected to lead sales efforts to increase efficiency and more profitable sales results.

III.1.2 Selected Tools

Since data is on separated AS/400 databases, queries of the system and WinSQL are mainly used to retrieve data from the main systems. These tools are selected because of easy and convenient data selection and settling relations in between in order to take the data needed.

While exploring the data, MS Office tools like Access and Excel are used widely. These easy to find tools helped to work on large amounts of data, direct ODBC links to the main systems and set detailed relations. Moreover, computed fields or groupings are convenient on these tools. Although the data mining tool selected supports these phases, the locational conditions forced to prepare data on MS Office tools.

On the analysis aspect, Clementine; data mining tool of SPSS, is utilized. This tool supports all steps of the data mining process on CRISP-DM standard. In Clementine, the process is represented as data flows, called streams, which can be created by the user through a visual programming interface. The user can select icons which represent data sources, data manipulation algorithms, data mining algorithms, graphical and statistical data analysis techniques. These icons are combined into streams. Clementine is a well-established and functional tool. Moreover, beyond wide range capabilities of the tool, SPSS and Clementine are one of the most preferred tools for data mining. In a survey conducted by Kdnuggets in 2006, 29% of the responders voted for SPSS and Clementine as the most preferred tool. [45] It was again at the top in 2005 with 27%. [46]

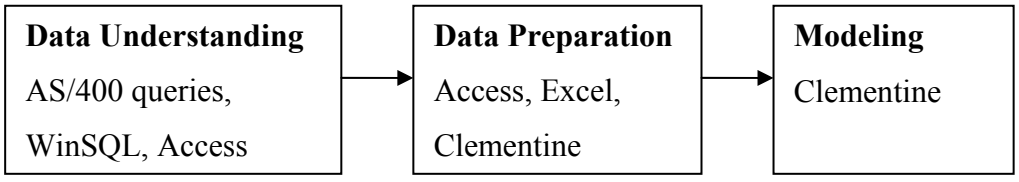

Figure III-1 Phases and Tools Utilized

III.1.3 Problem Definition

According to the process approach of data mining, business problem should be translated to a data mining problem to be solved. Since the business objective

aims better marketing strategies on acquisition of new customers, wide understanding of present customers is required to address good potential customers. Customer clustering and classification are two of the most important data mining methodologies used in marketing and customer relationship management. The methodologies generally use customer purchase transaction data and demographic data to track buying behavior and create strategic business initiatives. The analysis is made to divide customers into segments based on shareholder value like customer profitability, risk, a measure of the lifetime value of a customer or retention probability. Creating customer segments based on such variables highlights obvious marketing opportunities. In the light of shareholder value, good customer in our study covers notions of having more potential to buy, to be loyal, to pay high premium and to be profitable. High-profit, high-value, and low-risk customers are the ones a company wants to keep and get. Demographic data in segmentation steer marketers in selecting appropriate advertising, marketing channels and campaigns to satisfy the strategic behavioral segment initiatives.

On the other hand, on the product side, promoting the right product(s) to the right customer requires understanding of relations among products and customers. Moreover, the relations among products for cross-selling are also added to the problems to be solved, as chances at the time of first purchase for proactive strategies although is generally located to customer phase. Another target is to eliminate customers having possible high risk of generating loss in order to increase profitability of sales activities.

In detail, below mentioned questions are data mining questions to reach business objective.

Main questions:

- Who are loyal customers? For defining good customers.
- Who are profitable customers? (Claims/Premium) For defining good customers.
- Who are likely to buy? Potential buyers.
- Who are high premium paying customers? For defining good customers.
- Who are most likely to leave? For retaining good customers.

- Who buys which product? For acquiring new customers.
- Are there interdependencies among products? For increasing value of customers, cross-selling.
- Who are likely to be risky customers? For defining bad customers.

In coordination with the methods mentioned in previous chapter, the questions can be converted to data mining problems as below:

- Customer Segmentation
- Loyal customer description
- Profitable customer description
- Prediction of loss generating customer
- Association between health and P/C products

As a first step to data mining, profitable customer description is selected as the aim of the study.

III.1.4 Beginning Strategy

According to our problem definition generally the below mentioned data is needed to analyze.

- Personal Information about customers (Sex, age, occupation, territory etc.)
- Purchased policy information (Product, premium, claims, insured etc.)
- Other insurance branch info (Cross-sales, product etc.)

As a beginning strategy as many attributes as possible would be reached and a defining set would be retrieved to form model set. Health insurance data in the period of 2001-2004 is used as main data set and Property/Casualty insurance data is also appended to the main data. The decision about the period is taken inevitably due to organizational change in 2000 affecting data and 2005 data couldn't be taken due to the characteristics of insurance industry that claims are still accepted in the next year period.

III.2 Data Understanding

This process is the most time consuming step with data preparation step in a data mining study in literature. In addition to generic problems of data leading these phases to be the most time consuming, distributed and unstructured data sources doubled the effort of understanding the data in the study.

III.2.1 Selection of Appropriate Data

The data in the company database was passed through a detailed investigation to bring out what is available. After many meetings with business and IT departments, it is understood that data is needed to be deciphered from all over the tables used for transactions. The data containing customer personal information is distributed in three different files and nothing about purchase information could be reached in these databases. The attributes in files are generally operational information including Bank information, Addresses, Phone numbers etc. The parameters defining the customer such as age, sex etc. are selected from these files.

Since the company hasn't got a structured data warehouse, the transactions should be summarized to retrieve sales and claims information using SQL scripts and querying and further processing is done on MS Access. Every claim record for the policies beginning between 2001 and 2004, more than 2 million records, are processed to reach claims data, and policy information files are used to get premium records. Additional data which are collected for medical insurance purposes, like height, weight, risk etc., are gathered from different locations on databases. The data including the information whether the client is subject to a discount or extra additional premium is also joined the data from other sources.

Property and Casualty insurance information of clients are also subject to work on different systems. Data on both systems gathered and affiliated to form the data set. Since it is hard to find data as many attributes as possible are tried to be joined to data set. Moreover, all of the records are collected because this is the first data mining analysis carried out in the company and information about the entire portfolio is needed.

During the data gathering phase more than 20 different transactional tables in company databases which are sometimes in different environments are connected, filtered and summarized to build a data set.

In health insurance, policies are generally made for one year period. Since analysis is carried on a time horizon, the data is prepared on policy basis.

III.2.2 Exploring the Data

A database is only as useful as the quality of data it contains. Whenever half of the fields have greater than 60% missing values or there are discrepancies and wrong entries in the data, it is difficult to extract meaningful and helpful value. This may be even misleading. As a result, data quality is the most important issue in data mining.

The business is a changing environment. Since the data contains a 4 year period, there are inconsistencies in the information stored in data. During this period company changed regional organization, some products are put on the market and some do not exist anymore etc. Hence exploring the data inconsistencies is crucial and they are to be solved and deeper understanding of data should be reached. Actions taken are summarized below: (Since the exploration of data is closely related to the consequent phases like cleaning and transformation, the below mentioned explanations also cover some actions belonging to these forthcoming phases not to lose relation.)

III.2.2.1 Policy Number

Policy number is a key for the policy. Detailed data (such as payments, claims) gathered under policy number which is unique and records are formed on policy basis.

III.2.2.2 Beginning Year

This field indicates the year that policy coverage begins. The beginning year is used as the year of policy in analysis. The data covers the policies beginning in 2001-2004 period.

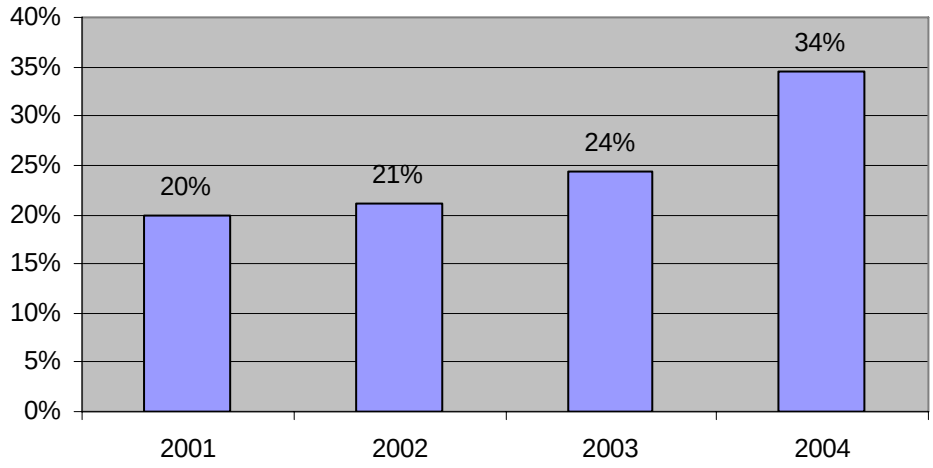

Figure III-2 Distribution of Policies by Beginning Year

III.2.2.3 Product

Usually the products drawn from market are predecessors of new products. Variation by details can be ignored among similar product groups. With this information, products are grouped and predecessors are joined to related groups. Moreover some products differentiated to give customer more flexibility, these differentiated products by minor changes are also grouped in order to have enough data in each product group. As a result, a major grouping is formed for first level analysis and a minor level grouping is kept for the instance of more detailed analysis. The Major product groups are explained below:

- CORTA: Economic class low coverage products
- CALT: Medium class standard coverage products
- CUST: High class extended coverage products
- S: Short term obligatory travel products
- W: Upper high class full coverage products

When the data is analyzed, it is seen that more than half of the sales belongs to high class products, this is also a clue that high income class people are still the main client segment.

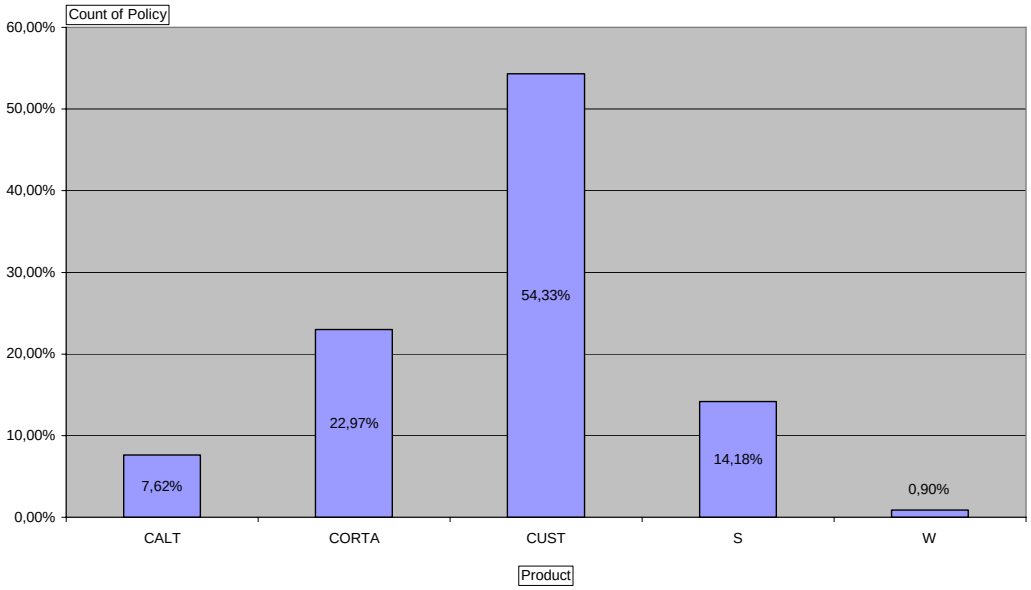

Figure III-3 Number of Policies Sold by Product

When the consistency of data is analyzed yearly, there is a jump from low class to higher class products between 2001 and 2002 which takes attention. The reasons of this situation are analyzed and it is found out that a product change took place in these years. In 2002, coverage of products was changed and high class products are differentiated to be more flexible. Most of the middle class product insureds preferred to go on with high class products.

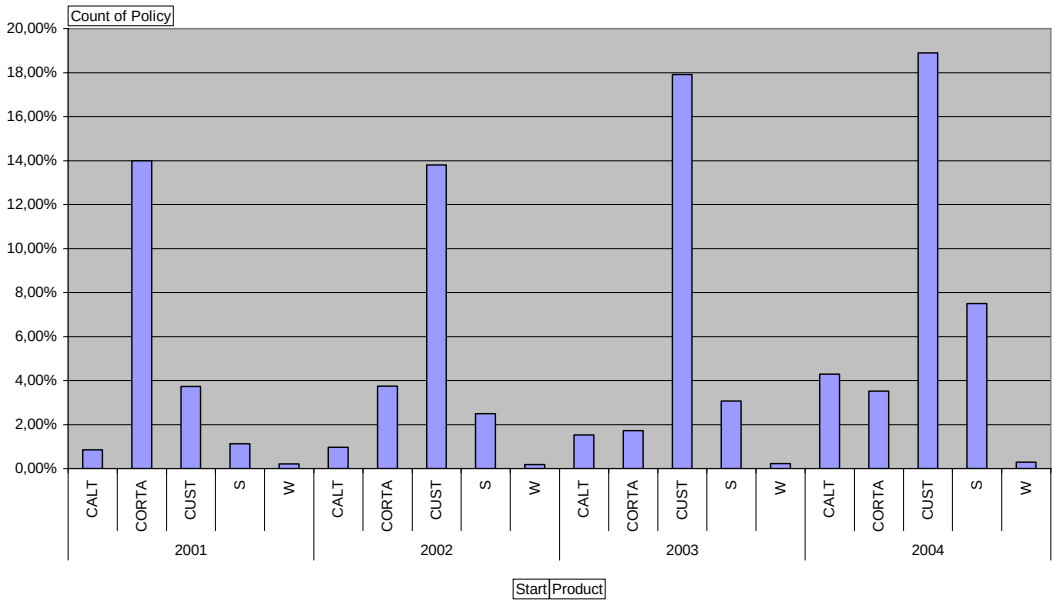

Figure III-4 Annual Number of Policies Sold by Products

III.2.2.4 Renewal Years

Health policies are yearly contracts except short-term travel insurance and policies are renewed every year. Hence, renewal year is a direct indicator of loyalty. The company is in the market since health branch is settled in 1990. As a consequence, up-to 15-16 years of renewal is possible in the data.

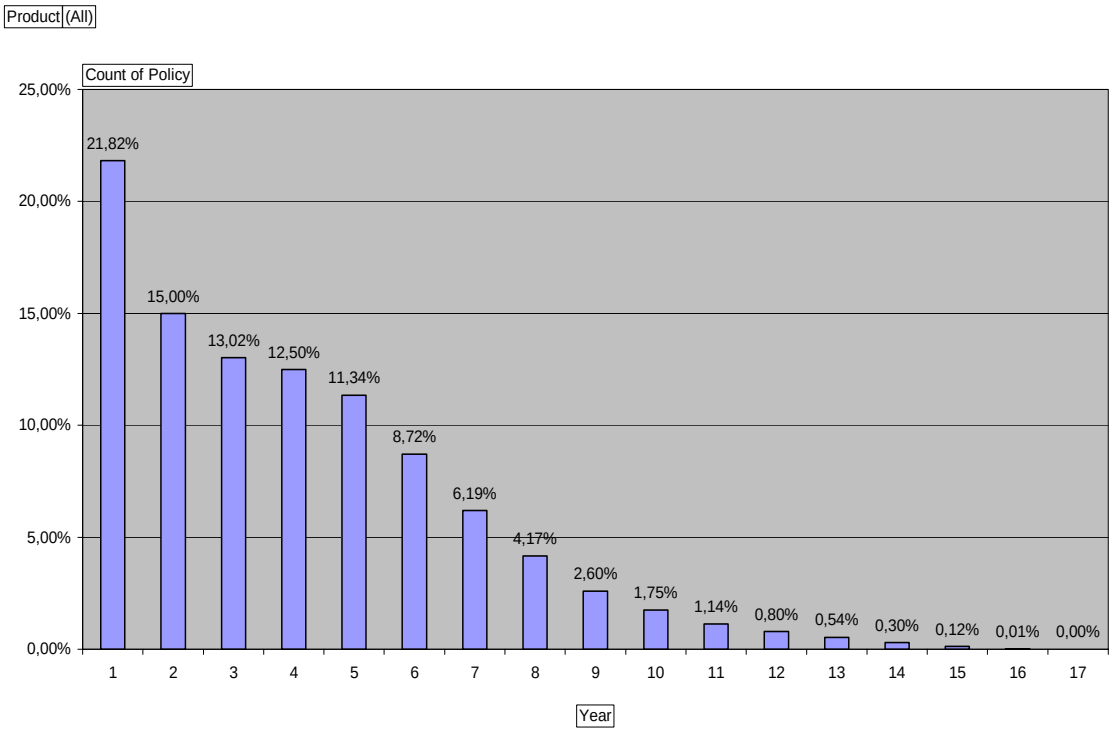

Figure III-5 Number of Policies Sold by Renewal Year

In the above mentioned graph it is seen that, although they are almost negligible in number, there are customers who are with the company since the establishment date. Also the distribution of the histogram is as expected.

III.2.2.5 Customer ID

Customer Number is the identifier of customer. In 4 year time period, the same customer may have 1 to 4 or more policies (Purchase of more than one product). In order to be able to trace the customer which is the main concern, Customer ID is kept in data set.

III.2.2.6 Sales Region

The data covers a four year period. In a living organization, this is a long time and many changes can be occurred during the period. When the changes are

investigated, it is founded that regional coding, product coding and some organizational changes were taken place. In order to standardize the data, former regional records are changed with the present organizational regions. Organizational changes are synchronized with the present organization.

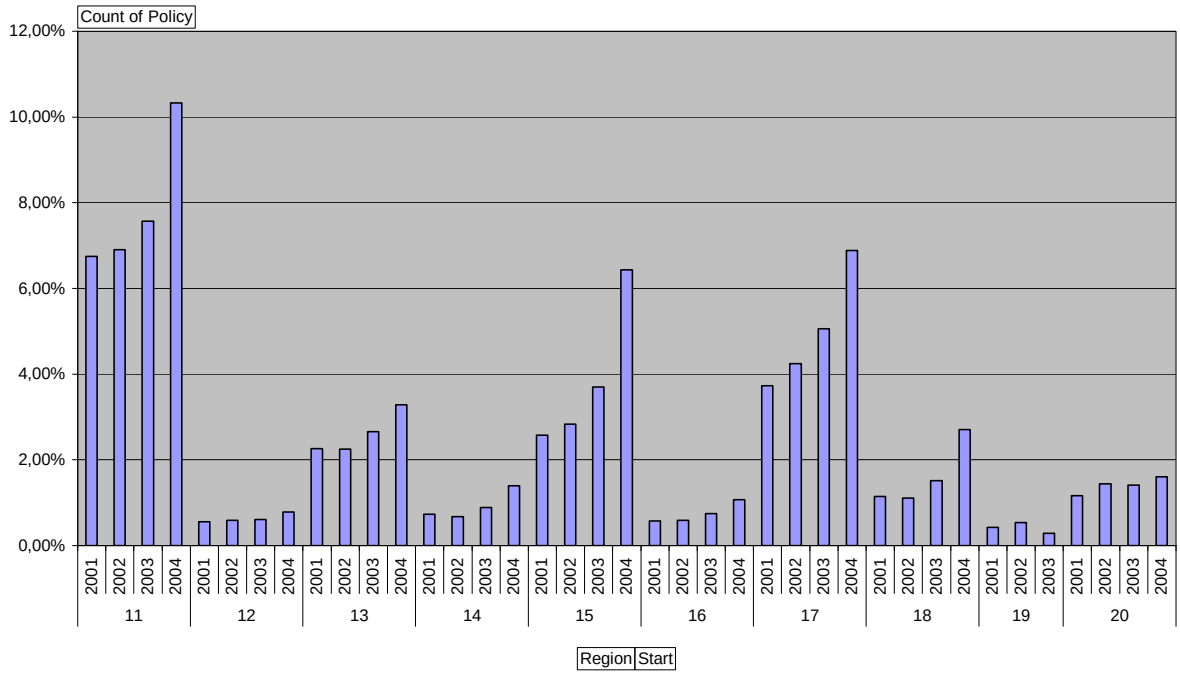

Figure III-6 Annual Distribution of Policies Sold by Regions

Hence, when the data is analyzed by regional time series, it is found that a region's production is zero in 2004. It's found out that this region stopped sales in 2003. Therefore, previous records should be synchronized with the new ones. In order to find the new region of the customers, the renewals of the policies are traced, and former policy regions are replaced by new policy's region. However, if the policy is not renewed, the new region can not be found. Total number of No.19 region clients is less than 1% of company profile and 73% of the records are corrected by tracing renewals. When renewal regions are analyzed 70% of renewals are from Region No.1. Hence, remaining clients consisting 2.7‰ of portfolio is assumed to be from Region No.1. Regional distribution of data appeared to be as expected.

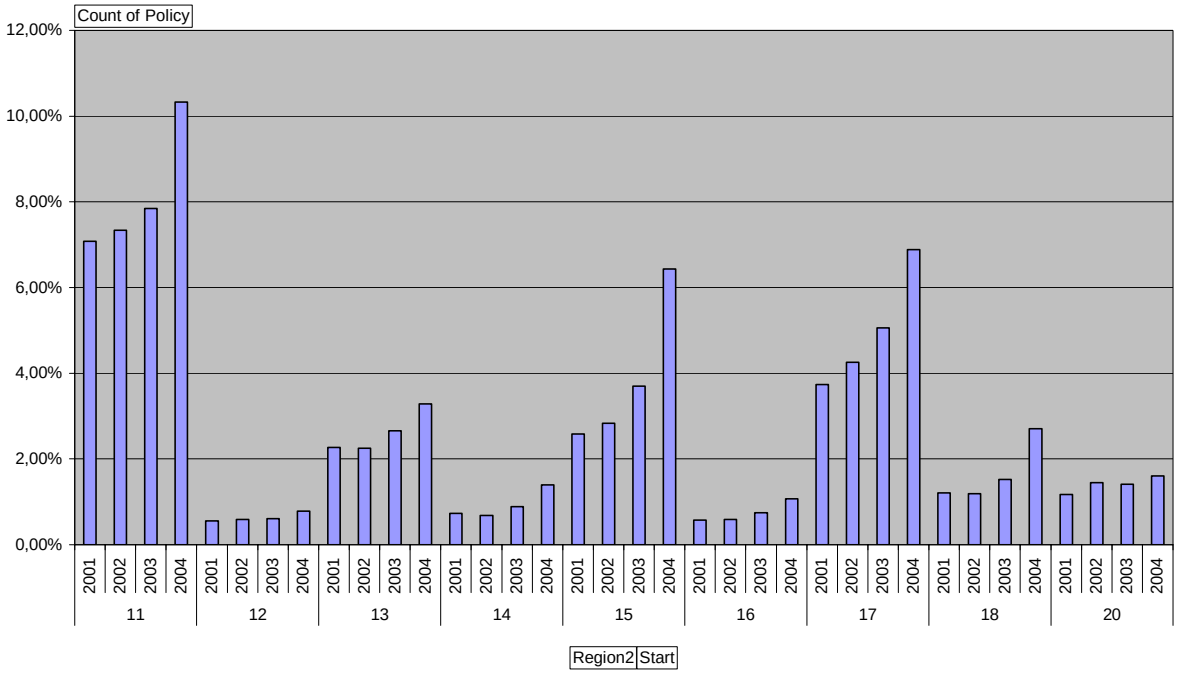

Figure III-7 Annual Distribution of Policies Sold by Regions (Changed)

III.2.2.7 Sales Channel

Since in this study finding out general understanding of behavior is aimed, detailed analysis of each channel is not covered. Main sales channels are the real concern of the study, hence channels are grouped as follows:

- A: Agencies
- B: Banks/Financial Institutions
- D: Direct Sales Teams

In general insurance industry is mainly dependent to agency sales channel. However, as it can be seen from the below chart, in health insurance agency sales channel is not as dominant as other branches.

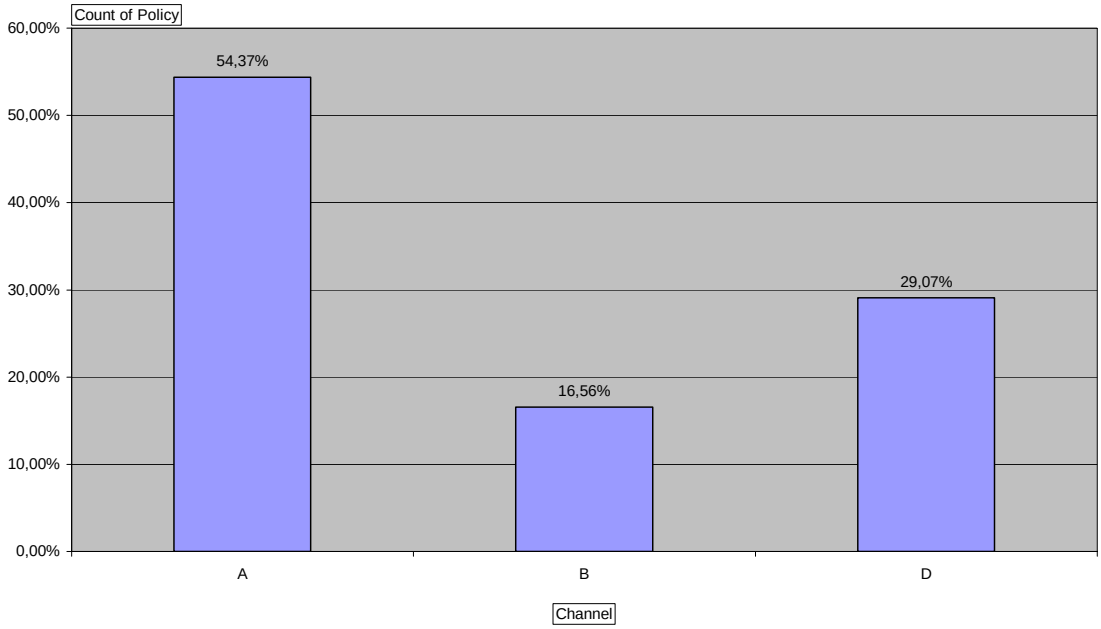

Figure III-8 Distribution of Policies by Sales Channel

III.2.2.8 Age

Age is one of the crucial information in medical insurance. The ages of insureds are graphed below:

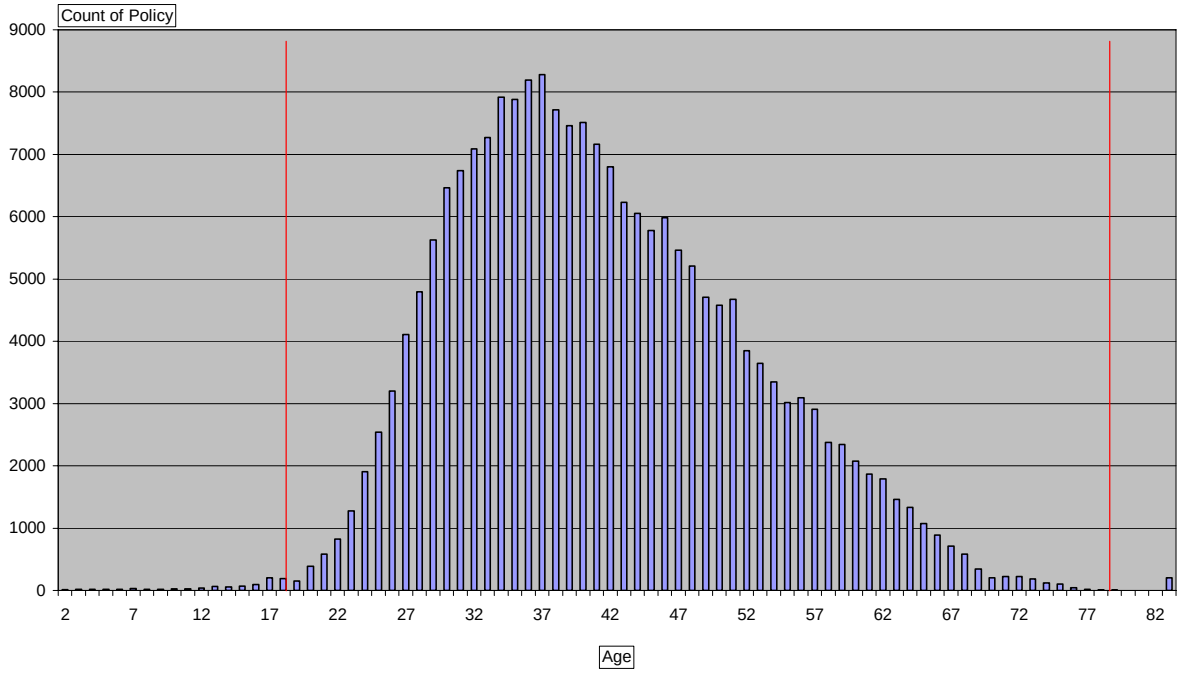

Figure III-9 Number of Policies by Age

A child under 18 can not have an individual insurance unless he/she is under coverage of his parents' family policy. The records having age less than 18 are erroneous records and having age over 80 are also suspicious since companies do not prefer to insure in these ages unless the insured has renewal guarantee. Moreover renewal guarantee concept came to market only 11 years ago.

When these records are analyzed, it is found out that almost all of the policies are travel obligatory insurance products. Since these policies are obligatory, children and elder people have to be insured. If effect of this short term insurance product is eliminated, graph changes as below:

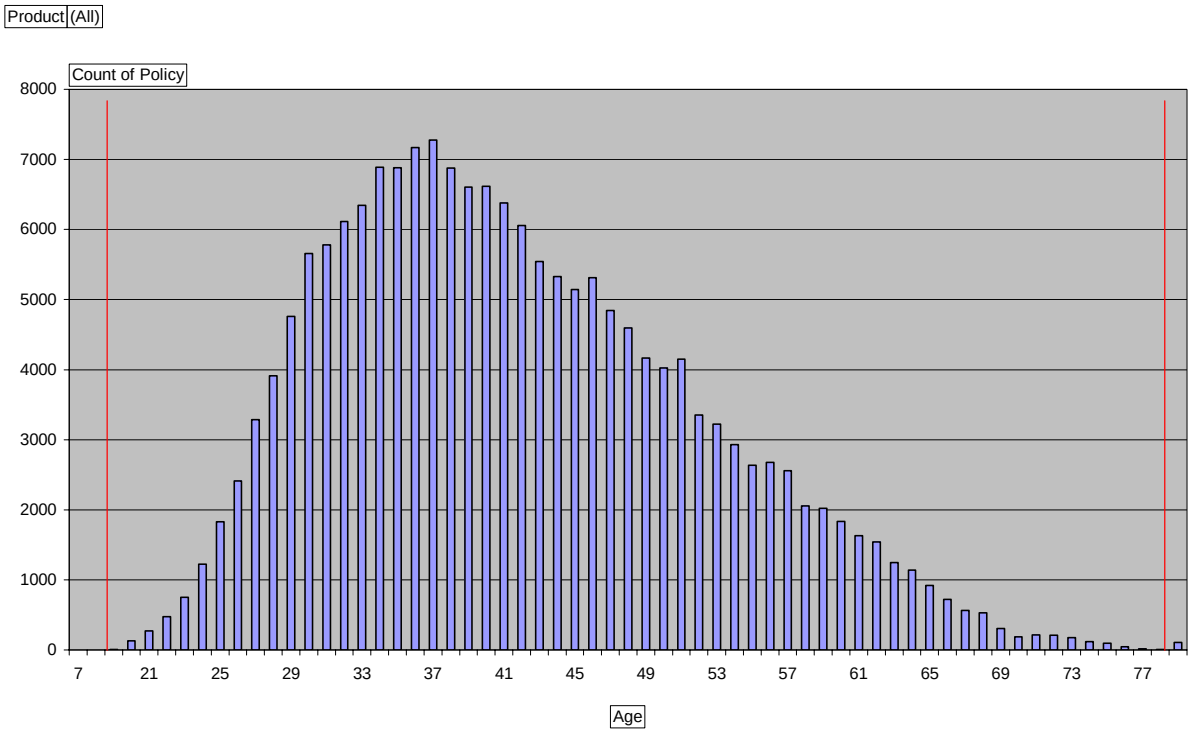

Figure III-10 Number of Policies by Age (Product Limited)

Only 2 records remained under age of 18 and 110 records with no information. Moreover about 200 records are missing age value. Since these records constitutes only 1 ‰ of total, these records are neglected in analysis.

III.2.2.9 Gender/Sex

Sex is another important attribute in health insurance. However, the data is stored in databases as male, female and child. Since a child can not buy an insurance policy except travel insurance, the data accepted to be fairly clean. The products except travel insurance are traced for child customers that are wrong. Only

5 records are detected and corrected manually by getting use of their name. About 200 missing records are disregarded in analysis.

When data set analyzed except travel policies, 60% of insureds are males. This is not a surprising situation that females are more risky than males and as a consequence of this premiums are higher for females. This situation may change the equilibrium to males hand side in number records. Moreover, another reality for Turkish society is that generally economical decisions are made by males. As a result, generally family policies are on male although policies cover their wives and children.

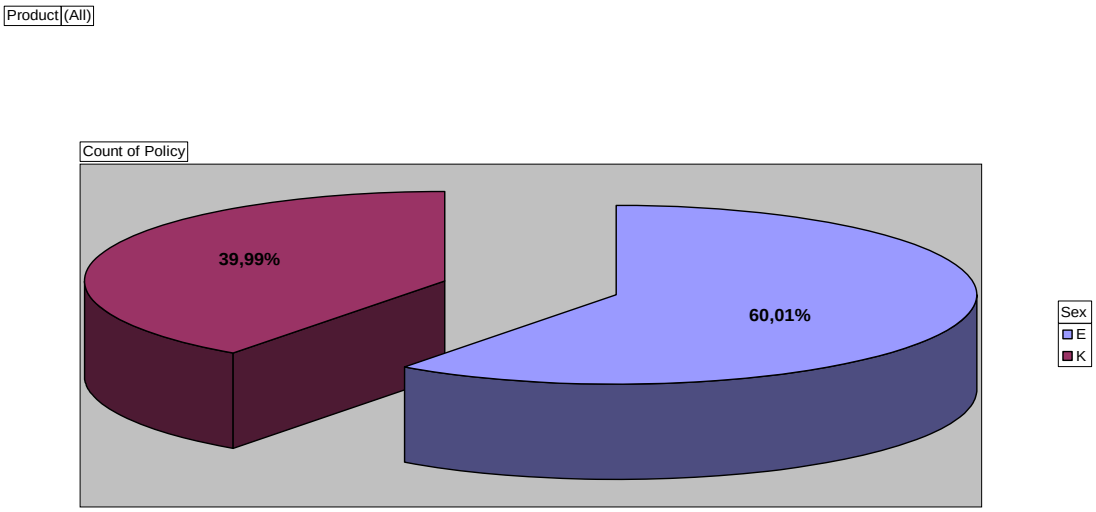

Figure III-11 Distribution of Policies by Sex

III.2.2.10 Marital Status

The attribute indicates if the insured is married or not. The categories are:

- E: Married
- B: Single
- D: Divorced

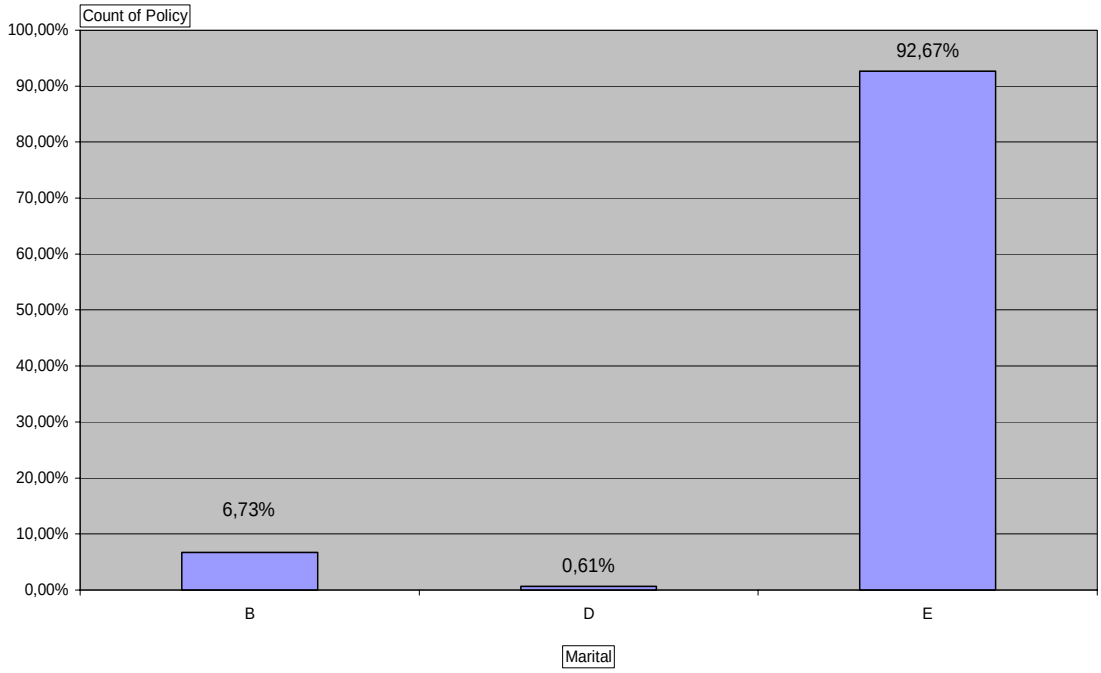

Figure III-12 Distribution of Policies by Marital Status

However, as it can be seen in the graph, great majority that is 93% is married. In order to have a notion that the data is reliable or not, number of insureds under those policies is also graphed below.

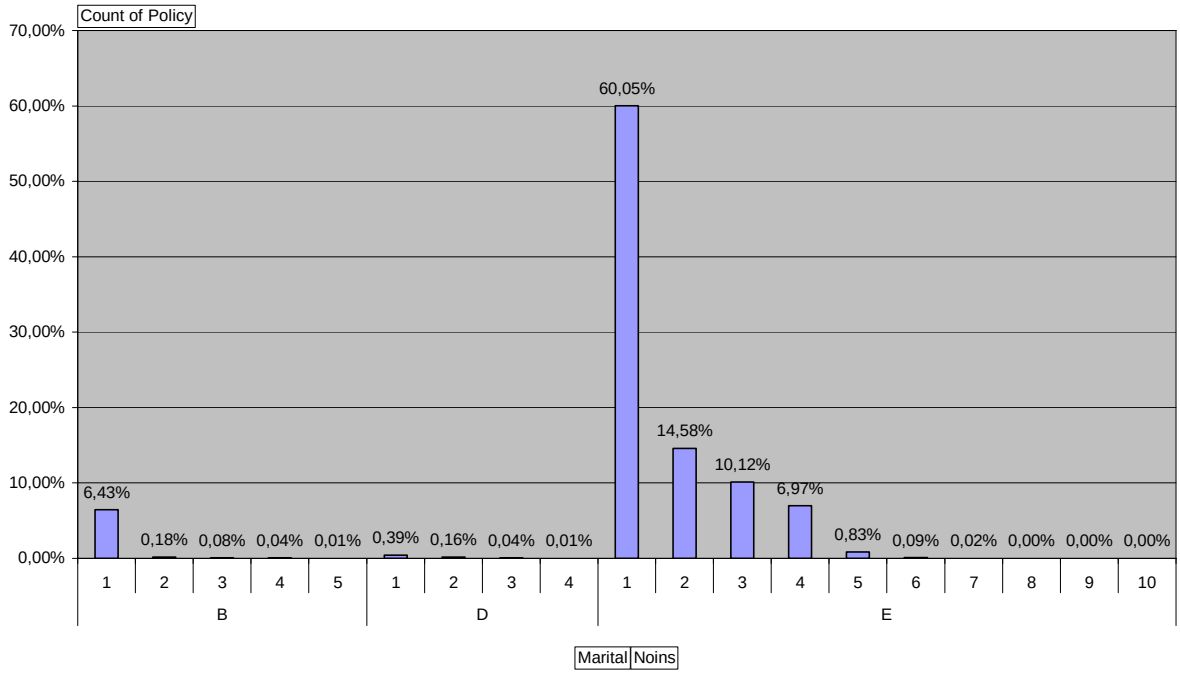

Figure III-13 Number of Insureds per Policy by Marital Status

When singles are concerned, minority is possible to have more than one insured. These are clients insuring their parents generally. Also divorced clients also generally cover their children under their policy. However, 60% of majority having no other insured under policy but married clients are suspicious. The "married" is coded in the main system with "0" and it is found that zero is the default value. Because of the reliability of attribute is low, the variable is excluded from data set.

III.2.2.11 Property/Casualty (P/C) Insurance

Health insurance and P/C insurance differ from each other by the process, claims frequency, customer behavior etc. Moreover, the regulations in 2002 which are mentioned in "II.7 About Private Health Insurance" section that separated health branch from life branch companies lead companies to go on health insurance in P/C insurance companies. As a result, health system was transferred from Life companies to P/C companies in these years and health branch became a branch having a completely different system in these companies.

Since health and P/C insurance branches have separate database and system, Tax Identification Number (TIN) is decided to be the best identifier of cross-sale customers. In order to understand how successful can be this TIN matching method, TIN field in both databases are analyzed. Every year there are 0.6-1.0 million records to be analyzed. It is seen that Property/Casualty branch in 2001 half of the records do not contain a TIN and in a decreasing order less than 10% is missing in 2004. Since yearly analysis is impossible with this data, four year records are joined to form a Property/Casualty insurance client list within anytime in 2001-2004 years. Also in health insurance database, since especially prior years' policy data has considerable amount of missing TIN values, missing values are tried to be completed as much as possible using Company's Customer Identification Number (CID). However, another problem is also recognized that a TIN can repeat on different client records being a clue for non-ignorable number of incorrect TIN records. In order to cope with this situation, only both name and TIN matching records are considered as cross-sale clients. This situation ensures correct matching but contains a risk of biased sampling of both branch customers by missing customers having incorrect TIN and name pair. However, by this operation about 13% of policies found to be belonging to cross-sales customers.

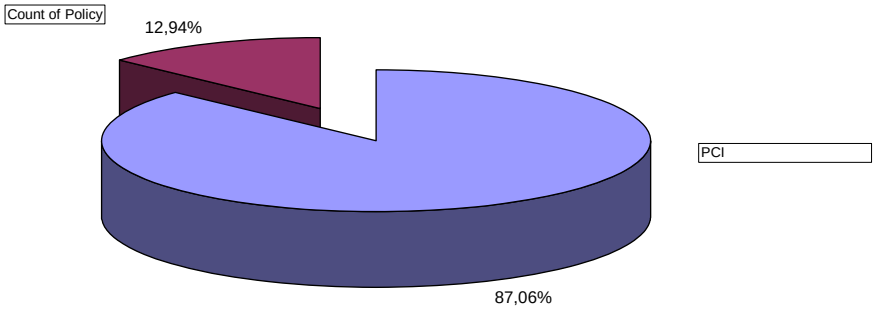

Figure III-14 Distribution of Policies by P/C Cross-sales

When the percentage of clients is graphed, 10% of health insureds are also Property/Casualty client.

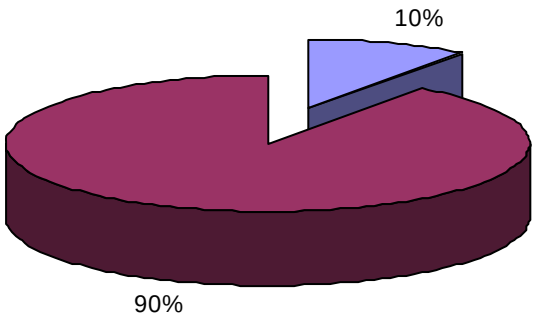

Figure III-15 Distribution of Clients by P/C Cross-sales

III.2.2.12 Location

Although Region generally gives information about the location, more detailed attributes City and District are also focused as a candidate to data set. City attribute has only 1‰ of missing values. The abnormalities ranging out of 1-81 city codes are corrected since these were Post Codes written in the wrong field. The resulting graph below indicates the overwhelming dominance of Istanbul with about 60% and totaling together with other big cities İzmir and Ankara to 82%.

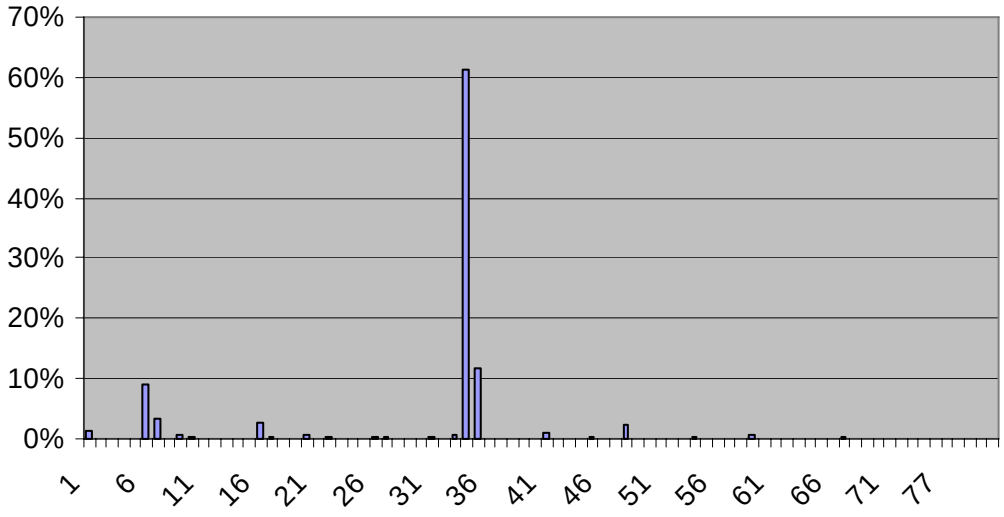

Figure III-16 Distribution of Policies by City

District attribute is more problematic with 23% missing values. Furthermore, the field is free text variable causing many typing errors and different styles like 1.Levent, 1Levent, Levent etc. Since further cleaning process is needed for district attribute and region/city attributes are adequate for this level of analysis, district variable is not considered in the analysis.

III.2.2.13 Occupation

Occupation attribute is optional to get the information from customer and both personnel inputting the data and sales channels are not sensitive to this information. However, since the field shelters valuable information, the field is explored. After abnormal data is regarded as missing, manual corrections are performed to obvious typing errors like extra or missing characters in occupation coding and the similar occupational codes are grouped to obtain understandable data, the below mentioned graph is reached. As it can be seen 41% of data is missing and with 23% self-employed customers are major clients.

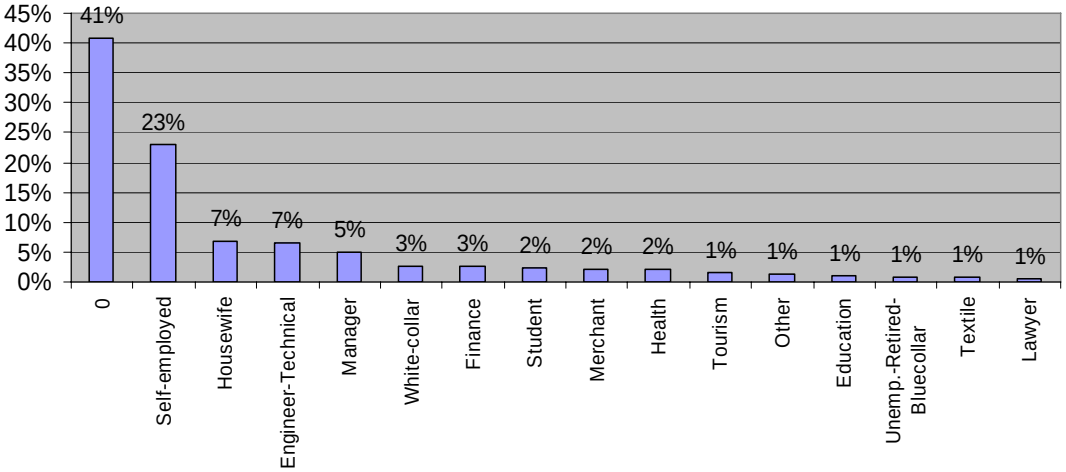

Figure III-17 Distribution of Policies by Occupation

III.2.2.14 VIP

VIP flag indicates the customer is a "Very Important Person" to be served specially. As an indicator of a different group from the population the information of VIP is also added to data set although VIP customers constitute a small portion of the total (about 2% of policies).

III.2.2.15 Number of Insureds

The policies may contain more than one insured like family policies. The number of insureds affects the premium and claims. The distribution is given below indicating almost 2/3 of policies have single insured while others family policies.

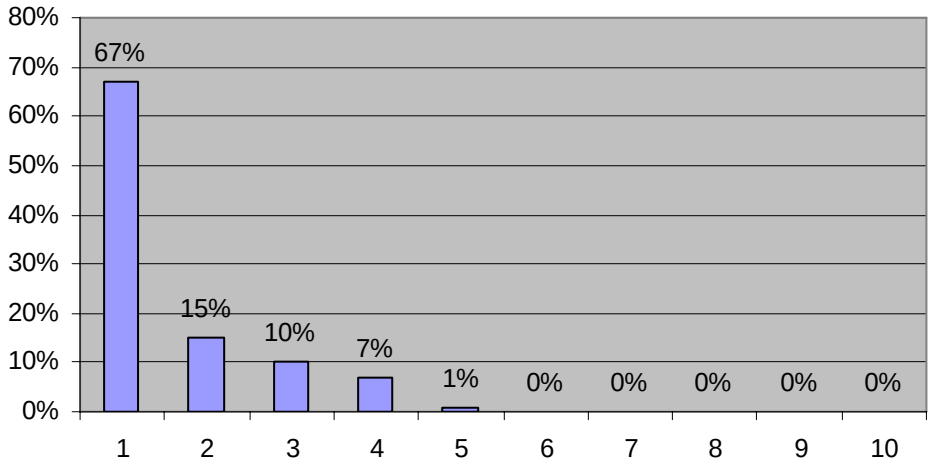

Figure III-18 Distribution of Policies by Number of Insureds

III.2.2.16 Premium

The attribute gives the price of the policy. The records are traced for abnormal data with less than or equal to zero premium and with very high premium. Only 8 records are detected to be abnormal and 6 of them identified as trial entries by IT department. While these records are extracted from data set, other two erroneous records are corrected by other system records. The premiums are rounded to be integer to ease computation.

III.2.2.17 Payment term

The payment transactions are counted for each policy to reach the number of installments realized. Since partial payments change the number of installments; policies having one or two payments are accepted as paid in advance and others by installments.

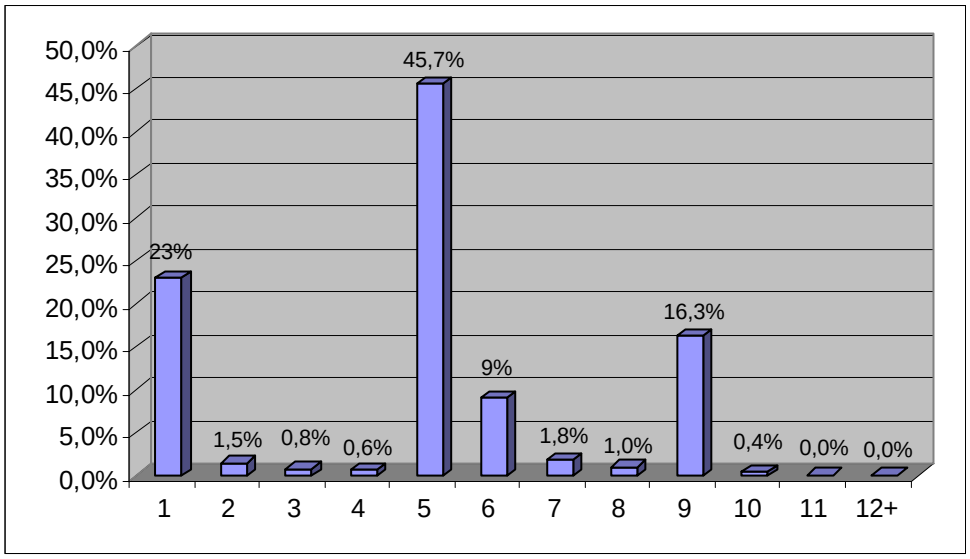

Figure III-19 Distribution of Number of Installments of Policies

Payment term is kept in two fields for further analysis. One contains the indicator of in advance payment (P) and payment by installments (T). The other attribute contains the detail of number of installments.

III.2.2.18 Payment Method

Since every payment can be made by a different payment method, each policy may have different payment type for each installment in the data. About 4 % of

policies had been used more than one payment method and only less than 1 ‰ has more than two payment methods.

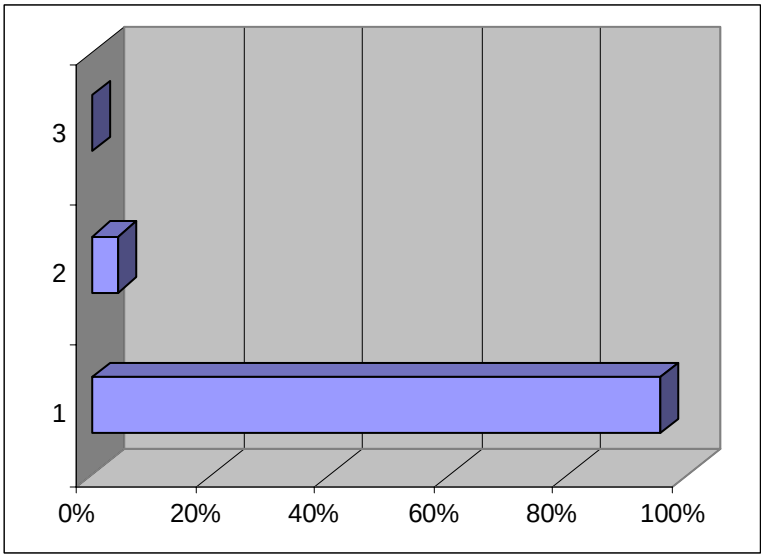

Figure III-20 Number of Payment Methods Used per Policy

In order to set one payment type for each policy, it is accepted that the most used payment method for a policy is the customer's payment method for the policy. The rational for this simplification is due to the different payment method is usually only the first payment that needed to be paid for the insurance take effect and other installments in majority has single payment method. About 95% of the multiple payment method records had been used a different payment method only once. Records having equal number of usage of two payment methods were negligible, about 1 ‰ of data, hence randomly one most used payment method is selected.

The payment methods used are:

- H: Bank transfer
- K: Credit Card
- N: Cash
- O: Other (Cheque, Security, automatic bank payment etc.)

The mostly used methods are credit card, transfer and cash, other methods are around 1% constituting others class.

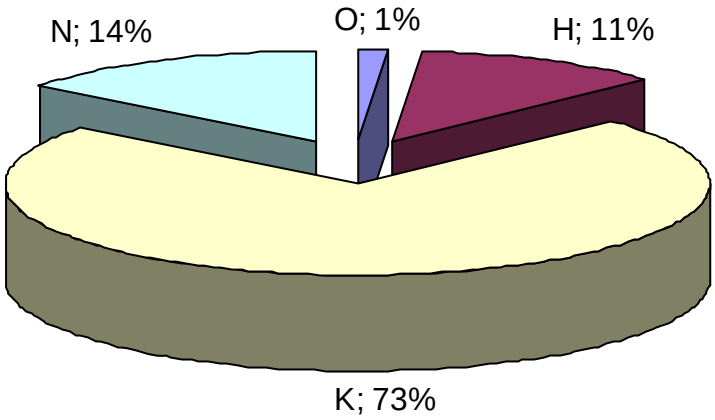

Figure III-21 Distribution of Payment Methods by Policies

III.2.2.19 Number of Claims

Claim records of each coverage are counted for number of claims. The distribution is as expected with majority in no claims area and others are in logarithmic order smoothly decreasing while number of claims increasing.

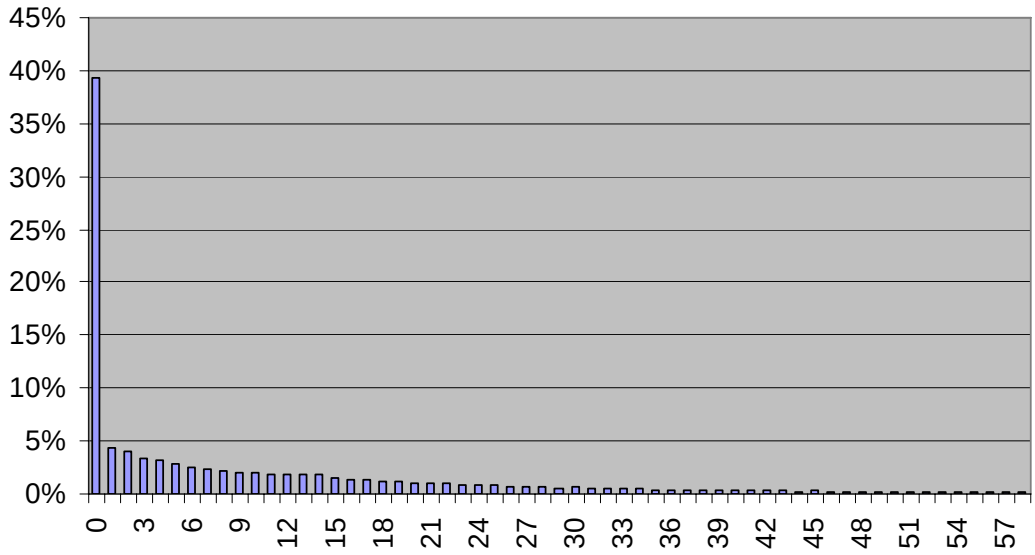

Figure III-22 Distribution of Number of Claims per Policy

III.2.2.20 Claims

The attribute indicates the claims paid to the policy. The records are traced for abnormal data with negative values and extreme values. The records constitute zero values equal to the number of no claims records extracted in the previous section

accordingly. No abnormal values could be detected in the first sight. Also the claims are rounded to be integer to ease computation.

III.2.2.21 Medical Information

When the cases of additional premium and discounts due to many reasons like risk, loyalty, claim amount, profitability, habits (alcohol, cigarette etc.), way of life (extreme sports, risky job etc.) are analyzed, since these cases are processed before the risk acceptance with additional notes, generally analyzable data is very rare. Although there are about 40 attributes, many of them do not contain data. The attributes are analyzed with company officials, as a consequence, keeping the special no claims discount attribute separate, all discounts and additional premiums are gathered under discount and additional premium. These fields are designed as Yes/No variables.

In addition height and weight values which are critical indicators in health industry are extracted but it is found out that 98.3% are missing or abnormal. Consequently, it is not possible to get use of these attributes.

III.3 Data Preparation

III.3.1 Model Set

Model set is the set of data to be used in data mining analysis. Through exploration, model set to start analysis is formed with below mentioned attributes and more than 200,000 records.

Table III-2 Model Set Attributes

Attribute	Explanation
Policy	Policy No
Start	Policy beginning year
Product	Product group
Year	Renewal Year
Region	Up-to-date Sales region
Channel	Grouped Sales Channel
CID	Customer Identification No
Age	Age
Sex	Sex
City	City Code
Occup	Occupation code

Attribute	Explanation
VIP	Very Important Person code
Noins	Number of family members insured
Premium	Policy premium
Install	Number of payments (Installments)
Payment	Payment method: In advance-Installments
Paytool	Payment tool
ClaimNo	Number of claims
Claims	Claims amount
PCI	Property/Casualty client
NCD	No claim discount
AddPr	Additional Premium Charged (Risky)
Discount	Discount granted

The data sets are generally separated into two or three sets. Model Set includes "Training Set" that is used to build an initial set of models, "Validation Set" that is used to choose best data mining model or to adjust the initial model to make it more general and eliminate over fitting in training set and "Test Set" that is used to determine how the model performs. Generally 60% of main data set is separated for Training, 30% to Validation and 10% to Test if the data set is large. In many cases the validation set is disregarded and validation and test is carried on test set with a separation of 80% to training and 20% to test.

Since a classification analysis is carried out and not to lessen the training data, only test set is separated in our model. It is preferred to select the test set from the last year since it is expected to present today most. Moreover, in order not to reduce the effect of recent data in training set, 20,000 records are selected randomly which constitutes about 10% of data.

III.3.2 Data Cleaning

Dirty or inaccurate data must be avoided if results are to be accurate and useful. All data has problems however problem importance varies with the data mining technique. For example, missing values and outliers do not cause too much trouble for decision trees but it is not the case for neural networks. [10]

Working with missing or incorrect values is a difficult process. Some data mining algorithms are capable of treating "missing" as a value and incorporating it into rules but some not. None of the possible solutions can preserve the true distribution of the variable. Some solutions are listed below.

- Extracting records with missing values from data set causes bias because it is possible that such records are distributed randomly.
- Replacing the value with a value such as the mean, median or the most common value, changes distribution of data.
- Replacing the missing value with an unlikely value is worse because the data mining algorithms will take into account this abnormal value. i.e. age=999
- Predicting the missing value using a decision tree or neural network may be a solution.

In this study, the incorrect or missing values are tried to be filled by more reliable data from another database. These corrections increased the quality of data and also usable data set. Remaining problematic data is neglected to simplify the analysis although there is a risk of biased sampling.

Since exploring and cleaning the data are interrelated, the cleaning job performed is explained in Data Understanding section. Some examples of work done are under sex attribute, missing and not appropriate (other than M or F) records are corrected by using policy insured data table in the main system. City codes are corrected using post codes. Organizational change effects are reflected to the previous years' records in the region attribute.

III.3.3 Data Transformation

After setting model set and fixing major data problems, the data is not prepared for analysis. Transformation includes further detailed cleaning and adding derived fields to bring information to the surface. This stage may also involve removing outliers, binning numeric variables, grouping classes for categorical variables, applying transformations such as logarithms, turning counts into proportions, and the like.

As previously explained in exploring stage products and sales channels are grouped for the analysis. The reason for grouping products mainly stems from the product changes in time. The products can be differentiated with minor changes to adopt conditions or radically changes may occur in accordance with marketing

decisions. The standardization of these products and consolidating products with minor differentiations is done by grouping. Moreover, sales channels are not subject to be analyzed in detail, the major sales channels are adequate for this level of analysis and accordingly sales channels are grouped to main branches as Agencies, Banks and Direct sales.

Since the quality of occupation attribute is low due to changing behavior of the selection of occupations in the system in time, the corrections and simpler classification is constructed manually to prepare the attribute to analysis.

Payment term is a derived field from the data of payments whether it is paid in advance or by installments. The aim of the study does not require the installment detail at this stage although the field kept in the data set. Summarized information of payment can be a variable in descriptive models.

Many insignificant records are consolidated under “Other” group like rare payment methods and rare occupations to simplify the analysis.

Age is another attribute to be grouped since analyzing each age separately does not provide any extra information but reducing the occurrences in each class. The intervals are determined by business people according to the generally accepted intervals in health insurance as level-1 0-20, level-2 21-35, level-3 36-45, level-4 46-60 and level-5 61-.

To view the histogram of premium and claims attributes the numbers are classed to 100 YTL intervals. When the claims paid per policy is analyzed, although almost 2/3 of the claims are under 500 YTL and 80% is under 1200 YTL, there are extreme claims reaching 290,000 YTL. The frequency decreases and the intensity increases while claim amount increases. For example while 80% of policies have less than 1200 YTL claims amount, they only add up to 22% of the total claim. The claim total reaches 80% of overall at 9600 YTL claims amount while at this rate over 99% of the policies are summed.

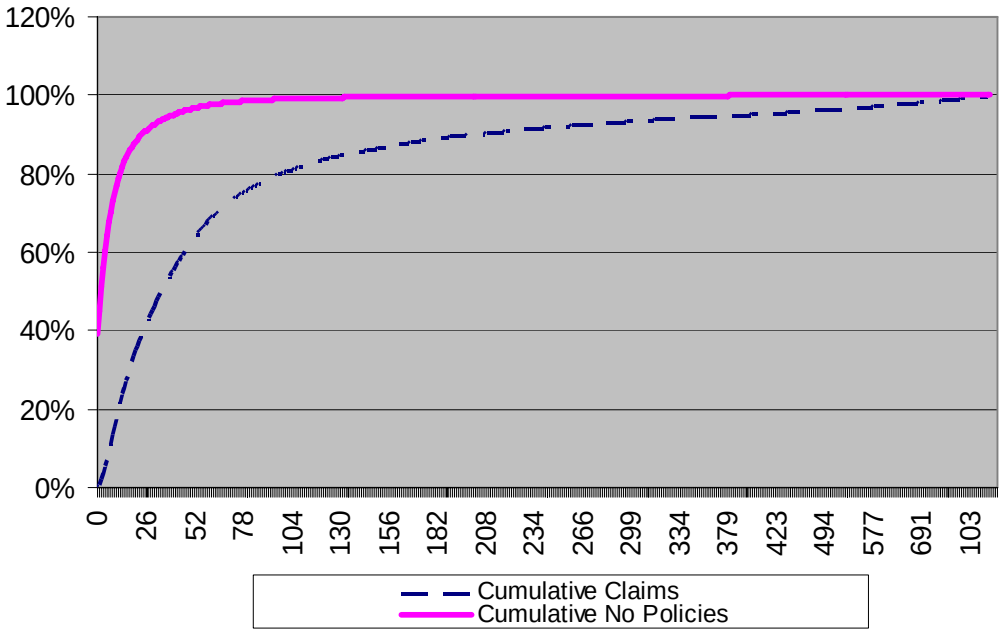

Figure III-23 Cumulative Claims vs No of Policies by Claims Amount per Policy

Similarly premium attribute has approximately same formation. The policies paying under 1900 YTL are 80% of policies these constitute half of total production.

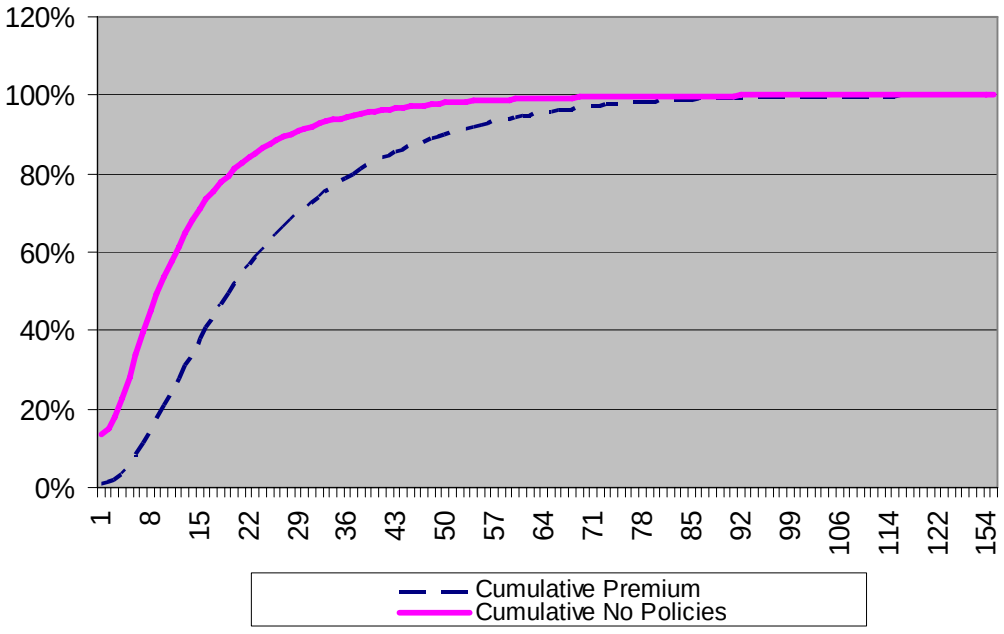

Figure III-24 Cumulative Premium vs No of Policies by Premium Amount per Policy

In insurance industry, a simple ratio of claims to premiums called loss ratio is used for the sense of profitability. Although this number only shows technical

results and disregards the overheads, personnel and other expenses, if a portion on premium is left to these expenses approximately, it is a good indicator of profitability. Profitability is one of the main concerns in the study and Loss Ratio is derived from Claims and Premium attributes to form a new variable in data set. For the convenience of reading the ratio is multiplied by 100, meaning when loss ratio is 100; premium received is equal to the claims paid.

III.4 Modeling

In this study, initial classification of customers is aimed in limitation of time and resources. Almost every marketing strategy based analysis begins with segmentation. The initial concerns of the company are profitability and revenue. In the light of this information, the definition of profitable customer profile is taken as of first degree importance.

III.4.1 Technique Selection

Decision trees are directed (supervised) method of learning used for classification. Since the data mining problem is classification, the decision trees are selected as data mining technique. Simple is the best in business, even at the first stage of a new method. The technique should be easy to explain to business people to convince them to the success of the analysis. Decision trees are easy to understand and easily converted to a set of rules. The data set contains generally classified and easy to classify data which is suitable for decision tree application. Moreover, decision trees can classify both categorical and numerical data. However, the output attribute must be categorical. This is also parallel to our aim of distinguishing profitable customer profile. Since there are no a priori assumptions in decision trees about the nature of the data, this also fits our case.

Multiple output attributes are not allowed in decision trees which is not the aim of our study at this stage. However, decision tree algorithms are unstable, slight variations in the training data can result it different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent sub-trees. This characteristic emerges the need of periodical maintenance of the model.

J. Ross Quinlan has popularized the decision tree approach with his research spanning more than 15 years. The selected decision tree algorithm C5.0 is the latest commercial algorithm of Quinlan. Since the algorithm is commercial, no published detail about the algorithm is available. C5.0 has been designed to analyze substantial databases containing thousands to hundreds of thousands of records and tens to hundreds of numeric, time, date, or nominal fields. To maximize interpretability, C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand than neural networks. The algorithm is easy to use that does not need any special knowledge of statistics or data mining. [48]

When the C5.0 algorithm (Black) is compared with previous C4.5 algorithm (Grey), the below mentioned graphs are stated for a similar sized data to the data set of this study. The C5.0 provides dramatic performance improvement according to the graphs in Figure III-25 C4.5 vs C5.0 Algorithm. Boosting is another addition to the new algorithm. Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy. In tests, the 10-trial boosting decreased the error rate by 10 to 40%. However, the method may decrease accuracy when the data is noisy. [49]

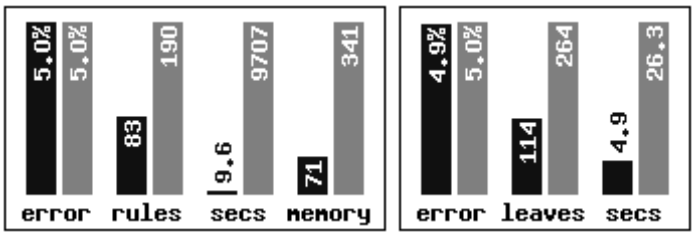

Figure III-25 C4.5 vs C5.0 Algorithm

A general approach of C5.0 algorithm can be summarized below [51]:

- Step 1: An attribute is selected to differentiate the output attribute values best. The C5.0 algorithm uses the Max-Gain method of selecting the best attribute that is choosing the attribute that has the largest expected information gain. In other words, algorithm tries to select the attribute that will result in the smallest expected size of the sub-trees rooted at its children. [50]

- Step 2: A separate tree branch is created for each value or range of the chosen attribute.
- Step 3: The instances are divided into subgroups so as to reflect the attribute values of the chosen node.
- Step 4: For each subgroup, the attribute selection process is terminated if:
  - All members of a subgroup have the same value for the output attribute, terminate the attribute selection process for the current path and label the branch on the current path with the specified value.
  - The subgroup contains a single node or no further distinguishing attributes can be determined. As above, the branch is labeled with the output value seen by the majority of remaining instances.
- Step 5: For each subgroup created in Step 3 that has not been labeled as terminal in Step 4, the above process repeated beginning from Step 1.

Some general parameters are explained below in C5.0 algorithm:

- Output and Input variables

Output can not be continuous since C5.0 allows only categorical attributes for output however input variables can be continuous and only one output variable should be chosen.

- Confidence

The confidence value is used when pruning the tree that removing branches that are deemed to provide little or no gain in statistical accuracy of the model. 25% confidence works reasonably well in most cases. However, if the actual error rate on real data (or the error rate on cross-validation) is significantly higher than the error rate on the training data, the confidence factor can be decreased to cause more drastic pruning and a more general model of the data. To obtain a more specific modeling based on the training data, the confidence factor can be increased, which will decrease the amount of pruning that occurs.

- Limitation to the instances of leaf

It is generally the number of instances that must be present in the training data for a new leaf to be created in the decision tree. A higher number will create a more generalized tree and a lower number will create a more specialized tree.

- Folds for Cross-validation

If the test set does not exist, cross-validation is used to test. If the number of folds for cross-validation is x, then (x-1)/x of the training data is used to construct the model and 1/x of the training data is used to test the model. This process is then repeated x times so that all the training data is used exactly once in the test data. The x different error estimates are then averaged to yield an overall error estimate. While extensive tests on numerous datasets have shown that ten-fold cross-validation is one of the best numbers for getting the most accurate error estimate. Decreasing the number of folds will likely decrease the amount of time it takes for the decision tree to be generated, and increasing the number of folds will likely increase the amount of time it takes. Increasing the number of folds will create a larger dataset for the training data, which may increase accuracy of the decision tree; similarly, decreasing the number of folds will create a smaller dataset for the training data, which may decrease the accuracy of the decision tree.

- Variable misclassification cost

If the cost of each variable classified false is different, it is also possible to load it to the algorithm. The algorithm minimizes the cost of misclassification.

III.4.2 Building Model

In order to prepare the data for C5.0 algorithm, the target variable about profitability which is loss ratio derived variable is converted to a classified field. This continuous attribute is divided into 4 classes with respect to the business understanding.

- Class 1: Loss ratio<50%, which are profitable, most favorable customers
- Class 2: 50%=<Loss ratio<80%, which are less profitable, favorable customers

- Class 3: 80%=<Loss ratio<120%, which are unprofitable, unfavorable customers
- Class 4: Loss ratio=>120%, which are highly unprofitable, most unfavorable customers

The identification tags that Policy number and CID are excluded from the input set in the analysis. Moreover, the inputs of derived target variable that premium and claims data are also excluded.

In the model, group symbolics is selected to combine symbolic values that have similar patterns with respect to the output field to have easy to understand rules. The model is run on the whole data and cross-validation is done with 10 folds. The algorithm is run on seeking generality instead of accuracy in order to get more generalized tree and generalized rules accordingly.

The tree obtained from the C5.0 algorithm is presented below:

ProductL2 in ["CALT" "CORTA" "CUST" "W"] [Mode: 1] (179.877)
Noins <= 1 [Mode: 1] => 1 (113.197, 0,698)
Noins > 1 [Mode: 1] (66.680)
AgeGr in [1 2] [Mode: 1] (11.610)
Sex in ["E"] [Mode: 1] => 1 (3.729, 0,537)
Sex in ["K"] [Mode: 1] (7.881)
Addpr = 1 [Mode: 1] => 1 (4.039, 0,422)
Addpr = 0 [Mode: 4] (3.842)
NCD = 1 [Mode: 4] => 4 (1.204, 0,488)
NCD = 0 [Mode: 1] => 1 (2.638, 0,363)
AgeGr in [3 4 5] [Mode: 1] => 1 (55.070, 0,592)
ProductL2 in ["S"] [Mode: 1] => 1 (29.711, 0,998)

Graphical representation is presented below:

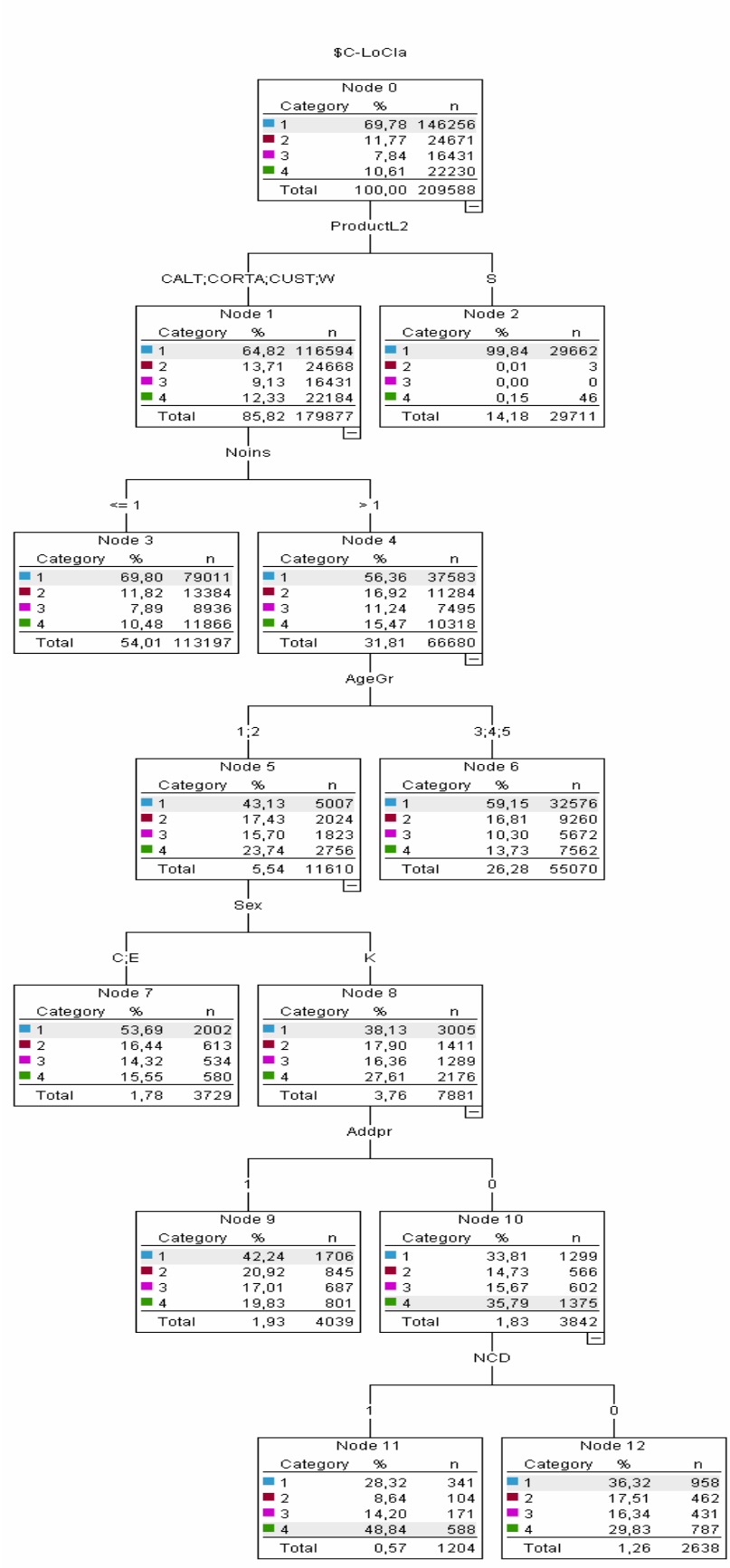

Figure III-26 Tree Presentation

The tree depth is 6 levels which is quite general and cross-validation mean is 69.9 with 0 standard error. However, although 69.9 % of the records are classified correctly, a high bias is seen in the tree that almost every record is resulted as 1 in the algorithm. The coincidence matrix shows the situation better:

Table III-3 Coincidence Matrix

Real values/Calculated	1	4
1	145915	341
2	24567	104
3	16260	171
4	21642	588

The overwhelming majority of Class 1 in training set leads the algorithm to result in calculating almost every record to be Class 1. Many modeling techniques have trouble with such biased data because they will tend to learn only the frequent data and ignore the rare and also the decision tree faces the same problem in the study with Class 1 which is 69.78% of the whole data in this case. If the data are well-balanced with approximately equal numbers of classes, models will have a better chance of finding patterns to distinguish profiles. In order to overcome this problem and no sequential analysis will be performed on this data, balancing is applied to the classes with factors equalizing the incidences of values. Since balancing by boosting generates about 600,000 records to operate and capacity of the PC used couldn't cope with the amount of data, the data balanced by reducing.

Table III-4 Balancing

Value	Proportion	Balance Factor
1	69.78%	0.11234410895963243
2	11.77%	0.6660046208098577
3	7.84%	1.0
4	10.61%	0.739136302294197

The resulting tree has 19 levels. However, the tree performance is poor with 46.33% of the records are correctly defined, although coincidence matrix shows the bias is reduced.

Table III-5 Coincidance Matrix (Balanced Set)

Real values/Calculated	1	2	3	4	%Correct
1	73600	29030	21414	22212	50%
2	4765	9861	5361	4684	40%
3	2874	4192	6113	3252	37%
4	4809	5187	4698	7536	34%

The accuracy parameter is increased to see the limitations of accuracy and only 47% correctness could be reached. In order to increase accuracy, reclassification is performed on Loss Ratio to reduce the classes to profitable, medium and unprofitable. Moreover, misclassification costs are set again since defining an unprofitable customer as profitable or vice versa are dangerous instead can not define and leave neutral. The cost of classifying these conditions is predicted to be twice the other misclassifications. Even the misclassification cost changes could not reduce the bias of 74.4% majority of Class 1 leading classification results to 98% Class 1 and balancing performed again.

An evaluation ratio is determined with respect to the important errors of misclassification. The ratio is the class 3 members assigned to class 1 to the correctly defined class 1 members. For class 3, the ratio is vice versa. The cost of misclassification between these classes directly affects profitability.

The aim of the remaining part of study is to increase the records in neutral part in order to reach more accurate profiles in Class 1 and 3 while reducing coincidences of these two classes. Some iterations with cost values are given in below tables.

Table III-6 Misclassification Cost of [1->3] and [3->1]=2, Balanced C5.0 Coincidance Matrix

Class	True	False	False/True
1	23%	2%	10%
2	96%		
3	5%	2%	42%

Table III-7 Misclassification Cost of [1->3] and [3->1]=1.5, Balanced C5.0 Coincidance Matrix

Class	True	False	False/True
1	41%	8%	21%
2	74%		
3	22%	14%	63%

Table III-8 Misclassification Cost of [1->3] and [3->1]=1, Balanced C5.0 Coincidance Matrix

Class	True	False	False/True
1	52%	17%	33%
2	59%		
3	34%	23%	69%

Although the True values increase False values increase faster when the cost is reduced. In order to determine the optimal point of cost, the profit per policy is analyzed. The profit is estimated by the formulation ((80%xPremium)-Claims). The mean of Class 1 is 6.3 while of Class 3 -27.8. One unprofitable policy can be neutralized by approximately 4.4 profitable policies.

A further improvement is tried by using boosting in the model. The mean increases slightly but the error rates defined during study that is Class3 defined as Class1 divided by Class1 defined true (Rate value for 1) and Class1 defined as Class3 divided by defined true (Rate value for 3) acts differently. The tree reduces a few levels by boosting and Class1 Rate performance improves by 4% to 36% while Class3 Rate performance decreases by 9% to 60%. When Class3 Rate performance is evaluated, it was acceptable with 30% and 36% in Cost=2 and Cost=1.5 cases respectively, however bad performance with about 50% rate in boosting. On the other side, Class1 improves by 36% in Cost=2 case and not as much improvement could be obtained in other cases.

Table III-9 Rates for Different Misclassification Costs

Models	Cost=2			Cost=1.5			Cost=1		
	True	False	Rate	True	False	Rate	True	False	Rate
Boosted 1	22%	1%	6%	40%	11%	27%	51%	17%	33%
Normal 1	22%	2%	10%	33%	9%	28%	50%	18%	37%

Models	Cost=2			Cost=1.5			Cost=1		
	True	False	Rate	True	False	Rate	True	False	Rate
Boosted 2	98%			70%			60%		
Normal 2	97%			79%			59%		
Boosted 3	3%	1%	48%	27%	14%	52%	34%	22%	66%
Normal 3	5%	1%	30%	23%	9%	36%	36%	22%	60%

The real profit averages of profitable and unprofitable classes are analyzed to determine the magnitudes of the profit, since loss ratio provides only information whether the policy is profitable or not. The average loss of unprofitable policies is 4.4 times of the average profit of profitable policies. In order to be profitable at the end, the true classifications should be at least 4.4 times of the false classifications in Class1. Since the accuracy of Class3 is low in every case, the Class1 Rate is the primary concern. From low cost levels to high, the model loses almost half of the correctly defined Class1 members. However, the false classifications are improved significantly from 18% to 2% which is desirable. Boosting the model improves the model perfomance in both increasing correctly defineds and reducing incorrectly defineds. As a consequence, boosted and misclassification=2 balanced model which has only 6% False-True Rate is selected.The tree is presented in Appendix B.

The resulting tree depth is 21 levels and the model is cross-validated by 10-folds. The cross-validation mean is 42.6 with standard error of 0.2. In general understanding the mean 42.6 seems to be low and slightly worse than previous models, but the risk of defining unprofitable customers as profitable and vice versa is significantly reduced which is a critical risk that can sweep all of the profit gained from analysis. The usage of rate is preferred to reduce costly critical mistakes. As a consequence, the evaluation will be carried out with the rate of false classifications to true classifications and the percentage of correctly determined true values.

III.4.3 Further Modeling

In the first modeling effort above, some problems of biased real data reduces the accuracy of the model. Moreover, tree produced is further branched to increase accuracy resulting in too detailed, hard to interpret and apply rules. Hence, an

alternative model is developed to overcome generality, profit magnitude and bias problems.

Balancing by under sampling and cost-sensitive learning methods are used to cope with unbalanced data in the first model. Eliminating redundant information, re-sampling, combination, analyzing classes separately are some other ways for unbalanced data. Hybrid method are also tried to overcome the problem. Moreover, the distribution in each class is another problem of imbalance. Highly variant data makes difficult to derive accurate results. [53][54][55][56]

In order to increase generality and having less branched tree, inputs are reviewed again to combine, to disregard irrelevant and to reduce variances in data. Renewal year is transformed to new policy/new customer/old customer set to increase the information that no claims discount brings and distinguishing new and old customers. If renewal year is 1 Policy is new, if renewal year is 2 or 3 the customer is new and if renewal year is higher customer is loyal. Moreover, a customer new/old flag is generated. Region field is disregarded since city variable better explains the location and region usually becomes redundant. VIP attribute constitutes a small group of customers and in a generalized model, the field brings less information to be eliminated. Payment related fields that are payment tool and method are eliminated since the information is not usable at acquisition level of analysis. While no claims discount is kept in data set, discount variable that has general discount information is disregarded.

As known from the knowledge obtained from the first model, product S is purely profitable. In order to reduce the number of records and run the algorithm effectively with less unbalanced data, the product S is deselected from data.

A solution for high variance in classes especially in the minority class and class imbalance, shortly imbalances within or among the classes, is proposed by undirected learning within classes to both reduce in-class imbalance and among classes imbalances and further running classification algorithms. [57]

The classes are fed into clustering algorithm to minimize the in-class imbalance. The K-Means algorithm, which was explained in “II.5.2.5.1 K-means Clustering”, is used for clustering. After each class is clustered into subclasses, the sizes of the clusters are equalized in balancing. Since balancing with factors carries

the variances to the new set, the similarities are increased by using clustering based re-sampling. After subclasses are equalized in number, the classes are balanced in number generating less varied data set.

Generally sample size affects the variance, bias or error. While sample size increases the variance tend to decrease statistically significantly but it can not be said to bias either. Although bias seems to decrease there can not be found adequate evidence in the study of Brain and Webb. For the complexity of results aspect, it is surely expected that as training sets increase there will be more information with less variance available to the classification algorithm, and this will allow a more complex theory to be produced. The dramatic increase in complexity is also faced in the study of Brain and Webb. [52] Fairly large data used in our model is also a reason for complexity while bias could not be reduced by sample size significantly.

Although statistically enough clue can not be found in bias reduction in above mentioned study, mostly bias is reduced in samples. Furthermore, variance decreases significantly in the study. Hence, instead of under sampling, boosting to the highest number of incidence class is preferred this time. While re-sampling the rare classes above mentioned cluster supported re-sampling is used.

Table III-10 Subclass Clusters

Class	Cluster	Incidence	% in Class
2	1	5065	20%
	2	5857	24%
	3	4345	18%
	4	4260	17%
	5	5251	21%
3	1	9019	31%
	2	3992	14%
	3	4105	14%
	4	6062	21%
	5	5599	19%

The clusters produced by K-Means are represented in above table. While class 2 has fairly equal number of members, class 3 clusters have distant number of members. The cluster members in each class are equalized and then class incidence

number is equalized with class 1, which constitutes majority. The re-sampled data set is fed to the C5.0 algorithm for classification again.

The algorithm is run for the costs used in the first model and boosted algorithm is also tried for this data set. The results are given below.

Table III-11 Rates for Different Misclassification Costs (2nd model)

Models	Cost=2			Cost=1.5			Cost=1		
	True	False	Rate	True	False	Rate	True	False	Rate
Boosted 1	18%	8%	42%	57%	19%	34%	57%	21%	38%
Normal 1	18%	8%	46%	56%	18%	33%	57%	22%	39%
Boosted 2	86%			49%			49%		
Normal 2	85%			51%			49%		
Boosted 3	26%	5%	21%	51%	24%	47%	54%	25%	46%
Normal 3	27%	5%	19%	49%	23%	48%	56%	24%	44%

However, since product S is extracted and determined as Class 1 before the model is run, the success of this prior assumption is also added to the results of Class 1 in the below table.

Table III-12 Rates for Different Misclassification Costs (2nd model-Merged)

Models	Cost=2			Cost=1.5			Cost=1		
	True	False	Rate	True	False	Rate	True	False	Rate
Boosted 1	29%	6%	22%	63%	17%	26%	63%	18%	29%
Normal 1	29%	7%	24%	62%	16%	25%	63%	19%	31%
Boosted 2	86%			49%			49%		
Normal 2	85%			51%			49%		
Boosted 3	26%	5%	21%	51%	24%	47%	54%	25%	46%
Normal 3	27%	5%	19%	49%	23%	48%	56%	24%	44%

The tree depth doesn't change, however since the variables are lessened understandability is increased. The re-sampling by the help of clustering increased the accuracy of Class-3 drastically. While the correctly detected Class-3 members are increased by 1.6 to 8.5 times, the ratio is improved up to 57%. On the other hand, Class-1 presents a performance not as good as Class-3. Although, in all cases

the correctly detected incidences are higher than model 1 and generally ratio is better in cost equals 1 and 1.5 cases, the model 1 ratio far away better in cost equals 2 case which was selected in model 1 phase. While the cost parameter increases the performance of model 1 faster than model 2 leading a better score at cost equals 2. Moreover, with the information of one false is approximately equal to 4.4 true, the risk is high with these ratios in Class 1.

To summarize, although the variance is reduced by pre-clustering in re-sampling in Class-3 and Class-2 leading better accuracy, the variance in Class-1 becomes more remarkable in the model leading worse ratios at higher costs. Since the cost of a false is very high, the risk of false classifications in Class-1 can be very costly to the company though far away better results are reached in Class 3. The model achieves better results in rare incidences but the performance of majority class is not enough to be selected. As a result, the company needs more accurate Class-1 definition in this case and model -1 becomes more rational instead of 2.

With the conclusion of better results in model 1, the revised data set for model 2 is used in model 1 expecting less complexity (less level in tree) and better definition of the Class 1. In the previous efforts, it is seen that penalty of defining Class 3 incidence as Class1 is very high forcing the ratio of the selected model to be low, meanwhile the success of the Class 3 selection suffers from the enforcement. Aiming the success of Class 1 becomes more rational at the case. The revised data set is modeled with the same parameters in model 1. To note that, cost value of 1.7 is used as highest, since the cost value reaches a good level of performance and increasing cost to 2 will generate worse results by reducing true classifications. The model and results are presented below.

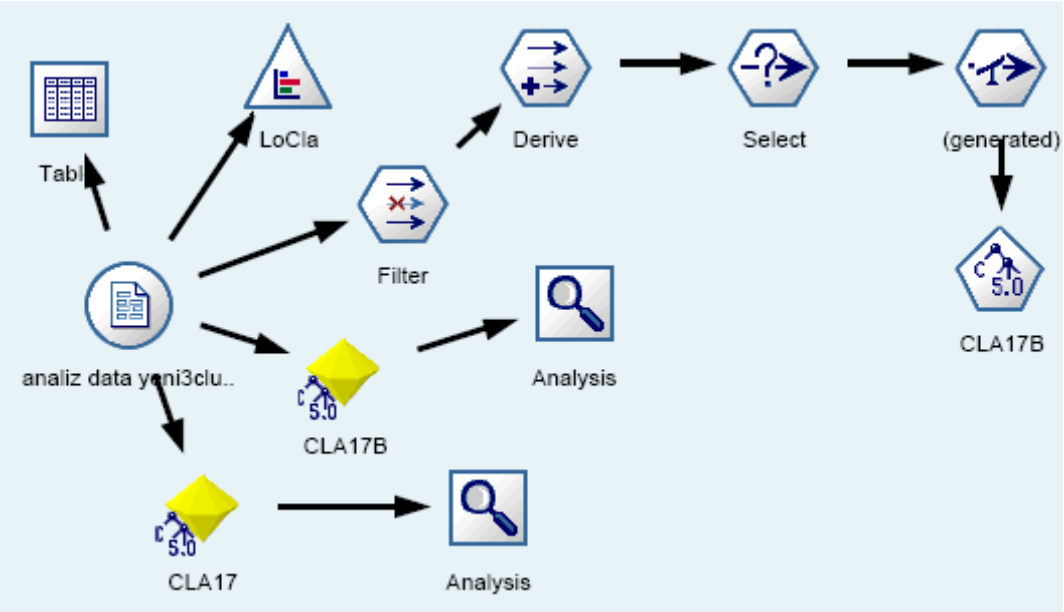

Figure III-27 Clementine Model Diagram

Table III-13 Rates for Different Misclassification Costs (3rd model)

Models	Cost=1.7			Cost=1.5			Cost=1		
	True	False	Rate	True	False	Rate	True	False	Rate
Boosted 1	52%	4%	7%	47%	6%	13%	64%	16%	24%
Normal 1	52%	4%	8%	49%	6%	13%	66%	16%	24%
Boosted 2	66%			68%			43%		
Normal 2	66%			66%			40%		
Boosted 3	14%	24%	175%	19%	23%	119%	34%	35%	100%
Normal 3	15%	23%	154%	19%	22%	115%	35%	37%	103%

The results show a better performance than the previous ones in Class 1. In general the correct values are increased by the model and false values are decreased. A continuous decrease is achieved in ratio values by the cost changes. Especially in cost level 1.7, over than half of the Class 1 members are defined correctly while having only 4% False, Class 3 values. At this level very close ratios are obtained to model 1 with almost 2.5 times incidences are defined correctly. With this model (Model3, Boosted, cost=1.7), slightly worse ratios are obtained, but much more incidences are classified. Moreover, the generalization and understandability is improved and 21 leveled tree is reduced to 14 leveled decision tree with less attributes. The decision tree of model 3 is presented in Appendix C.

III.4.4 Test and Validation

Two models are over-performed others in the modeling phase. The boosted algorithm having the cost of 2 for assigning Class 3 to 1 or vice versa and using first data set and again a boosted algorithm having the cost of assigning Class 3 to 1 or vice versa is 1.7 and using revised, combined, eliminated and generalized data set. The clustering based re-sampling model couldn't perform well in Class 1 determination which is critical.

The previously separated test set contains randomly selected 20,000 records from the most recent year which is the year 2004. The best two models which are model 1 and model 3 are tested in the test set and Class 1 performances are indicated below table.

Table III-14 Performance Rates of Models (Test Set)

Models	True	False	Rate
Model 3	26%	4%	13%
Model 1	28%	3%	10%

Classification algorithms are sensitive to the data that small changes can result different conclusions. In training sets, model 3 out-performed in detection of corrects but in test set, it is slightly worse than model 1. The other performance indicator, ratio, was slightly worse than model 1 in training set and the results are alike in test set. In the respect of generality and understandability, in the training set 14 to 21 levels, model 3 was simpler to understand. In parallel, model 3 is simpler with 6 levels while model 1 has 12 levels.

In the cost point of view, both models are still rational though performance level reduces. The cost of false classifications is 4.4 against true classifications leading ratio under 22% to be profitable in total. The models have ratios of 10% and 13%.

IV RESULTS

IV.1 Current Business

Current business has the deficiency of past data usage in developing customer life cycle business strategies. Product based campaigns are developed triggering by the demand of sales channels or with respect to the general production reports or profitability reports. The strategies are only built on these undetailed performance reports and individuals' knowledge coming from experience. Customer analytics are occasionally performed only when demanded and only on the demanded point of view. The profitability is not at the expected level and improvement can not be achieved in recent years. The strategies on acquisition of new customers are not developed.

IV.2 Results and New Strategies

The modeled part of objectives in this study aims to form a systematic approach to increase profit by profiling profitable and unprofitable customers in order to be used in customer acquisition. The study targeted to profile extremes that is medium customers which are around zero profit are not profiled.

The results of the models could not determine the minority class 3 which are unprofitable records adequately. Moreover, the cost of Class 3 is 4.4 times of the gain of Class 1. For the purpose of acquisition of new customers, defining profitable customers with less error of unprofitable ones is the main purpose. Since the Class 3 definition is not adequate, Class 1 performance is the aim and focused subject in the analysis.

IV.2.1 Interpreting Results

IV.2.1.1 Model 1

At the first glance in decision trees, one product is separate from others which is S. The product constitutes about 14% of policies and the proportion of the product in total increases every year. The product needs special attention with more than 99% profitability independent of other variables. However, when the premiums are analyzed for the product, it is seen that the premium per policy is far below the other products. Although the profitability ratio is high since profit per policy is low the product is not suitable for costly campaigns.

After a branching with number of insured, the age groups 1 and 2 are clustered to medium customers. This is a heterogeneous group having customers from each class with Class3 highest. However, the cost of the Class1 members in this cluster leads grouping in medium class which bring no information about the aim. Medium group branches will not be deducted in detail in the remaining part of interpretation.

The middle aged or higher customers having more than 1 insured (meaning they are families) buying lowest and highest coverage products in cities [19 23 26 31 39 45 46 50 52 53 55 68 81] are profitable by 66%. The middle aged or higher families in these cities are candidates of profitable customers. Since these products have the lowest (CORTA) and highest coverage (W) and respectively prices, it is important to perform directed campaign with respect to income class. However, another result shows that the customers having exactly the same characteristics in other cities representing extremely different behavior. The customers of CORTA and CALT in cities [2 3 4 5 8 11 12 13 14 15 21 24 25 27 28 29 30 32 36 37 40 43 44 49 51 56 57 58 60 62 65 66 69 70 71 72 73 76 78 79 80] are unprofitable. Only 10% of this group is profitable. This result supports the idea of city based analysis. Since every city does not have the same level of private medical health infrastructure, this may be one of the reasons.

The single insured customers of CORTA and CALT products with medical or past claims data risk are risky to be unprofitable. These customers should not be taken under coverage or premiums should be determined accordingly.

The male, single insured customers of CUST and CALT products without any medical or past claims data risk who are occupied as lawyer or in Health sector are profitable. Medium and High class products are suitable to promote to the single lawyers and health sector personnel. The female, single insured customers of CUST and CALT products without any medical or past claims data risk and living in cities [5 10 15 53 65 72] are also profitable with 59% accuracy whereas same characteristic customers living in cities [2 4 8 11 12 13 18 19 24 25 28 29 30 32 36 37 38 40 43 44 49 50 51 55 56 57 58 60 62 63 66 68 69 70 71 73 74 76 79 80] are unprofitable with 65% accuracy.

The middle aged or above CALT and CUST customers with family without no claims discount (either new customer or previous year's unprofitable or medium class customer) in cities [2 4 8 11 12 13 14 15 18 19 21 24 25 28 29 30 36 43 49 56 57 58 60 62 65 66 68 69 70 71 72 73 76 79] are unprofitable with 70% accuracy. On the other hand the customers living in [22 47 52 53 64] cities are profitable.

The medium and higher aged family customers of CORTA and W in city 6 using regional sales channels of [14 16 17 20] are unprofitable. The city 6 customers using other regional sales organizations are 60% unprofitable.

The medium and higher aged family customers of W in city 7 are profitable.

An interesting result, the medium and higher aged family customers of CORTA and W in city 10 are profitable if they have medical risk and paying extra premium whereas unprofitable if they have no medical risk and they don't pay any extra premium. Similarly customers in 17 are unprofitable if they get no claims discount whereas profitable without no claims discount. The city 20 customers are profitable if they mentioned the occupation and they are nor self-employed. The 99% of customers in city 54 paying the policy in advance are unprofitable. The customers in city 59 covered with their children are profitable in 64% of cases.

Single female customers of CALT and CUST in city 26 without any medical risk are profitable while they get no claims discount and unprofitable without no claims discount. Customers in city 27 are unprofitable unless they are self-employed and working in education industry. In city 45, customers are profitable they get no claims discount. In city 48, VIP customers are unprofitable by 80%.

Middle or higher aged CALT and CUST family customers without no claims discount in city 1, 6 and 26 are unprofitable when they have more than 2 children in their policy. In city 33, the customers are unprofitable unless they also have P/C insurance from the company. City 48 customers using sales channels in Istanbul are profitable. Middle or higher aged CALT and CUST family customers without no claims discount in city 67 are profitable.

Middle or higher aged CORTA and W family customers with 2 insureds in city 59 are profitable if in age group 4, unprofitable otherwise.

Some results of the model 1 up to level 8 of decision tree are explained in this section, and further results are not mentioned not to lose continuity.

IV.2.1.2 Model 3

This model also begins with the acceptance of product S is profitable with 99%. Other products are of distributed behaviors.

The loyalty variable is a distinguisher of the first branches. The policies of new customers who are customers not having a policy before, are profitable by 47.5% in products CORTA, CUST and W. The remaining product CALT has a behavior which is harder to distinguish.

The CALT customers in age group 4 are if they have number of insureds less than 4 and are male, they are profitable by 52%. Age group 4 CALT customers having more than 3 insured on the other hand, are profitable if they are in occupational groups [2 3 4 7 9 10 11 12 13 15] with 54%.

The CALT customers in other age groups with only 1 insured are profitable if they are members of occupational groups 13 and 15 and sold by sales channels A and B. Furthermore, the CALT customers in other age groups with only 1 insured and not a member of occupational group 13 and 15 are profitable if they are in age group 5 or if they are in age group 2 and have no claims discount.

However, CALT customers in other age groups with more than 1 insured are unprofitable if they have medical risk with no claims discount or if they have medical risk without no claims discount and having less than 3 insured.

The loyalty levels of LC and NC customers, who are previously customers, with more than 1 insured are unprofitable if they are in age groups 1 and 2. In other age groups, if the customer does not have a no claims discount, male age group 5 customers of product W are unprofitable by 61% or age group 3 customers of product W without a P/C insurance are unprofitable by 50% whereas the customers P/C insurance are profitable with a percentage of 73%. If the customer has a no claims discount however, the location matters. The cities [2 4 5 8 11 13 14 15 19 23 24 25 28 30 36 37 40 43 49 53 57 58 60 62 65 66 70 72 77 79 80] are unprofitable whereas the cities [10 17 18 21 32 46 47 50 51 52 63 64 78 81] are profitable. In city 1, interesting but medically risky customers are profitable while others are unprofitable. In city 6, age group 5 is profitable, but in age group 3 and 4 occupational group [5 7 13 14 15] are unprofitable while 11 is profitable. In city 34 product W is profitable while age group 3 and 5 CALT customers are unprofitable. In city 35, age group 5 is unprofitable unless the product is CALT.

When single insured NC and LC customers are analyzed, male customers having no previous medical risk are profitable they have no claims discount or CORTA and W customer without no claims discount. On female side, no claim discount seems to affect reversely, leading no profitable results. Female CALT customers without no claims discount in age group 5 is profitable.

IV.2.2 Strategies

The strategy development is a process greedy to information. Making decisions on what to do require understanding the knowledge of past and correctly predict the future. Description of what is going on and why provides the information to understand now and future events. Enough understanding of past, today and future leads the steps in the future.

Analysis of data is a sophisticated process and limited with analyzer ability, data and tools used. Customer data mining is one of the strongest tools to derive information from data which contains the application of descriptive and predictive analytics (such as clustering, segmentation, estimation, prediction and affinity analysis) to support the marketing, sales and service functions. The wide spectrum tool set of data mining provides advantage of wide range of analysis according to

the aim. The analysis may occur as an integrated part of an operational application or as a stand-alone analytical application. Generic data mining or statistical tools can be used to support marketing, sales and service applications. However, the growing interest in deploying this capability beyond the traditional analyzer team to organizations without an established background in data mining is attracting attention for packaged applications that provide more guidance and a process for users to perform customer analysis. These packages and defined guidance of process makes data mining analysis easier to use and widespread everyday. Better understanding supports better decisions.

A beginning analysis is carried out to describe profitable customers in this study. Product recommendation and personalized customer service are some advantages of customer profiling. However, the application range is much wider from cross-sales to time-series predictions, providing information to decision makers in different levels of customer cycle.

Basic analysis of internal customer, sales and operational data by data mining can be used to play a role even in brand development, new product development, product re-design, campaign management, customer specialized pricing etc. Moreover, combination of data mining with other marketing tools like surveys and focus group research can be used to create successful marketing plan and competitive advantage business strategy. Through these tools it is easier to target best customers and finding which advertising and promotion methods are most efficient and effective. For decision makers involved in creating a marketing strategy one of the first steps in opportunity scanning is data mining. For example, data mining analysis before a major external market research study to scan for market opportunities provides insights to increase the success.

Data mining can uncover emerging usage patterns within a newly defined market segment, high margin customers with increasing buying velocity, or an undetected trend in customer account losses. Data mining is often used to gauge the customer lifetime value in direct marketing efforts.

The results of the classification application with decision tree are explained in the previous section. The profitable customer profile gives a strong clue for customer acquisition and even for further stages of customer life cycle. The analysis

by machine learning systems, data mining, gives the chance of studying on many attributes and uncovering the relations to be used in decision making and strategy setting. Although the method used in the study is a relatively easy and understandable method, the analysis with too many attributes is almost impossible with only human effort. The data mining software tools can handle many attributes with many records in time of minutes. Since data mining is an iterative process, processing time is an important parameter. The profiles described by the analysis are the main inputs to marketing decisions.

When a model is selected strategies are developed by the use of information provided by the analysis. After interpreting the results of analysis, some examples of the outputs for developing new strategies of customer relations can be summarized as follows:

- Product S is a stand alone product which is purely profitable. With respect to this result, strategies can be generated to increase sales of S, new similar products can be designed, existing products can be enriched by the coverage concept of this product to increase profitability.
- New customers of CORTA, CUST and W products are generally profitable. New acquisition of customers in these products should be promoted. Moreover, CALT product should be analyzed to find out why it is not profitable while all other products are profitable.
- The products, loyalty, number of insureds, age group and city are main identifiers. The premium calculations are generally carried out on these except city, instead sales region is used as area identifier. In this study, it's seen that city is one of the dominant identifiers and differs from sales region.
- It is generally hard to say females are unprofitable while they buy the insurance first time although it is known as females are of higher risk. There is no clue about that. Moreover, they are not distinguishably different when the policy has more than insured. Possibly this effect is a consequence of males in the policy insureds. Because, when females are single insured in the policy, no rule could be found to be profitable. Single female insureds are not profitable. As a result,

strategies should be developed to include family members to her policy or pricing or coverage limits should be reconsidered.

- The customers purchased insurance before and insuring only him/herself with medical risk can not be fit into profitable profile in any case. The situation should be analyzed for different solutions like pricing policy changes, coverage arrangements or turning them to possibly profitable profiles by adding family members to the policy.
- The profitability classes does not have city attribute in new customers, for future profitability of renewed policies, the campaign strategy focus can be to the cities that policies are profitable in the renewal years.
- Direct marketing activities can be performed on the profitable profile potential customers.
- Special deductions to retain or acquire profitable customers can be performed.
- The definition of profitable customer profile is the customers to focus on to acquire. Target marketing strategies should be developed to attract potential customers fitting in the profitable profile.

V EVALUATION

The increased availability of value added information throughout the firm helps to augment the knowledge of the business to individuals. Decision makers are able to use information to better devise or implement the strategies to more effectively manage available sources in order to best meet the needs of customers. The growth in productivity due to informational revolution has been at highest rates since post World War-2 era and the effect is so huge that it is expressed in numbers of general economy as about half point decrease in inflation and one full point increase in growth in early 2000's. [58]

Successfully competing in the new global economy requires immediate decision capability. This immediate decision capability requires quick analysis of both timely and relevant data to derive accurate and useful conclusions. Organizations are now on the stage of getting use of data in hand. To support this analysis, organizations give more effort to data mining. Those organizations making swift, fact-based decisions by optimally leveraging their data resources will outperform those organizations that do not in the near future.

Data mining is a set of techniques to bring information from large data sets to the surface. Bernstein says if you torture the data enough they will say what you want. [59] Many statistical and AI techniques are used in data mining. In addition, data mining has a process approach making DM a systematic solution. In this study, it is intended to use data mining methods to derive conclusions from a large set of real insurance company data to be used strategy setting and decision making process in highly competitive market.

In a data mining study, practice is beyond the theory. In the real world, nothing is independent, many causes affect many results and even these results affect other results. The hardest part of a real study is to be able to imagine the cause and effect relations as much as possible. This needs great understanding of

the analyzed business and its environment including economics, social changes etc. from regional habitats to worldwide.

Another important issue in a real practice is nothing is ready and waiting for a data miner to analyze. Some companies have structured databases some don't have, some have data marts to work on some don't have even a database, some have many attributes in databases but have no data in them. Data mining is an "iceberg" having so much work to do before the analysis stage. It is very time consuming to reach analysis stage and data miner can be left in his hands nothing meaningful when it is achieved. In the life-cycle of a data mining project, in the Business Understanding stage, it is crucial to find out what we need and what we have. This is the connection of a data miner with life and further analysis. This is a continuous process and needed data that is agreed on with the business side should be collected by business departments.

This is a beginning study to have a general knowledge from existing data. However, working with a stable (can not be improved and changed) and unstructured data forces us to change the way of analysis and the result is "what you have is what you get". This is one of the main problems encountered during study.

The process approach is modeled in the study in a small subset of environment. The general approach to the problems and related business objectives and filtering them to the specific data mining targets is the main structure of the study. However, due to the limited time and sources at each step the target is filtered to a specific subset while leaving other probable targets to other analysis. This structured a concrete base for future studies on the data and complete solutions to all problems mentioned.

The model generated on primary success parameter which is profitability. The initial focusing point is selected as acquisition of customers. The model is formed to classify the profitable and unprofitable customers to determine profitable and profitable customer profiles. The model is directly related to focused marketing strategy development which meets the business objectives.

The unbalanced data problem which also caused problem by leading less accurate results in our study is pervasive and causing trouble to a large segment of the data mining community. Moreover, the nature of the insurance dealing with risk

increases the variance in predefined classes, which is difficult to interpret. These problems are tried to be reduced by balancing, reclassification and misclassification costs to some extend. Moreover, it was emphasized that the use of common evaluation measures such as accuracy can yield misleading conclusions in unbalanced data sets. [53] A different accuracy measure of False classifications (Class1 members clustered to Class 3 or vice versa) per True classifications is used to evaluate the results. This is also a need for the magnitude difference between true values and false values. Hence, evaluation of models is carried out on this rate and correctly defined records.

The problem of unbalanced data is tried to be solved by different methods. Balanced re-sampling by factors, application of clustering in re-sampling before classification, combination and simplification of attributes are applied. None of the methods yield a satisfactory result for the rare class, unprofitable customers. However, acceptable results are achieved in profiling profitable customers.

The usage of real data which has deficiencies with the aim of less error in defining profitable customers lead less accurate results. However, low level of error is achieved. As generally decision trees also the C5.0 algorithm is a greedy algorithm leading rules with high complexity in especially unbalanced data conditions. This situation affected the results of model 1 especially, and even in the simpler model-3 algorithm generated 14 branches. However, simpler and less number of attributes in model 3 resulted in more understandability as expected.

The classification work needs to be performed periodically since small changes in data may result different results. While doing this, further balancing methods can be used to reduce imbalance in data to define unprofitable customers better and in addition, to define unprofitable customers. Furthermore, a less greedy algorithm may lead better results in imbalanced environment.

The study contains only a first step to data mining world. As a further study, many analyses exampled in “II.8 Insurance and Data Mining” like association rule discovery for cross-sales, predictions for future by predictive analysis, fraud detection in extremeties can be performed.

VI REFERENCES

[1] Konrad R.: "Data mining: Digging user info for gold", http://zdnet.com.com/2100-11-528032.html?legacy=zdnn, *ZDNET News* February 7, **(2001)**

[2] The Technology Review Ten, *MIT Technology Review* January/February, **(2001)**

[3] Moxon B.: "Defining Data Mining", *DBMS Data Warehouse Supplement*, August, **(1996)**

[4] Seifert J.W.: "Data Mining: An Overview", *CRS Report for Congress* December 16, **(2004)**

[5] Shaw M.J.; Subramaniam C.; Tan G.W.; Welge M.E.: "Knowledge Management and Data Mining for Marketing", **(2001)**

[6] Mannila, H.: "Methods and Problems in Data Mining", *Proceedings of the 6th International Conference on Database Theory*, **(1997)**

[7] Thearling K.: "More Than Algorithms: Data Mining in the Real World", http://www.thearling.com/umass/index_frame.htm (Access date: December **2005**)

[8] Hand D.J.: "Data Mining: Statistics and More?", *The American Statistician* Vol.52 No:2, May, **(1998)**, 117-118

[9] Thearling K.: "An Introduction to Data Mining", http://www.thearling.com/text/dmwhite/dmwhite.htm (Access date: January **2006**)

[10] Berry M.; Linoff G.: "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management", Second Edition, *John Wiley & Sons*, **(2004)**

[11] Two Crows Corporation: "Introduction to Data Mining and Knowledge Discovery", Third Edition, **(1999)**

[12] Sharp D.E.: "Customer Relationship Management Systems Handbook", *CRC Press*, **(2003)**, 114-115

[13] Bukhbinder G.; Krumenaker M.; Phillips A.: "Insurance Industry Decision Support: Data Marts, OLAP and Predictive Analytics", **(2005)**

[14] Galfond G.: "Data Mining Can Unearth A Competitive Edge", October 6, **(1997)**

[15] http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm, **(2004)**

[16] Wirth R.; Hipp J.: "CRISP-DM: Towards a Standard Process Model for Data Mining", *Proceedings of the Fourth International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pp. 29-39, Manchester, UK, **(2000)**

[17] Chapman P.; Clinton J.; Kerber R.; Khabaza T.; Reinart T.; Shearer C.; Wirth R.: "CRISP–DM 1.0 Step-by-Step Data Mining Guide", http://www.crisp-dm.org, **(2000)**

[18] Edelstein H.A.: "Pan For Gold In The Clickstream", *InformationWeek* March 12, http://www.informationweek.com/828/prmining.htm, **(2001)**

[19] Larose D.T.: "Discovering Knowledge in Data: An Introduction to Data Mining", *John Wiley & Sons*, **(2005)**

[20] Hand D.; Mannila H.; Smyth P.: "Principles of Data Mining", *The MIT Press*, **(2001)**, 352-353, 384-391

[21] Guo L.: "Applying Data Mining Techniques in Property-Casualty Insurance", *Forums of the Casualty Actuarial Society*, **(2003)**

[22] Eker H.: "Veri Madenciliği veya Bilgi Keşfi", **(2005)**, 3

[23] Edelstein H.: "Building Profitable Customer Relationships With Data Mining", *SPSS White Paper*, **(2000)**

[24] Ashdown S.P.: "An Investigation of the Structure of Sizing Systems: A Comparison of Three Multidimensional Optimized Sizing Systems Generated from Anthropometric Data", *International Journal of Clothing Science and Technology*, Vol. 10 No:5, **(1998)**, 324-341

[25] Kaski S.: "Data exploration using self-organizing maps", Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No:82, Espoo, *Finnish Academy of Technology*, **(1997)**, 57

[26] Dunham M.H.: "Data Mining Introductory and Advanced Topics", *Prentice Hall*, http://engr.smu.edu/~mhd/ , **(2002),** Appendix 1

[27] Agrawal R.; Imielinski T.; Swami A.N.: "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.*, **(1993)**, 26-28, 207–216

[28] Collier K.; Carey B.; Sautter D.; Marjaniemi C.: "A Methodology for Evaluating and Selecting Data Mining Software", *Proceedings of the 32nd*

Hawaii International Conference on System Sciences, IEEE Computer Society, **(1999)**

[29] Thomasson M.: "Health Insurance in the United States", *EH.Net Encyclopedia*, http://eh.net/encyclopedia/article/thomasson.insurance.health.us , April 18, **(2003)**

[30] Alexander D.; Hilary N.; Shah S.: "Private Medical Insurance", *Faculty and Institute of Actuaries*, **(2001)**

[31] Swiss Re: "Insurance in Emerging Markets", *Sigma No:5*, **(2005)**

[32] http://www.bankaciyiz.biz/article373.html and http://www.ntvmsnbc.com/news/380959.asp , **(2006)**

[33] WHO Regional Office for Europe: "Health care systems in transition Turkey", **(1996)**

[34] Smith L.; Tossani L.S.: "Advanced Modeling, Visualization, and Data Mining Techniques for a New Risk Landscpape", *Casualty Actuarial Society Risk and Capital Management Seminar*, July 28-29, **(2003)**

[35] Insurance Information Institute: "Insurance Fraud", http://www.iii.org/media/hottopics/insurance/fraud/ , **(2006)**

[36] Hazine Sigorta Denetleme Kurulu: "2005 Yılında Türkiye'de Sigorta ve Bireysel Emeklilik Sektörü", **(2006)**, 6

[37] West D.: "Enhancing Value Through Data Mining", www.insurancenetworking.com/protected/article.cfm?articleId=3631 , **(2005)**

[38] SAS White Paper: "Data Mining in the Insurance Industry", **(2002)**

[39] Elkan C.: "Magical Thinking in Data Mining Lessons from CoIL Challenge 2000", http://www-cse.ucsd.edu/users/elkan/kddcoil.pdf , **(2001)**

[40] Viveros M.S.; Nearhos J.P.; Rothman M.J.: "Applying Data Mining Techniques to a Health Insurance Information System", *Proceedings of the 22nd VLDB Conference Mumbai (Bombay), India*, **(1996)**

[41] Kietz J.U.; Reimer U.; Staudt M.: "Mining Insurance Data at Swiss Life", *Proceedings of the 23rd VLDB Conference Athens, Greece*, **(1997)**

[42] Rossini I.: "Data Mining Applications in the Italian Insurance Market", *Lecture notes Politecnico di Milano*, **(2004)**

[43] Bologna Data Mining Center: "Database Marketing: A Data Mining Application For An Insurance Company", June, **(1998)**

[44] Chen M.S.; Han J.; Yu P.S.: "Data Mining: An Overview from Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, December, **(1996)**

[45] http://www.kdnuggets.com/polls/2006/data_mining_analytic_tools.htm , **(2006)**

[46] http://www.kdnuggets.com/polls/2005/data_mining_tools.htm , **(2005)**

[47] Cerny P.A.: "Data mining and Neural Networks from a Commercial Perspective", *Proceedings of the ORSNZ Conference Twenty Naught One*, University of Canterbury, NZ, **(2001)**

[48] Rulequest Research: http://www.rulequest.com/see5-info.html , May, **(2006)**

[49] Rulequest Research: http://www.rulequest.com/see5-comparison.html , May, **(2006)**

[50] Dyer C.R.: http://www.cs.wisc.edu/~dyer/cs540/notes/learning.html , *University of Wisconsin Madison*, **(2003)**

[51] http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_Trees.html , *Minnesota State University*, **(2002)**

[52] Brain D.; Webb G.I.: "On the Effect of Data Set Size on Bias and Variance in Classification Learning", *Proceedings of the Fourth Australian Knowledge Acquisition Workshop (AKAW '99)*, **(1999)**, 117-128

[53] Chawla N.V.; Japkowicz N.; Kolcz A.: "Special Issue on Learning from Imbalanced Data Sets", *SIGKDD Explorations*, Volume 6, Issue 1, **(2004)**, 1-6

[54] Han H.; Wang W.Y.; Mao B.H.: "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", *ICIC 2005*, Part I, LNCS 3644, **(2005)**, 878 – 887

[55] Visa S.; Ralescu A.: "Issues in Mining Imbalanced Data Sets - A Review Paper", *University of Cincinnati*, **(2005)**

[56] Weiss G.M.: "Mining with Rarity: A Unifying Framework", *SIGKDD Explorations, Volume 6, Issue 1*, **(2004)**, 7-19

[57] Nickerson A.S.; Japkowicz N.; Milios E.: "Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets", *Proceedings of the Eighth International Workshop on AI and Statistics*, **(2001)**

[58] Kudyba S.; Hoptroff R.: "Data Mining and Business Intelligence: A Guide to Productivity", *Idea Group Publishing*, **(2001)**

[59] Bernstein P.L.: "Against the Gods: the Remarkable Story of Risk", **(1998)**

VII APPENDICES

APPENDIX A

The list of data mining software including company name, web address, methods and techniques.

Software	Company	Web Address	Functions	Techniques
4Thought	Cognos Inc.	www.cognos.com/products/4thought	Prediction	Neural networks (MLP)
AC^2	ISoft	www.alice-soft.com/html/prod ac2.htm	Clustering, classification, prediction, segmentation	Decision trees
Affinium Model	Unica Corporation	www.unicacorp.com	Classification, segmentation	
AI Trilogy	Ward Systems Group, Inc.	www.wardsystems.com	Classification, forecasting, prediction	Genetic algorithms, jackknife estimation, neural networks
Alice d'ISoft & Alice Server	ISoft	www.alice-soft.com/html/prod alice.htm	Clustering, correlation, segmentation	Decision trees
AnswerTree 3.0	SPSS Inc.	www.spss.com/answertree	Classification	Decision trees [CHAID, Exhaustive CHAID, C&RT (a variation of CART), QUEST]
Braincel	Jurik Research and Consulting	www.jurikres.com/catalog/ms bcel.htm	Forecasting	Neural networks (back-prepogation)
BrainMaker	California Scientific Software	www.calsci.com	Forecasting	Neural networks

Software	Company	Web Address	Functions	Techniques
CART	Salford Systems	www.salford-systems.com	Classification	Decision tree (CART)
Clementine	SPSS Inc. (formerly Integral Solutions, Ltd.)	www.spss.com/clementine	Association rules, classification, clustering, factor analysis, forecasting, prediction, sequence discovery	Apriori, BIRCH, CARMA, Decision trees (C5.0, C&RT a variation of CART), K-means clustering, neural networks (Kohonen, MLP, RBFN), regression (linear, logistic) rule induction (C5.0, GRI)
Cubist 1.10	RuleQuest Research Pty Ltd	www.rulequest.com/cubist-info.html	Numerical modeling	Regression, rules
Darwin	Oracle Corporation	www.oracle.com/ip/analyze/warehouse/datamining/content.html	Clustering, prediction, classification, association rules	Decision trees (CART), K-means, K-nearest neighbors, neural networks, regression (linear, logistic)
Data Mining Suite	Information Discovery, Inc.	www.datamining.com/dmsuite.htm	Association rules, forecasting, prediction, rules	Rule induction
DataEngine	Management Intelligenter Technologien GmbH	www.dataengine.de	Classification, clustering, decision trees, time series analysis	Decision trees, fuzzy rules, K-means, neural networks (MLP, Kohonen), regression (linear)
DataMite	Logic Programming Associates Ltd.	www.lpa.co.uk/ind pro.htm	Rules	Decision trees, rule induction
DBMiner Analytical System	DBMiner Technologies Inc.	www.dbminer.com	Association rules, classification, clustering	Decision trees, K-means
Decider and Decider-Online	Neural Technologies	www.neuralt.com	Classification, prediction, rules	Neural networks
DecisionTime and WhatIf	SPSS Inc.	www.spss.com/decisiontime	Time series forecasting and analysis	ARIMA, exponential smoothing
Enterprise Miner	SAS Institute Inc.	www.sas.com/products/miner	Association rules, classification, clustering, prediction, time series	Decision trees (CART, CHAID), K nearest neighbors, regression (linear, logistic), memory-based reasoning, neural networks (Kohonen, MLP, RBF, SOM)

Software	Company	Web Address	Functions	Techniques
GainSmarts	Urban Science	www.urbanscience.com/main/gainpage.htm	Classification, clustering, prediction	Bayesian classification, decision trees (AID, CHAID), genetic algorithms, K-means, neural networks, regression (linear, logistic)
HNC Risk Suite	HNC Software	www.hnc.com	Prediction	Neural networks, rules
Intelligent Miner	IBM Corporation	www.software.ibm.com/data/iminer	Association rules, clustering, classification, prediction, sequential patterns, time series	Decision trees (modified CART), K-means, neural networks (MLP, back-propagation, RBF), regression (linear)
JDA Intellect (formerly Decision Series)	JDA Software Group, Inc.	www.jda.com/jdad/ProductsServices/Intellect facts.htm	Association rules, classification, clustering, prediction	Naive Bayes, decision trees, K-means, K nearest neighbors, neural networks (back-propagation, RBF)
JDBCMine	Intelligent Systems Research	www.intsysr.com/JDBCMine.htm	Classification	Decision trees (C4.5)
Kaidara Advisor (formerly KATE)	Kaidara	www.kaidara.com	Classification, clustering, prediction	Case-based reasoning, decision trees
KnowledgeSEEKER	ANGOSS Software Corporation	www.angoss.com/ProdServ/AnalyticalTools/index.html	Classification	Decision trees (CHAID, XAID, entropy-based algorithms)
KnowledgeSTUDIO	ANGOSS Software Corporation	www.angoss.com/ProdServ/AnalyticalTools/index.html	Classification, clustering, prediction, rules	Decision trees (CHAID), expectation-maximization, K-means, neural networks (MLP, RBF), regression (linear, logistic)
LOGIT	Salford Systems	www.salford-systems.com	Forecasting, hypothesis testing	Regression (logistic)
Magnify	Magnify	www.magnify.com	Prediction	
Magnum Opus 1.3	RuleQuest Research Pty Ltd	www.rulequest.com/MagnumOpus-info.html	Association Rules	Opus

Software	Company	Web Address	Functions	Techniques
Mantas	Mantas, Inc.	www.mantas.com	Association rules, classification, clustering, link analysis, prediction, sequence analysis, time series	Decision trees, neural networks
MarketMiner	MarketMiner Inc.	www.marketminer.com	Classification, pattern recognition, prediction, segmentation	Decision trees (C4.5), KNN, regression (linear, logistic), statistical
MARS	Salford Systems	www.salford-systems.com	Forecasting	Regression
Minotaur and Minotaur Transcure	Neural Technologies	www.neuralt.com	Classification, prediction, rules	Neural networks
Net Perceptions Retail Analyst (formerly KD1)	Net Perceptions, Inc.	www.netperceptions.com	Association rules, clustering	
Oracle9i Database	Oracle Corporation	www.oracle.com/ip/deploy/database/oracle9i/bi dm.html	Association rules, classification, prediction	Naive Bayes
Partek	Partek Incorporated	www.partek.com	Clustering, prediction	Genetic algorithms, neural networks, regression
PolyAnalyst 4.4	Megaputer Intelligence Inc.	www.megaputer.com/products/pa/index.php3	Association rules, classification, clustering, prediction	Decision trees, neural networks
Quadstone System	Quadstone	www.quadstone.com	Prediction, profiling, segmentation	Decision trees, regression
Re	Lumio Limited	www.lumio.com/products.reorder	Sequential patterns	
Scenario	Cognos Inc.	www.cognos.com/products/scenario	Classification, clustering, outlier detection	Decision trees
See5/C5.0 1.15	RuleQuest Research Pty Ltd	www.rulequest.com/see5-info.html	Classification	Decision trees, rules

Software	Company	Web Address	Functions	Techniques
S-Plus	Insightful Corporation	www.insightful.com	Classification, clustering, hypothesis testing, prediction, time series analysis	ARIMA, correlation (Pearson), decision trees, hierarchical clustering (agglomerative, divisive), K-means, regression (linear, logistic, nonlinear, polynomial), statistical techniques (Jackknife, Monte Carlo)
STATISTICA Data Miner	StatSoft, Inc.	www.statsoft.com/dataminer.html	Classification, clustering, prediction	ARIMA, decision trees (CART, CHAID), exponential smoothing, neural networks (Back-propagation, MLP, RBF, SOM), regression
SuperQuery	AZMY Thinkware, Inc.	www.azmy.com	Rules	Rule induciton
SurfAid Analytics	IBM Corporation	www.ibm.com/surfaid	Web mining	
Visualizer Workstation	Computer Science Innovations, Inc.	www.csi-inc.com	Classification, clustering, correlations, prediction, rules	Neural networks(RBF)
WebAnalyst	Megaputer Intelligence Inc.	www.megaputer.com/products/wa/index.php3	Web mining (prediction, patterns)	
WebMiner ASP	Webminer	www.webminer.com	Prediction, web mining	
WizWhy	WizSoft	www.wizsoft.com/Why.html	Prediction, rules (if-then and if-and-only-if)	Rule induction
XpertRule Miner	Attar Software Ltd.	www.attar.com	Association rules, classification, clustering	Decision trees, rules

APPENDIX B

The decision tree of model 1.

```
ProductL2 in [ "CALT" "CORTA" "CUST" "W" ] [ Mode: 2 ] (69.681,656)
      Noins <= 1 [ Mode: 2 ] (40.296,831)
            Addpr = 1 [ Mode: 2 ] (5.348,981)
                  NCD = 1 [ Mode: 2 ] (713)
                        VIP = 1 [ Mode: 2 ] => 2 (22, 0,591)
                        VIP = 0 [ Mode: 2 ] (691)
                              Sex in [ "E" "K" ] [ Mode: 2 ] => 2 (691, 0,339)
                              Sex in [ "" "C" ] [ Mode: 1 ] => 1 (0)
                  NCD = 0 [ Mode: 2 ] (4.635,981)
                        ProductL2 in [ "CALT" "CORTA" ] [ Mode: 2 ] (1.414,998)
                              Paytool in [ "H" "K" "O" ] [ Mode: 2 ] (1.394,998)
                                    AgeGr in [ 2 3 ] [ Mode: 2 ] => 2 (856,998, 0,396)
                                    AgeGr in [ 4 5 ] [ Mode: 2 ] (538)
                                          ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (62, 0,613)
                                          ProductL2 in [ "CORTA" ] [ Mode: 2 ] (476)
                                                Region in [ 11 12 13 14 15 16 ] [ Mode: 2 ] => 2 (332, 0,386)
                                                Region in [ 17 ] [ Mode: 2 ] (89)
                                                      Sex in [ "E" ] [ Mode: 2 ] => 2 (43, 0,372)
                                                      Sex in [ "K" ] [ Mode: 1 ] => 1 (45, 0,533)
                                                      Sex in [ "" "C" ] [ Mode: 3 ] => 3 (1, 1,0)
                                                Region in [ 18 20 ] [ Mode: 1 ] => 1 (55, 0,527)
                                    AgeGr in [ 1 ] [ Mode: 3 ] => 3 (0)
                              Paytool in [ "N" ] [ Mode: 3 ] => 3 (20, 0,6)
                        ProductL2 in [ "CUST" "W" ] [ Mode: 2 ] => 2 (3.220,983, 0,396)
            Addpr = 0 [ Mode: 2 ] (34.947,85)
                  ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] (26.663,907)
                        Sex in [ "" ] [ Mode: 2 ] (19)
                              Discount = 1 [ Mode: 2 ] => 2 (13, 0,615)
                              Discount = 0 [ Mode: 2 ] => 2 (6, 0,333)
                        Sex in [ "E" ] [ Mode: 2 ] (14.680,907)
                              OccupGr in [ 0 1 2 3 4 5 6 7 8 10 11 12 13 14 ] [ Mode: 2 ] (14.369,907)
```

```
                                                NCD = 1 [ Mode: 2 ] (6.977,91)
                                                      VIP = 1 [ Mode: 2 ] => 2 (137, 0,321)
                                                      VIP = 0 [ Mode: 2 ] (6.840,91)
                                                            ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (6.840,91,
0,274)

                                                NCD = 0 [ Mode: 2 ] (7.391,997)
                                                      City in [ 1 ] [ Mode: 2 ] (83)
                                                            Region in [ 11 ] [ Mode: 1 ] => 1 (11, 0,727)
                                                            Region in [ 12 18 ] [ Mode: 2 ] => 2 (71, 0,324)
                                                            Region in [ 13 14 15 16 17 20 ] [ Mode: 3 ] => 3 (1, 1,0)
                                                      City in [ 3 9 14 17 21 22 23 25 28 33 38 42 44 45 47 52 54 55 63 68
74 ] [ Mode: 2 ] => 2 (225, 0,351)

                                                      City in [ 5 19 30 39 53 64 71 72 78 80 81 ] [ Mode: 1 ] => 1 (32,
0,75)

                                                      City in [ 6 ] [ Mode: 2 ] (1.013,999)
                                                            Region in [ 11 14 ] [ Mode: 2 ] => 2 (23, 0,565)
                                                            Region in [ 12 16 20 ] [ Mode: 3 ] => 3 (2, 1,0)
                                                            Region in [ 13 ] [ Mode: 2 ] (956,999)
                                                                  AgeGr in [ 1 5 ] [ Mode: 1 ] => 1 (13, 0,538)
                                                                  AgeGr in [ 2 3 ] [ Mode: 2 ] => 2 (774,999, 0,347)
                                                                  AgeGr in [ 4 ] [ Mode: 2 ] (169)
                                                                        OccupGr in [ 0 ] [ Mode: 2 ] (86)
                                                                              Channel in [ "A" ] [ Mode: 2 ] => 2 (38,
0,632)

                                                                              Channel in [ "B" ] [ Mode: 1 ] => 1 (18,
0,611)

                                                                              Channel in [ "D" ] [ Mode: 2 ] (30)
                                                                                    ProductL2 in [ "CALT" ] [ Mode: 1 ] =>
1 (11, 0,636)

                                                                                    ProductL2 in [ "CUST" ] [ Mode: 2 ]
(19)

                                                                                          Paytool in [ "H" ] [ Mode: 1 ]
=> 1 (4, 0,5)

                                                                                          Paytool in [ "K" ] [ Mode: 2 ]
(15)
```

```
                                                    Discount = 1 [ Mode: 2 ]

                                                    Discount = 0 [ Mode: 2 ]

                                                Paytool in [ "N" "O" ] [ Mode: 3

                                    OccupGr in [ 1 ] [ Mode: 2 ] (49)
                                          ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ]

                                    OccupGr in [ 2 4 5 6 7 11 14 ] [ Mode: 1 ] => 1

                                    OccupGr in [ 3 8 12 13 ] [ Mode: 2 ] => 2 (25,

                                    OccupGr in [ 10 ] [ Mode: 3 ] => 3 (1, 1,0)
                         Region in [ 15 18 ] [ Mode: 1 ] => 1 (8, 0,5)
                         Region in [ 17 ] [ Mode: 2 ] (24)
                               AgeGr in [ 1 2 5 ] [ Mode: 1 ] => 1 (15, 0,533)
                               AgeGr in [ 3 4 ] [ Mode: 2 ] => 2 (9, 0,778)
City in [ 7 ] [ Mode: 2 ] (319)
      Payment = T [ Mode: 2 ] => 2 (292, 0,312)
      Payment = P [ Mode: 2 ] (27)
            Paytool in [ "H" "K" "O" ] [ Mode: 2 ] => 2 (27, 0,481)
            Paytool in [ "N" ] [ Mode: 1 ] => 1 (0)
City in [ 10 ] [ Mode: 2 ] (29)
      ProductL2 in [ "CALT" ] [ Mode: 1 ] => 1 (4, 0,75)
      ProductL2 in [ "CUST" ] [ Mode: 2 ] (25)
            AgeGr in [ 2 3 ] [ Mode: 2 ] => 2 (23, 0,435)
            AgeGr in [ 4 ] [ Mode: 1 ] => 1 (2, 1,0)
            AgeGr in [ 1 5 ] [ Mode: 3 ] => 3 (0)
City in [ 16 ] [ Mode: 2 ] (204)
      Channel in [ "A" ] [ Mode: 2 ] => 2 (69, 0,42)
      Channel in [ "B" "D" ] [ Mode: 2 ] (135)
            Payment = T [ Mode: 2 ] => 2 (122, 0,393)
            Payment = P [ Mode: 2 ] => 2 (13, 0,462)
```

```
=> 2 (9, 0,333)

=> 2 (6, 0,833)

] => 3 (0)


=> 2 (49, 0,347)

(8, 0,625)

0,44)
```

```
                                                City in [ 2 4 8 11 12 13 15 18 24 29 36 37 40 43 46 49 50 51 56 57
58 59 60 61 62 65 66 69 70 73 76 79 ] [ Mode: 3 ] => 3 (57, 0,579)
                                                City in [ 20 ] [ Mode: 2 ] (72)
                                                      OccupGr in [ 0 3 5 11 ] [ Mode: 2 ] => 2 (33, 0,303)
                                                      OccupGr in [ 1 ] [ Mode: 2 ] (35)
                                                            AgeGr in [ 2 ] [ Mode: 2 ] => 2 (18, 0,222)
                                                            AgeGr in [ 3 ] [ Mode: 2 ] (16)
                                                                  Channel in [ "A" ] [ Mode: 2 ] => 2 (6, 0,5)
                                                                  Channel in [ "B" "D" ] [ Mode: 1 ] => 1 (10, 0,7)
                                                            AgeGr in [ 1 4 5 ] [ Mode: 3 ] => 3 (1, 1,0)
                                                      OccupGr in [ 2 4 6 7 8 10 12 13 ] [ Mode: 1 ] => 1 (2, 1,0)
                                                      OccupGr in [ 14 ] [ Mode: 3 ] => 3 (2, 1,0)
                                                City in [ 26 ] [ Mode: 2 ] (17)
                                                      AgeGr in [ 2 ] [ Mode: 3 ] => 3 (6, 0,667)
                                                      AgeGr in [ 1 3 4 5 ] [ Mode: 1 ] => 1 (11, 0,636)
                                                City in [ 27 ] [ Mode: 2 ] (22)
                                                      OccupGr in [ 0 2 4 5 6 8 10 11 12 13 14 ] [ Mode: 3 ] => 3
(14, 0,571)
                                                      OccupGr in [ 1 3 7 ] [ Mode: 1 ] => 1 (8, 0,625)
                                                City in [ 31 ] [ Mode: 2 ] (9)
                                                      Discount = 1 [ Mode: 3 ] => 3 (4, 0,75)
                                                      Discount = 0 [ Mode: 2 ] => 2 (5, 0,6)
                                                City in [ 32 ] [ Mode: 3 ] (9)
                                                      Discount = 1 [ Mode: 1 ] => 1 (5, 0,6)
                                                      Discount = 0 [ Mode: 3 ] => 3 (4, 1,0)
                                                City in [ 34 ] [ Mode: 2 ] (4.232,998)
                                                      VIP = 1 [ Mode: 2 ] => 2 (94, 0,223)
                                                      VIP = 0 [ Mode: 2 ] (4.138,998)
                                                            Channel in [ "A" "B" ] [ Mode: 2 ] (3.022)
                                                                  OccupGr in [ 0 ] [ Mode: 2 ] (944)
                                                                        PCI = 1 [ Mode: 2 ] (158)
                                                                              Paytool in [ "H" "K" "O" ] [ Mode: 2 ]
=> 2 (158, 0,373)
                                                                              Paytool in [ "N" ] [ Mode: 3 ] => 3
(0)
```

```
=> 2 (440, 0,33)

(4, 1,0)


=> 2 (48, 0,417)

(39)

"CALT" ] [ Mode: 3 ] => 3 (7, 0,571)

"CUST" ] [ Mode: 2 ] => 2 (32, 0,5)

(34)

] [ Mode: 2 ] => 2 (34, 0,529)

] => 1 (1, 1,0)


2 ] (138)

"CALT" "CUST" ] [ Mode: 2 ] => 2 (138, 0,254)

2 ] => 2 (23, 0,391)


(51)

=> 2 (24, 0,292)

=> 2 (27, 0,593)
```

```
PCI = 0 [ Mode: 2 ] (786)
      Region in [ 11 13 15 18 ] [ Mode: 2 ]

      Region in [ 12 14 ] [ Mode: 1 ] => 1

      Region in [ 17 ] [ Mode: 2 ] (283)
            Discount = 1 [ Mode: 2 ] (122)
                  AgeGr in [ 2 ] [ Mode: 2 ]

                  AgeGr in [ 3 ] [ Mode: 2 ]

                        ProductL2 in [

                        ProductL2 in [

                  AgeGr in [ 4 ] [ Mode: 2 ]

                        Channel in [ "A" "B"

                  AgeGr in [ 1 5 ] [ Mode: 1

            Discount = 0 [ Mode: 2 ] (161)
                  Channel in [ "A" ] [ Mode:

                        ProductL2 in [

                  Channel in [ "B" ] [ Mode:

      Region in [ 20 ] [ Mode: 2 ] (59)
            Channel in [ "A" ] [ Mode: 2 ]

                  Discount = 1 [ Mode: 2 ]

                  Discount = 0 [ Mode: 2 ]
```

```
=> 1 (8, 0,5)




3 ] => 3 (7, 0,714)

1 ] => 1 (40, 0,5)



] => 2 (150, 0,24)

Mode: 1 ] => 1 (1, 1,0)

=> 3 (19, 0,579)


=> 2 (6, 0,333)

(108)

"CUST" ] [ Mode: 2 ] => 2 (108, 0,389)

=> 1 (2, 1,0)

=> 3 (0)


1 (12, 0,5)
```

```
                  Channel in [ "B" ] [ Mode: 1 ]

            Region in [ 16 ] [ Mode: 3 ] => 3 (0)
OccupGr in [ 1 ] [ Mode: 2 ] (896)
      AgeGr in [ 1 ] [ Mode: 2 ] => 2 (1, 1,0)
      AgeGr in [ 2 ] [ Mode: 2 ] (359)
            PCI = 1 [ Mode: 2 ] (47)
                  ProductL2 in [ "CALT" ] [ Mode:

                  ProductL2 in [ "CUST" ] [ Mode:

            PCI = 0 [ Mode: 2 ] => 2 (312, 0,292)
      AgeGr in [ 3 ] [ Mode: 2 ] (286)
            Channel in [ "A" ] [ Mode: 2 ] (170)
                  Region in [ 11 13 17 ] [ Mode: 2

                  Region in [ 12 14 15 16 ] [

                  Region in [ 18 20 ] [ Mode: 3 ]

            Channel in [ "B" ] [ Mode: 2 ] (116)
                  Paytool in [ "H" ] [ Mode: 2 ]

                  Paytool in [ "K" ] [ Mode: 2 ]

                        ProductL2 in [ "CALT"

                  Paytool in [ "O" ] [ Mode: 1 ]

                  Paytool in [ "N" ] [ Mode: 3 ]

      AgeGr in [ 4 ] [ Mode: 2 ] (211)
            ProductL2 in [ "CALT" ] [ Mode: 1 ] =>
```

```
(199)


2 ] => 2 (30, 0,533)

1 ] => 1 (10, 0,5)

0,34)



0,326)



(30)

=> 2 (16, 0,312)

=> 2 (14, 0,643)

=> 1 (7, 0,571)

2 (1, 1,0)

Mode: 3 ] => 3 (0)


0,277)
```

```
                ProductL2 in [ "CUST" ] [ Mode: 2 ]

                      PCI = 1 [ Mode: 2 ] (40)
                            Channel in [ "A" ] [ Mode:

                            Channel in [ "B" ] [ Mode:

                      PCI = 0 [ Mode: 2 ] => 2 (159,

      AgeGr in [ 5 ] [ Mode: 2 ] (39)
            PCI = 1 [ Mode: 3 ] => 3 (4, 0,75)
            PCI = 0 [ Mode: 2 ] => 2 (35, 0,314)
OccupGr in [ 2 4 5 7 14 ] [ Mode: 2 ] => 2 (576,

OccupGr in [ 3 ] [ Mode: 2 ] (291)
      PCI = 1 [ Mode: 2 ] (43)
            Payment = T [ Mode: 2 ] (38)
                  Region in [ 11 17 ] [ Mode: 2 ]

                        Discount = 1 [ Mode: 2 ]

                        Discount = 0 [ Mode: 2 ]

                  Region in [ 13 20 ] [ Mode: 1 ]

                  Region in [ 18 ] [ Mode: 2 ] =>

                  Region in [ 12 14 15 16 ] [

            Payment = P [ Mode: 3 ] => 3 (5, 0,6)
      PCI = 0 [ Mode: 2 ] (248)
            Payment = T [ Mode: 2 ] => 2 (195,

            Payment = P [ Mode: 2 ] (53)
```

```
(37)
=> 1 (17, 0,529)
(20)
] [ Mode: 2 ] => 2 (20, 0,45)
2 (12, 0,333)
1 (1, 1,0)
[ Mode: 3 ] => 3 (3, 1,0)

0,529)


(59)
"CUST" ] [ Mode: 2 ] => 2 (59, 0,339)
(11)
=> 2 (5, 0,8)
Mode: 1 ] => 1 (6, 1,0)
0,545)


0,279)
```

```
                  Region in [ 11 ] [ Mode: 2 ]
                        Discount = 1 [ Mode: 1 ]
                        Discount = 0 [ Mode: 2 ]
                              Channel in [ "A" "B"
                  Region in [ 17 ] [ Mode: 2 ] =>
                  Region in [ 18 ] [ Mode: 1 ] =>
                  Region in [ 12 13 14 15 16 20 ]
OccupGr in [ 6 ] [ Mode: 2 ] (99)
      Paytool in [ "H" "N" ] [ Mode: 1 ] => 1 (17,
      Paytool in [ "K" ] [ Mode: 2 ] (81)
            Payment = T [ Mode: 2 ] (70)
                  Channel in [ "A" ] [ Mode: 2 ]
                        ProductL2 in [ "CALT"
                  Channel in [ "B" ] [ Mode: 1 ]
                        AgeGr in [ 2 ] [ Mode: 2 ]
                        AgeGr in [ 1 3 4 5 ] [
            Payment = P [ Mode: 3 ] => 3 (11,
      Paytool in [ "O" ] [ Mode: 2 ] => 2 (1, 1,0)
OccupGr in [ 8 ] [ Mode: 2 ] (110)
      AgeGr in [ 2 4 ] [ Mode: 2 ] => 2 (61,
```

```
                                                AgeGr in [ 1 3 ] [ Mode: 3 ] => 3 (43,
0,558)
                                                AgeGr in [ 5 ] [ Mode: 1 ] => 1 (6, 0,667)
                                          OccupGr in [ 10 ] [ Mode: 2 ] (34)
                                                ProductL2 in [ "CALT" ] [ Mode: 3 ] => 3 (5,
0,6)
                                                ProductL2 in [ "CUST" ] [ Mode: 2 ] (29)
                                                      AgeGr in [ 1 2 5 ] [ Mode: 1 ] => 1
(15, 0,667)
                                                      AgeGr in [ 3 4 ] [ Mode: 2 ] => 2 (14,
0,357)
                                          OccupGr in [ 11 ] [ Mode: 2 ] (32)
                                                PCI = 1 [ Mode: 2 ] => 2 (6, 0,667)
                                                PCI = 0 [ Mode: 2 ] (26)
                                                      Discount = 1 [ Mode: 2 ] => 2 (17,
0,412)
                                                      Discount = 0 [ Mode: 1 ] => 1 (9,
0,556)
                                          OccupGr in [ 12 ] [ Mode: 2 ] (26)
                                                Region in [ 11 12 13 14 15 16 ] [ Mode: 3 ]
=> 3 (15, 0,533)
                                                Region in [ 17 20 ] [ Mode: 2 ] => 2 (10,
0,4)
                                                Region in [ 18 ] [ Mode: 1 ] => 1 (1, 1,0)
                                          OccupGr in [ 13 ] [ Mode: 2 ] (14)
                                                Paytool in [ "H" "N" ] [ Mode: 3 ] => 3 (4,
0,75)
                                                Paytool in [ "K" "O" ] [ Mode: 2 ] => 2 (10,
0,5)
                                      Channel in [ "D" ] [ Mode: 2 ] => 2 (1.116,998, 0,338)
                                City in [ 35 ] [ Mode: 2 ] (741)
                                      Paytool in [ "H" ] [ Mode: 2 ] (90)
                                      ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (90,
0,311)
                                      Paytool in [ "K" ] [ Mode: 2 ] (640)
```

```
                                    Channel in [ "A" "D" ] [ Mode: 2 ] (499)
                                          AgeGr in [ 2 ] [ Mode: 2 ] (217)
                                                Discount = 1 [ Mode: 2 ] => 2 (150, 0,293)
                                                Discount = 0 [ Mode: 2 ] => 2 (67, 0,313)
                                          AgeGr in [ 3 ] [ Mode: 2 ] (165)
                                                Channel in [ "A" "D" ] [ Mode: 2 ] => 2
(165, 0,37)

                                                AgeGr in [ 1 4 ] [ Mode: 2 ] => 2 (104, 0,337)
                                                AgeGr in [ 5 ] [ Mode: 2 ] (13)
                                                      Channel in [ "A" ] [ Mode: 3 ] => 3 (8,
0,625)

                                                      Channel in [ "D" ] [ Mode: 1 ] => 1 (5, 0,6)
                                          Channel in [ "B" ] [ Mode: 2 ] => 2 (141, 0,496)
                                    Paytool in [ "N" ] [ Mode: 1 ] => 1 (4, 0,5)
                                    Paytool in [ "O" ] [ Mode: 2 ] => 2 (7, 0,714)
                              City in [ 41 ] [ Mode: 2 ] (74)
                                    OccupGr in [ 0 2 3 7 8 10 11 12 13 14 ] [ Mode: 2 ] => 2 (44,
0,455)

                                    OccupGr in [ 1 ] [ Mode: 2 ] (25)
                                          ProductL2 in [ "CALT" ] [ Mode: 1 ] => 1 (6, 0,667)
                                          ProductL2 in [ "CUST" ] [ Mode: 2 ] => 2 (19, 0,579)
                                    OccupGr in [ 4 5 ] [ Mode: 1 ] => 1 (4, 0,75)
                                    OccupGr in [ 6 ] [ Mode: 3 ] => 3 (1, 1,0)
                              City in [ 48 ] [ Mode: 2 ] (212)
                                    Paytool in [ "H" "N" ] [ Mode: 1 ] => 1 (14, 0,714)
                                    Paytool in [ "K" ] [ Mode: 2 ] (195)
                                          OccupGr in [ 0 3 4 7 ] [ Mode: 2 ] => 2 (103, 0,34)
                                          OccupGr in [ 1 ] [ Mode: 2 ] (75)
                                                Payment = T [ Mode: 2 ] (65)
                                                      PCI = 1 [ Mode: 2 ] (12)
                                                            AgeGr in [ 2 ] [ Mode: 1 ] => 1 (3,
0,667)

                                                            AgeGr in [ 3 ] [ Mode: 2 ] => 2 (5,
0,6)
```

```
                                                                  AgeGr in [ 1 4 5 ] [ Mode: 3 ] => 3
(4, 0,75)
                                                                        PCI = 0 [ Mode: 2 ] => 2 (53, 0,377)
                                                               Payment = P [ Mode: 2 ] => 2 (10, 0,4)
                                                          OccupGr in [ 5 6 8 10 12 13 ] [ Mode: 1 ] => 1 (11,
0,727)
                                                                OccupGr in [ 2 11 14 ] [ Mode: 3 ] => 3 (6, 0,667)
                                                     Paytool in [ "O" ] [ Mode: 2 ] => 2 (3, 0,333)
                                                City in [ 67 ] [ Mode: 2 ] (26)
                                                     PCI = 1 [ Mode: 3 ] => 3 (4, 0,75)
                                                     PCI = 0 [ Mode: 2 ] => 2 (22, 0,318)
                                                City in [ 77 ] [ Mode: 2 ] (14)
                                                     PCI = 1 [ Mode: 2 ] => 2 (6, 0,667)
                                                     PCI = 0 [ Mode: 1 ] (8)
                                                          Discount = 1 [ Mode: 1 ] => 1 (4, 0,75)
                                                          Discount = 0 [ Mode: 3 ] => 3 (4, 0,75)
                                    OccupGr in [ 9 15 ] [ Mode: 1 ] => 1 (311, 0,537)
                               Sex in [ "K" ] [ Mode: 2 ] (11.964)
                                    City in [ 1 ] [ Mode: 2 ] (90,015)
                                         AgeGr in [ 2 ] [ Mode: 2 ] (41)
                                              OccupGr in [ 0 5 ] [ Mode: 3 ] => 3 (25, 0,52)
                                              OccupGr in [ 1 2 3 4 6 7 8 9 10 11 12 13 14 15 ] [ Mode: 2 ] => 2
(16, 0,812)
                                         AgeGr in [ 1 3 4 5 ] [ Mode: 2 ] => 2 (49,015, 0,367)
                                    City in [ 3 14 21 23 31 39 42 46 47 52 54 59 64 67 77 78 81 ] [ Mode: 2 ] => 2
(162,027, 0,389)
                                    City in [ 5 10 15 53 65 72 ] [ Mode: 1 ] => 1 (41,007, 0,585)
                                    City in [ 6 ] [ Mode: 2 ] (1.135,19)
                                         NCD = 1 [ Mode: 2 ] => 2 (417,095, 0,355)
                                         NCD = 0 [ Mode: 2 ] => 2 (718,095, 0,511)
                                    City in [ 7 ] [ Mode: 2 ] (305,052)
                                         PCI = 1 [ Mode: 2 ] => 2 (31, 0,581)
                                         PCI = 0 [ Mode: 2 ] => 2 (274,052, 0,347)
                                    City in [ 2 4 8 11 12 13 18 19 24 25 28 29 30 32 36 37 38 40 43 44 49 50 51 55
56 57 58 60 62 63 66 68 69 70 71 73 74 76 79 80 ] [ Mode: 3 ] => 3 (37,006, 0,649)
```

```
City in [ 9 ] [ Mode: 2 ] (58,01)
      Discount = 1 [ Mode: 2 ] => 2 (46,01, 0,304)
      Discount = 0 [ Mode: 1 ] (12)
            Channel in [ "A" "D" ] [ Mode: 1 ] => 1 (6, 1,0)
            Channel in [ "B" ] [ Mode: 2 ] => 2 (6, 0,167)
City in [ 16 ] [ Mode: 2 ] (251,042)
      Region in [ 11 ] [ Mode: 2 ] (20)
            Paytool in [ "H" "K" ] [ Mode: 2 ] => 2 (20, 0,3)
            Paytool in [ "N" "O" ] [ Mode: 3 ] => 3 (0)
      Region in [ 13 ] [ Mode: 3 ] => 3 (4, 0,75)
      Region in [ 14 ] [ Mode: 2 ] (216)
            Channel in [ "A" "B" ] [ Mode: 2 ] (131)
                  OccupGr in [ 0 ] [ Mode: 2 ] (33)
                        Discount = 1 [ Mode: 2 ] => 2 (24, 0,417)
                        Discount = 0 [ Mode: 2 ] => 2 (9, 0,667)
                  OccupGr in [ 1 ] [ Mode: 2 ] (20)
                        AgeGr in [ 1 2 4 5 ] [ Mode: 3 ] => 3 (14, 0,714)
                        AgeGr in [ 3 ] [ Mode: 2 ] => 2 (6, 1,0)
                  OccupGr in [ 2 4 5 6 7 8 9 10 11 13 15 ] [ Mode: 2 ] => 2
(65, 0,523)
                  OccupGr in [ 3 12 14 ] [ Mode: 1 ] => 1 (13, 0,692)
            Channel in [ "D" ] [ Mode: 2 ] => 2 (85, 0,329)
      Region in [ 15 20 ] [ Mode: 2 ] => 2 (5,042, 0,798)
      Region in [ 12 16 17 18 ] [ Mode: 1 ] => 1 (6, 0,5)
City in [ 17 ] [ Mode: 2 ] (31,005)
      Paytool in [ "H" ] [ Mode: 1 ] => 1 (4, 0,75)
      Paytool in [ "K" ] [ Mode: 2 ] (26,005)
            AgeGr in [ 2 ] [ Mode: 2 ] (12)
                  Channel in [ "A" "D" ] [ Mode: 1 ] => 1 (8, 0,5)
                  Channel in [ "B" ] [ Mode: 3 ] => 3 (4, 0,75)
            AgeGr in [ 1 3 4 ] [ Mode: 2 ] => 2 (13,005, 0,615)
            AgeGr in [ 5 ] [ Mode: 3 ] => 3 (1, 1,0)
      Paytool in [ "N" "O" ] [ Mode: 2 ] => 2 (1, 1,0)
City in [ 20 ] [ Mode: 2 ] (40,007)
      ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (40,007, 0,3)
```

```
City in [ 22 ] [ Mode: 2 ] (16,003)
      Channel in [ "A" ] [ Mode: 2 ] (10)
            NCD = 1 [ Mode: 1 ] => 1 (5, 0,6)
            NCD = 0 [ Mode: 3 ] => 3 (5, 0,6)
      Channel in [ "B" ] [ Mode: 1 ] => 1 (4, 0,75)
      Channel in [ "D" ] [ Mode: 2 ] => 2 (2,003, 0,5)
City in [ 26 ] [ Mode: 1 ] (15,003)
      NCD = 1 [ Mode: 1 ] => 1 (10,001, 0,7)
      NCD = 0 [ Mode: 3 ] => 3 (5,001, 0,6)
City in [ 27 ] [ Mode: 2 ] (18,003)
      OccupGr in [ 0 1 12 ] [ Mode: 2 ] => 2 (9,003, 0,778)
      OccupGr in [ 2 3 4 5 6 7 8 9 10 11 13 14 15 ] [ Mode: 3 ] => 3 (9, 0,778)
City in [ 33 ] [ Mode: 2 ] (82,014)
      AgeGr in [ 2 3 ] [ Mode: 2 ] => 2 (65, 0,385)
      AgeGr in [ 4 5 ] [ Mode: 1 ] => 1 (17,014, 0,647)
      AgeGr in [ 1 ] [ Mode: 3 ] => 3 (0)
City in [ 34 ] [ Mode: 2 ] (8.124,357)
      NCD = 1 [ Mode: 2 ] (3.412,678)
            AgeGr in [ 2 ] [ Mode: 2 ] => 2 (1.330,999, 0,349)
            AgeGr in [ 3 4 5 ] [ Mode: 2 ] (2.081,679)
                  ProductL2 in [ "CALT" ] [ Mode: 2 ] (54)
                        Payment = T [ Mode: 2 ] => 2 (50, 0,54)
                        Payment = P [ Mode: 1 ] => 1 (4, 0,5)
                  ProductL2 in [ "CUST" ] [ Mode: 2 ] (2.027,679)
                        OccupGr in [ 0 ] [ Mode: 2 ] (742,679)
                              VIP = 1 [ Mode: 2 ] (21)
                                    Payment = T [ Mode: 1 ] => 1 (15, 0,6)
                                    Payment = P [ Mode: 3 ] => 3 (6, 0,833)
                              VIP = 0 [ Mode: 2 ] => 2 (721,679, 0,364)
                        OccupGr in [ 1 ] [ Mode: 2 ] (284)
                              Paytool in [ "H" ] [ Mode: 2 ] (25)
                                    Channel in [ "A" ] [ Mode: 3 ] => 3 (16,
0,625)
                                    Channel in [ "B" ] [ Mode: 2 ] => 2 (1, 1,0)
                                    Channel in [ "D" ] [ Mode: 1 ] => 1 (8, 0,5)
```

```
0,249)

0,538)

(61, 0,246)

0,306)

0,833)

(13, 0,538)
```

```
      Paytool in [ "K" ] [ Mode: 2 ] (249)
            AgeGr in [ 3 4 ] [ Mode: 2 ] (223)
                  Payment = T [ Mode: 2 ] => 2 (197,

                  Payment = P [ Mode: 2 ] => 2 (26,

            AgeGr in [ 5 ] [ Mode: 2 ] => 2 (26, 0,385)
      Paytool in [ "N" "O" ] [ Mode: 2 ] => 2 (10, 0,5)
OccupGr in [ 2 ] [ Mode: 2 ] (409)
      Region in [ 11 ] [ Mode: 2 ] (216)
            PCI = 1 [ Mode: 2 ] => 2 (34, 0,471)
            PCI = 0 [ Mode: 2 ] => 2 (182, 0,341)
      Region in [ 15 ] [ Mode: 3 ] => 3 (4, 1,0)
      Region in [ 17 ] [ Mode: 2 ] (128)
            PCI = 1 [ Mode: 3 ] => 3 (8, 0,625)
            PCI = 0 [ Mode: 2 ] => 2 (120, 0,358)
      Region in [ 12 13 14 16 18 20 ] [ Mode: 2 ] => 2

OccupGr in [ 3 8 10 12 13 ] [ Mode: 2 ] => 2 (209,

OccupGr in [ 4 ] [ Mode: 2 ] (143)
      AgeGr in [ 3 5 ] [ Mode: 2 ] => 2 (85, 0,318)
      AgeGr in [ 4 ] [ Mode: 2 ] (58)
            Payment = T [ Mode: 2 ] => 2 (52, 0,519)
            Payment = P [ Mode: 1 ] => 1 (6, 0,5)
OccupGr in [ 5 ] [ Mode: 2 ] (73)
      AgeGr in [ 3 ] [ Mode: 2 ] => 2 (54, 0,259)
      AgeGr in [ 4 ] [ Mode: 2 ] (19)
            Paytool in [ "H" ] [ Mode: 2 ] => 2 (6,

            Paytool in [ "K" "N" "O" ] [ Mode: 1 ] => 1

      AgeGr in [ 5 ] [ Mode: 3 ] => 3 (0)
OccupGr in [ 6 ] [ Mode: 2 ] (61)
      PCI = 1 [ Mode: 2 ] => 2 (5, 0,8)
```

```
                                                PCI = 0 [ Mode: 2 ] (56)
                                                      AgeGr in [ 3 4 ] [ Mode: 2 ] (50)
                                                            Channel in [ "A" "B" ] [ Mode: 2 ] =>
2 (40, 0,4)

                                                            Channel in [ "D" ] [ Mode: 1 ] => 1
(10, 0,5)

                                                      AgeGr in [ 5 ] [ Mode: 2 ] => 2 (6, 0,833)
                                                OccupGr in [ 7 15 ] [ Mode: 1 ] => 1 (11, 0,636)
                                                OccupGr in [ 9 ] [ Mode: 2 ] (52)
                                                      AgeGr in [ 3 ] [ Mode: 1 ] (30)
                                                            PCI = 1 [ Mode: 2 ] => 2 (6, 0,333)
                                                            PCI = 0 [ Mode: 1 ] => 1 (24, 0,625)
                                                      AgeGr in [ 4 ] [ Mode: 2 ] (16)
                                                            Channel in [ "A" "B" ] [ Mode: 2 ] => 2 (11,
0,636)

                                                            Channel in [ "D" ] [ Mode: 1 ] => 1 (5, 0,6)
                                                      AgeGr in [ 5 ] [ Mode: 1 ] => 1 (6, 0,5)
                                                OccupGr in [ 11 ] [ Mode: 2 ] (27)
                                                      Channel in [ "A" ] [ Mode: 2 ] (22)
                                                            Region in [ 11 ] [ Mode: 2 ] => 2 (16,
0,562)

                                                            Region in [ 12 13 14 15 16 17 18 20 ] [
Mode: 1 ] => 1 (6, 0,5)

                                                      Channel in [ "B" ] [ Mode: 1 ] => 1 (1, 1,0)
                                                      Channel in [ "D" ] [ Mode: 3 ] => 3 (4, 0,75)
                                                OccupGr in [ 14 ] [ Mode: 2 ] (16)
                                                      Region in [ 11 12 13 14 15 16 ] [ Mode: 1 ] => 1

(8, 0,625)

                                                      Region in [ 17 18 20 ] [ Mode: 2 ] => 2 (8, 0,5)
                                          AgeGr in [ 1 ] [ Mode: 3 ] => 3 (0)
                                    NCD = 0 [ Mode: 2 ] (4.711,679)
                                          OccupGr in [ 0 ] [ Mode: 2 ] (1.334,679)
                                                Region in [ 11 ] [ Mode: 2 ] (736)
                                                      AgeGr in [ 1 2 ] [ Mode: 2 ] => 2 (322, 0,425)
                                                      AgeGr in [ 3 ] [ Mode: 2 ] (192)
```

```
                                    ProductL2 in [ "CALT" ] [ Mode: 1 ] => 1 (5, 0,6)
                                    ProductL2 in [ "CUST" ] [ Mode: 2 ] (187)
                                          Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2
(187, 0,513)
                              AgeGr in [ 4 ] [ Mode: 2 ] (177)
                                    VIP = 1 [ Mode: 2 ] => 2 (9, 0,556)
                                    VIP = 0 [ Mode: 2 ] => 2 (168, 0,429)
                              AgeGr in [ 5 ] [ Mode: 2 ] (45)
                                    Payment = T [ Mode: 2 ] (40)
                                          Discount = 1 [ Mode: 2 ] => 2 (25, 0,4)
                                          Discount = 0 [ Mode: 3 ] => 3 (15, 0,533)
                                    Payment = P [ Mode: 2 ] => 2 (5, 0,8)
                        Region in [ 12 13 15 16 ] [ Mode: 2 ] => 2 (59,679, 0,414)
                        Region in [ 14 ] [ Mode: 1 ] => 1 (2, 1,0)
                        Region in [ 17 ] [ Mode: 2 ] (337)
                              Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (337,
0,415)
                        Region in [ 18 ] [ Mode: 2 ] (118)
                              Channel in [ "A" ] [ Mode: 3 ] => 3 (15, 0,533)
                              Channel in [ "B" ] [ Mode: 2 ] (33)
                                    PCI = 1 [ Mode: 2 ] => 2 (6, 0,833)
                                    PCI = 0 [ Mode: 2 ] (27)
                                          AgeGr in [ 1 2 5 ] [ Mode: 1 ] => 1 (13,
0,615)
                                          AgeGr in [ 3 ] [ Mode: 2 ] => 2 (7, 0,714)
                                          AgeGr in [ 4 ] [ Mode: 3 ] => 3 (7, 0,571)
                              Channel in [ "D" ] [ Mode: 2 ] (70)
                                    AgeGr in [ 2 ] [ Mode: 3 ] => 3 (22, 0,545)
                                    AgeGr in [ 1 3 4 5 ] [ Mode: 2 ] => 2 (48, 0,625)
                        Region in [ 20 ] [ Mode: 2 ] (82)
                              Discount = 1 [ Mode: 2 ] => 2 (37, 0,514)
                              Discount = 0 [ Mode: 2 ] => 2 (45, 0,356)
                  OccupGr in [ 1 ] [ Mode: 2 ] (637)
                        VIP = 1 [ Mode: 3 ] => 3 (22, 0,591)
                        VIP = 0 [ Mode: 2 ] (615)
```

```
                                    PCI = 1 [ Mode: 2 ] => 2 (82, 0,366)
                                    PCI = 0 [ Mode: 2 ] (533)
                                          Paytool in [ "H" ] [ Mode: 2 ] (72)
                                                AgeGr in [ 2 3 ] [ Mode: 2 ] (61)
                                                      Region in [ 11 17 ] [ Mode: 2 ] (47)
                                                            Discount = 1 [ Mode: 2 ] => 2

(25, 0,28)
                                                            Discount = 0 [ Mode: 2 ] => 2

(22, 0,727)
                                                      Region in [ 12 13 14 15 16 18 20 ] [

Mode: 2 ] => 2 (14, 0,857)
                                                AgeGr in [ 4 ] [ Mode: 1 ] => 1 (10, 0,6)
                                                AgeGr in [ 5 ] [ Mode: 3 ] => 3 (1, 1,0)
                                                AgeGr in [ 1 ] [ Mode: 2 ] => 2 (0)
                                          Paytool in [ "K" "N" "O" ] [ Mode: 2 ] => 2 (461,

0,403)
                              OccupGr in [ 2 ] [ Mode: 2 ] (884)
                                    Payment = T [ Mode: 2 ] (748)
                                          AgeGr in [ 1 ] [ Mode: 3 ] => 3 (2, 1,0)
                                          AgeGr in [ 2 4 ] [ Mode: 2 ] => 2 (504, 0,46)
                                          AgeGr in [ 3 ] [ Mode: 2 ] (112)
                                                Paytool in [ "H" ] [ Mode: 2 ] (12)
                                                      ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ]
=> 2 (12, 0,583)
                                                Paytool in [ "K" "O" ] [ Mode: 2 ] => 2 (100,

0,43)
                                                Paytool in [ "N" ] [ Mode: 3 ] => 3 (0)
                                          AgeGr in [ 5 ] [ Mode: 2 ] (130)
                                                Channel in [ "A" "D" ] [ Mode: 2 ] (123)
                                                      Paytool in [ "H" "N" "O" ] [ Mode: 3 ] => 3

(16, 0,562)
                                                      Paytool in [ "K" ] [ Mode: 2 ] (107)
                                                            Region in [ 11 ] [ Mode: 2 ] (55)
                                                                  PCI = 1 [ Mode: 2 ] => 2 (8,

0,125)
```

```
                                                    PCI = 0 [ Mode: 2 ] => 2 (47,
0,553)
                                                  Region in [ 12 13 14 15 16 17 18 20 ]
[ Mode: 2 ] => 2 (52, 0,442)
                                                        Channel in [ "B" ] [ Mode: 3 ] => 3 (7, 0,571)
                                            Payment = P [ Mode: 2 ] => 2 (136, 0,471)
                                      OccupGr in [ 3 ] [ Mode: 2 ] (211)
                                            Payment = T [ Mode: 2 ] (176)
                                                  Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (176,
0,398)
                                            Payment = P [ Mode: 2 ] (35)
                                                  Channel in [ "A" ] [ Mode: 2 ] (18)
                                                        Paytool in [ "H" ] [ Mode: 3 ] => 3 (4, 0,75)
                                                        Paytool in [ "K" ] [ Mode: 2 ] => 2 (14, 0,5)
                                                        Paytool in [ "N" "O" ] [ Mode: 1 ] => 1 (0)
                                                  Channel in [ "B" ] [ Mode: 2 ] (10)
                                                        Paytool in [ "H" "N" "O" ] [ Mode: 2 ] => 2 (6,
1,0)
                                                        Paytool in [ "K" ] [ Mode: 3 ] => 3 (4, 0,75)
                                                  Channel in [ "D" ] [ Mode: 3 ] => 3 (7, 0,714)
                                      OccupGr in [ 4 ] [ Mode: 2 ] (272)
                                            Region in [ 11 ] [ Mode: 2 ] (153)
                                                  Payment = T [ Mode: 2 ] => 2 (136, 0,404)
                                                  Payment = P [ Mode: 2 ] (17)
                                                        Paytool in [ "H" ] [ Mode: 1 ] => 1 (7, 0,714)
                                                        Paytool in [ "K" "O" ] [ Mode: 2 ] => 2 (9,
0,889)
                                                        Paytool in [ "N" ] [ Mode: 3 ] => 3 (1, 1,0)
                                            Region in [ 12 16 ] [ Mode: 3 ] => 3 (3, 0,667)
                                            Region in [ 13 ] [ Mode: 1 ] => 1 (1, 1,0)
                                            Region in [ 14 15 18 20 ] [ Mode: 2 ] => 2 (51, 0,412)
                                            Region in [ 17 ] [ Mode: 2 ] (64)
                                                  AgeGr in [ 1 2 ] [ Mode: 2 ] => 2 (29, 0,414)
                                                  AgeGr in [ 3 4 5 ] [ Mode: 2 ] (35)
                                                        Paytool in [ "H" ] [ Mode: 3 ] => 3 (6, 0,667)
```

```
                  Paytool in [ "K" "N" "O" ] [ Mode: 2 ] => 2 (29,
0,621)
OccupGr in [ 5 ] [ Mode: 2 ] (331)
      Payment = T [ Mode: 2 ] (300)
            Region in [ 11 17 18 ] [ Mode: 2 ] (276)
                  Discount = 1 [ Mode: 2 ] (170)
                        VIP = 1 [ Mode: 3 ] => 3 (4, 0,75)
                        VIP = 0 [ Mode: 2 ] => 2 (166, 0,512)
                  Discount = 0 [ Mode: 2 ] => 2 (106, 0,368)
            Region in [ 12 15 20 ] [ Mode: 3 ] => 3 (24, 0,583)
            Region in [ 13 14 16 ] [ Mode: 1 ] => 1 (0)
      Payment = P [ Mode: 2 ] (31)
            Paytool in [ "H" ] [ Mode: 2 ] (17)
                  Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (10, 0,2)
                  Channel in [ "B" ] [ Mode: 1 ] => 1 (7, 0,571)
            Paytool in [ "K" ] [ Mode: 3 ] => 3 (12, 0,75)
            Paytool in [ "N" "O" ] [ Mode: 1 ] => 1 (2, 1,0)
OccupGr in [ 6 ] [ Mode: 2 ] (251)
      AgeGr in [ 2 3 ] [ Mode: 2 ] (237)
            Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (205, 0,429)
            Channel in [ "B" ] [ Mode: 2 ] (32)
                  Region in [ 11 12 13 14 15 16 18 20 ] [ Mode: 1 ]
=> 1 (22, 0,591)
                  Region in [ 17 ] [ Mode: 2 ] => 2 (10, 0,5)
      AgeGr in [ 4 ] [ Mode: 1 ] (12)
            Discount = 1 [ Mode: 1 ] => 1 (7, 0,714)
            Discount = 0 [ Mode: 2 ] => 2 (5, 0,8)
      AgeGr in [ 5 ] [ Mode: 1 ] => 1 (2, 0,5)
      AgeGr in [ 1 ] [ Mode: 2 ] => 2 (0)
OccupGr in [ 7 ] [ Mode: 2 ] (173)
      Paytool in [ "H" ] [ Mode: 1 ] => 1 (20, 0,5)
      Paytool in [ "K" "O" ] [ Mode: 2 ] => 2 (152, 0,5)
      Paytool in [ "N" ] [ Mode: 3 ] => 3 (1, 1,0)
OccupGr in [ 8 ] [ Mode: 2 ] => 2 (62, 0,339)
OccupGr in [ 9 ] [ Mode: 2 ] (96)
```

```
      AgeGr in [ 2 ] [ Mode: 2 ] => 2 (46, 0,348)
      AgeGr in [ 3 5 ] [ Mode: 1 ] => 1 (31, 0,548)
      AgeGr in [ 4 ] [ Mode: 2 ] (19)
            Channel in [ "A" "B" ] [ Mode: 2 ] => 2 (14, 0,643)
            Channel in [ "D" ] [ Mode: 3 ] => 3 (5, 0,8)
      AgeGr in [ 1 ] [ Mode: 3 ] => 3 (0)
OccupGr in [ 10 ] [ Mode: 2 ] (91)
      Paytool in [ "H" "N" ] [ Mode: 3 ] => 3 (16, 0,562)
      Paytool in [ "K" "O" ] [ Mode: 2 ] => 2 (75, 0,36)
OccupGr in [ 11 ] [ Mode: 2 ] (114)
      AgeGr in [ 2 ] [ Mode: 2 ] (76)
            Paytool in [ "H" "O" ] [ Mode: 2 ] => 2 (15, 0,667)
            Paytool in [ "K" ] [ Mode: 2 ] (60)
                  Discount = 1 [ Mode: 2 ] => 2 (39, 0,564)
                  Discount = 0 [ Mode: 2 ] => 2 (21, 0,381)
            Paytool in [ "N" ] [ Mode: 1 ] => 1 (1, 1,0)
      AgeGr in [ 3 ] [ Mode: 2 ] (31)
            Region in [ 11 ] [ Mode: 2 ] (18)
                  Discount = 1 [ Mode: 1 ] => 1 (9, 0,667)
                  Discount = 0 [ Mode: 2 ] => 2 (9, 0,667)
            Region in [ 17 18 ] [ Mode: 2 ] => 2 (12, 0,5)
            Region in [ 20 ] [ Mode: 3 ] => 3 (1, 1,0)
            Region in [ 12 13 14 15 16 ] [ Mode: 1 ] => 1 (0)
      AgeGr in [ 4 ] [ Mode: 3 ] => 3 (5, 0,8)
      AgeGr in [ 5 ] [ Mode: 1 ] => 1 (2, 0,5)
      AgeGr in [ 1 ] [ Mode: 2 ] => 2 (0)
OccupGr in [ 12 ] [ Mode: 2 ] (120)
      Region in [ 11 ] [ Mode: 2 ] (56)
            Channel in [ "A" ] [ Mode: 2 ] => 2 (37, 0,541)
            Channel in [ "B" "D" ] [ Mode: 1 ] => 1 (19, 0,579)
      Region in [ 12 13 14 15 16 17 ] [ Mode: 1 ] => 1 (35, 0,543)
      Region in [ 18 ] [ Mode: 2 ] => 2 (14, 0,429)
      Region in [ 20 ] [ Mode: 2 ] (15)
            Discount = 1 [ Mode: 3 ] => 3 (7, 0,571)
            Discount = 0 [ Mode: 2 ] => 2 (8, 0,75)
```

```
                        OccupGr in [ 13 ] [ Mode: 2 ] (46)
                              Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (46, 0,391)
                        OccupGr in [ 14 ] [ Mode: 2 ] (63)
                              Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (63, 0,444)
                        OccupGr in [ 15 ] [ Mode: 2 ] (26)
                              ProductL2 in [ "CALT" ] [ Mode: 1 ] => 1 (7, 1,0)
                              ProductL2 in [ "CUST" ] [ Mode: 2 ] => 2 (19, 0,474)
                  City in [ 35 ] [ Mode: 2 ] (1.153,194)
                        AgeGr in [ 1 ] [ Mode: 1 ] => 1 (6, 0,5)
                        AgeGr in [ 2 3 4 ] [ Mode: 2 ] (1.095,194)
                              Region in [ 11 12 17 ] [ Mode: 2 ] => 2 (46, 0,391)
                              Region in [ 13 ] [ Mode: 2 ] (10)
                                    Payment = T [ Mode: 2 ] => 2 (6, 0,833)
                                    Payment = P [ Mode: 3 ] => 3 (4, 0,75)
                              Region in [ 15 ] [ Mode: 2 ] (1.037,194)
                                    OccupGr in [ 0 3 4 12 14 15 ] [ Mode: 2 ] => 2 (528,194,
0,369)

                                    OccupGr in [ 1 ] [ Mode: 2 ] (180)
                                          Paytool in [ "H" "N" "O" ] [ Mode: 2 ] => 2 (20, 0,65)
                                          Paytool in [ "K" ] [ Mode: 2 ] (160)
                                                AgeGr in [ 2 3 ] [ Mode: 2 ] (138)
                                                      Payment = T [ Mode: 2 ] (123)
                                                            Channel in [ "A" ] [ Mode: 2 ] (58)
                                                                  PCI = 1 [ Mode: 1 ] (8)
                                                                        Discount = 1 [ Mode: 1 ]

=> 1 (4, 0,75)

                                                                        Discount = 0 [ Mode: 2 ]

=> 2 (4, 0,25)

                                                                  PCI = 0 [ Mode: 2 ] => 2 (50,

0,58)

                                                            Channel in [ "B" ] [ Mode: 2 ] => 2

(22, 0,409)

                                                            Channel in [ "D" ] [ Mode: 2 ] (43)
                                                                  NCD = 1 [ Mode: 1 ] => 1 (15,

0,533)
```

```
                                                                NCD = 0 [ Mode: 2 ] => 2 (28,
0,536)
                                                      Payment = P [ Mode: 1 ] => 1 (15, 0,533)
                                                AgeGr in [ 4 ] [ Mode: 2 ] (22)
                                                      NCD = 1 [ Mode: 1 ] => 1 (9, 0,667)
                                                      NCD = 0 [ Mode: 2 ] => 2 (13, 0,846)
                                          OccupGr in [ 2 ] [ Mode: 2 ] (100)
                                                Payment = T [ Mode: 2 ] => 2 (90, 0,233)
                                                Payment = P [ Mode: 3 ] => 3 (10, 0,6)
                                          OccupGr in [ 5 ] [ Mode: 2 ] (35)
                                                Discount = 1 [ Mode: 2 ] => 2 (30, 0,367)
                                                Discount = 0 [ Mode: 2 ] => 2 (5, 1,0)
                                          OccupGr in [ 6 ] [ Mode: 2 ] (48)
                                                Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (45, 0,378)
                                                Channel in [ "B" ] [ Mode: 1 ] => 1 (3, 1,0)
                                          OccupGr in [ 7 ] [ Mode: 2 ] (51)
                                                Channel in [ "A" ] [ Mode: 2 ] (29)
                                                      Discount = 1 [ Mode: 2 ] => 2 (24, 0,333)
                                                      Discount = 0 [ Mode: 1 ] => 1 (5, 0,8)
                                                Channel in [ "B" "D" ] [ Mode: 2 ] => 2 (22, 0,591)
                                          OccupGr in [ 8 ] [ Mode: 2 ] (13)
                                                NCD = 1 [ Mode: 2 ] => 2 (9, 0,444)
                                                NCD = 0 [ Mode: 1 ] => 1 (4, 0,75)
                                          OccupGr in [ 9 ] [ Mode: 2 ] (14)
                                                Channel in [ "A" ] [ Mode: 3 ] => 3 (4, 0,75)
                                                Channel in [ "B" "D" ] [ Mode: 1 ] => 1 (10, 0,6)
                                          OccupGr in [ 10 ] [ Mode: 2 ] (18)
                                                PCI = 1 [ Mode: 1 ] => 1 (6, 0,667)
                                                PCI = 0 [ Mode: 2 ] => 2 (12, 0,75)
                                          OccupGr in [ 11 ] [ Mode: 2 ] (42)
                                                AgeGr in [ 2 3 4 ] [ Mode: 2 ] => 2 (42, 0,476)
                                          OccupGr in [ 13 ] [ Mode: 1 ] => 1 (8, 0,5)
                                    Region in [ 16 20 ] [ Mode: 1 ] => 1 (2, 1,0)
                                    Region in [ 14 18 ] [ Mode: 3 ] => 3 (0)
                              AgeGr in [ 5 ] [ Mode: 2 ] (52)
```

```
                                    Discount = 1 [ Mode: 1 ] => 1 (46, 0,522)
                                    Discount = 0 [ Mode: 2 ] => 2 (6, 0,333)
                              City in [ 41 ] [ Mode: 2 ] (81,014)
                                    Payment = T [ Mode: 2 ] => 2 (66, 0,5)
                                    Payment = P [ Mode: 2 ] => 2 (15,014, 0,134)
                              City in [ 45 ] [ Mode: 2 ] (22,004)
                                    NCD = 1 [ Mode: 1 ] => 1 (12,002, 0,583)
                                    NCD = 0 [ Mode: 2 ] => 2 (10,002, 0,7)
                              City in [ 48 ] [ Mode: 2 ] (292,048)
                                    VIP = 1 [ Mode: 3 ] => 3 (5, 0,8)
                                    VIP = 0 [ Mode: 2 ] => 2 (287,048, 0,362)
                              City in [ 61 ] [ Mode: 2 ] (10,002)
                                    NCD = 1 [ Mode: 2 ] => 2 (5,001, 0,2)
                                    NCD = 0 [ Mode: 2 ] => 2 (5,001, 0,8)
                        Sex in [ "C" ] [ Mode: 3 ] => 3 (0)
                  ProductL2 in [ "CORTA" "W" ] [ Mode: 2 ] (8.283,943)
                        Sex in [ "" "C" ] [ Mode: 1 ] => 1 (6, 0,667)
                        Sex in [ "E" ] [ Mode: 2 ] => 2 (4.494,94, 0,224)
                        Sex in [ "K" ] [ Mode: 2 ] (3.783,003)
                              NCD = 1 [ Mode: 2 ] (1.540,999)
                                    AgeGr in [ 2 4 5 ] [ Mode: 2 ] => 2 (1.106,999, 0,346)
                                    AgeGr in [ 3 ] [ Mode: 2 ] (434)
                                          Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (434, 0,359)
                                    AgeGr in [ 1 ] [ Mode: 1 ] => 1 (0)
                              NCD = 0 [ Mode: 2 ] => 2 (2.242,004, 0,293)
Noins > 1 [ Mode: 2 ] (29.384,824)
      AgeGr in [ 1 2 ] [ Mode: 2 ] => 2 (6.321,066, 0,366)
      AgeGr in [ 3 4 5 ] [ Mode: 2 ] (23.063,758)
            ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] (16.779,777)
                  NCD = 1 [ Mode: 2 ] => 2 (6.322,96, 0,359)
                  NCD = 0 [ Mode: 2 ] (10.456,817)
                        City in [ 1 ] [ Mode: 2 ] (139,106)
                              Noins <= 4 [ Mode: 2 ] (124,106)
                                    Noins <= 3 [ Mode: 2 ] => 2 (81,08, 0,358)
                                    Noins > 3 [ Mode: 2 ] => 2 (43,027, 0,326)
```

```
                                    Noins > 4 [ Mode: 3 ] => 3 (15, 0,6)
                               City in [ 3 5 23 32 37 38 39 40 41 42 44 46 50 51 55 59 61 63 74 77 78 80 81 ]
[ Mode: 2 ] => 2 (232,178, 0,491)
                               City in [ 6 ] [ Mode: 2 ] (1.243,951)
                                    Noins <= 4 [ Mode: 2 ] (1.178,951)
                                         AgeGr in [ 3 ] [ Mode: 2 ] (644,119)
                                              PCI = 1 [ Mode: 2 ] => 2 (120,119, 0,55)
                                              PCI = 0 [ Mode: 2 ] => 2 (524, 0,452)
                                         AgeGr in [ 4 5 ] [ Mode: 2 ] => 2 (534,832, 0,551)
                                    Noins > 4 [ Mode: 3 ] => 3 (65, 0,538)
                               City in [ 7 ] [ Mode: 2 ] (373,25)
                                    Region in [ 11 12 15 18 20 ] [ Mode: 2 ] => 2 (23,25, 0,393)
                                    Region in [ 13 ] [ Mode: 2 ] (16)
                                         ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (16, 0,562)
                                    Region in [ 14 ] [ Mode: 3 ] => 3 (0)
                                    Region in [ 16 ] [ Mode: 2 ] (324)
                                         Sex in [ "E" ] [ Mode: 2 ] (242)
                                              Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (156, 0,513)
                                              Channel in [ "B" ] [ Mode: 2 ] (86)
                                                   Payment = T [ Mode: 2 ] (78)
                                                        ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (10,
0,5)
                                                        ProductL2 in [ "CUST" ] [ Mode: 2 ] (68)
                                                             Discount = 1 [ Mode: 2 ] => 2 (38, 0,289)
                                                             Discount = 0 [ Mode: 2 ] => 2 (30, 0,433)
                                                   Payment = P [ Mode: 1 ] => 1 (8, 0,75)
                                         Sex in [ "K" ] [ Mode: 2 ] => 2 (82, 0,439)
                                         Sex in [ "" "C" ] [ Mode: 3 ] => 3 (0)
                                    Region in [ 17 ] [ Mode: 2 ] (10)
                                         Sex in [ "" "C" "E" ] [ Mode: 3 ] => 3 (4, 0,75)
                                         Sex in [ "K" ] [ Mode: 2 ] => 2 (6, 0,833)
                               City in [ 9 ] [ Mode: 2 ] (68,013)
                                    Sex in [ "E" ] [ Mode: 2 ] (51,013)
                                         AgeGr in [ 3 4 ] [ Mode: 2 ] (46,006)
                                              Discount = 1 [ Mode: 2 ] => 2 (31,007, 0,323)
```

```
                                                Discount = 0 [ Mode: 2 ] => 2 (15, 0,6)
                                          AgeGr in [ 5 ] [ Mode: 3 ] => 3 (5,007, 0,599)
                                    Sex in [ "K" ] [ Mode: 2 ] (17)
                                          OccupGr in [ 0 4 ] [ Mode: 1 ] => 1 (4, 0,5)
                                          OccupGr in [ 1 2 3 5 6 7 8 9 10 11 12 13 14 15 ] [ Mode: 3 ] => 3
(13, 0,615)
                                    Sex in [ "" "C" ] [ Mode: 1 ] => 1 (0)
                              City in [ 10 ] [ Mode: 2 ] (37,028)
                                    ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (37,028, 0,325)
                              City in [ 2 4 8 11 12 13 14 15 18 19 21 24 25 28 29 30 36 43 49 56 57 58 60 62
65 66 68 69 70 71 72 73 76 79 ] [ Mode: 3 ] => 3 (10,008, 0,7)
                              City in [ 16 ] [ Mode: 2 ] (293,224)
                                    Paytool in [ "H" ] [ Mode: 2 ] (24)
                                          Noins <= 3 [ Mode: 1 ] => 1 (11, 0,636)
                                          Noins > 3 [ Mode: 3 ] => 3 (13, 0,538)
                                    Paytool in [ "K" ] [ Mode: 2 ] (265,224)
                                          OccupGr in [ 0 ] [ Mode: 2 ] (65,168)
                                                Addpr = 1 [ Mode: 2 ] (44,168)
                                                      AgeGr in [ 3 ] [ Mode: 2 ] (29)
                                                            Noins <= 3 [ Mode: 2 ] => 2 (15, 0,733)
                                                            Noins > 3 [ Mode: 3 ] => 3 (14, 0,643)
                                                      AgeGr in [ 4 ] [ Mode: 3 ] => 3 (13, 0,692)
                                                      AgeGr in [ 5 ] [ Mode: 2 ] => 2 (2,168, 0,961)
                                                Addpr = 0 [ Mode: 2 ] => 2 (21, 0,571)
                                          OccupGr in [ 1 ] [ Mode: 2 ] (84,028)
                                                Noins <= 3 [ Mode: 2 ] => 2 (55, 0,4)
                                                Noins > 3 [ Mode: 2 ] => 2 (29,028, 0,656)
                                          OccupGr in [ 2 6 9 10 12 14 ] [ Mode: 2 ] => 2 (41, 0,415)
                                          OccupGr in [ 3 ] [ Mode: 2 ] (23)
                                                PCI = 1 [ Mode: 3 ] => 3 (4, 1,0)
                                                PCI = 0 [ Mode: 2 ] => 2 (19, 0,474)
                                          OccupGr in [ 4 ] [ Mode: 2 ] (20,028)
                                                Channel in [ "A" ] [ Mode: 3 ] => 3 (5,028, 0,597)
                                                Channel in [ "B" ] [ Mode: 2 ] => 2 (7, 0,714)
                                                Channel in [ "D" ] [ Mode: 1 ] => 1 (8, 0,625)
```

```
            OccupGr in [ 5 15 ] [ Mode: 1 ] => 1 (11, 0,545)
            OccupGr in [ 8 ] [ Mode: 2 ] (16)
                  AgeGr in [ 3 ] [ Mode: 2 ] => 2 (10, 0,4)
                  AgeGr in [ 4 ] [ Mode: 1 ] => 1 (6, 0,667)
                  AgeGr in [ 5 ] [ Mode: 3 ] => 3 (0)
            OccupGr in [ 7 11 13 ] [ Mode: 3 ] => 3 (5, 0,6)
      Paytool in [ "N" "O" ] [ Mode: 2 ] => 2 (4, 1,0)
City in [ 17 ] [ Mode: 2 ] (19,015)
      Addpr = 1 [ Mode: 2 ] => 2 (15,013, 0,267)
      Addpr = 0 [ Mode: 2 ] => 2 (4,002, 0,5)
City in [ 20 ] [ Mode: 2 ] (42,032)
      Noins <= 2 [ Mode: 2 ] => 2 (9,02, 0,555)
      Noins > 2 [ Mode: 2 ] (33,012)
            Discount = 1 [ Mode: 2 ] => 2 (23,008, 0,348)
            Discount = 0 [ Mode: 2 ] => 2 (10,004, 0,8)
City in [ 22 47 52 53 64 ] [ Mode: 1 ] => 1 (19,015, 0,684)
City in [ 26 ] [ Mode: 2 ] (22,017)
      Noins <= 4 [ Mode: 2 ] (18,017)
            Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (18,017, 0,556)
      Noins > 4 [ Mode: 3 ] => 3 (4, 0,75)
City in [ 27 ] [ Mode: 2 ] (45,009)
      Sex in [ "" "C" ] [ Mode: 3 ] => 3 (0)
      Sex in [ "E" "K" ] [ Mode: 2 ] => 2 (45,009, 0,511)
City in [ 31 ] [ Mode: 2 ] (22,017)
      Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (22,017, 0,409)
City in [ 33 ] [ Mode: 2 ] (131,1)
      PCI = 1 [ Mode: 2 ] (18,013)
            AgeGr in [ 3 ] [ Mode: 2 ] => 2 (10,013, 0,9)
            AgeGr in [ 4 5 ] [ Mode: 3 ] (8)
                  Addpr = 1 [ Mode: 1 ] => 1 (4, 0,5)
                  Addpr = 0 [ Mode: 3 ] => 3 (4, 1,0)
      PCI = 0 [ Mode: 3 ] => 3 (113,088, 0,504)
City in [ 34 ] [ Mode: 2 ] (6.441,9)
      Noins <= 4 [ Mode: 2 ] (6.232,9)
```

```
                                    OccupGr in [ 0 1 2 3 4 5 6 7 8 10 11 12 13 ] [ Mode: 2 ] => 2
(6.026,9, 0,446)

                                    OccupGr in [ 9 14 15 ] [ Mode: 2 ] (206)
                                          Region in [ 11 ] [ Mode: 2 ] (91)
                                                Paytool in [ "H" ] [ Mode: 3 ] => 3 (5, 0,8)
                                                Paytool in [ "K" ] [ Mode: 2 ] (85)
                                                      Payment = T [ Mode: 2 ] => 2 (73, 0,274)
                                                      Payment = P [ Mode: 2 ] => 2 (12, 0,5)
                                                Paytool in [ "N" "O" ] [ Mode: 1 ] => 1 (1, 1,0)
                                          Region in [ 12 15 ] [ Mode: 2 ] => 2 (5, 0,6)
                                          Region in [ 17 ] [ Mode: 2 ] (68)
                                                Paytool in [ "H" ] [ Mode: 1 ] => 1 (8, 0,625)
                                                Paytool in [ "K" ] [ Mode: 2 ] (58)
                                                      Noins <= 2 [ Mode: 3 ] => 3 (25, 0,56)
                                                      Noins > 2 [ Mode: 2 ] (33)
                                                            OccupGr in [ 9 15 ] [ Mode: 2 ] (22)
                                                                  AgeGr in [ 3 5 ] [ Mode: 2 ] => 2 (14,
0,429)

                                                                  AgeGr in [ 4 ] [ Mode: 1 ] => 1 (8,
0,625)

                                                            OccupGr in [ 14 ] [ Mode: 2 ] => 2 (11,
0,727)

                                                Paytool in [ "N" "O" ] [ Mode: 3 ] => 3 (2, 1,0)
                                          Region in [ 18 ] [ Mode: 2 ] (24)
                                                AgeGr in [ 3 ] [ Mode: 2 ] (18)
                                                      Noins <= 2 [ Mode: 2 ] => 2 (12, 0,333)
                                                      Noins > 2 [ Mode: 2 ] => 2 (6, 0,833)
                                                AgeGr in [ 4 ] [ Mode: 1 ] => 1 (6, 0,667)
                                                AgeGr in [ 5 ] [ Mode: 2 ] => 2 (0)
                                          Region in [ 20 ] [ Mode: 2 ] (18)
                                                AgeGr in [ 3 5 ] [ Mode: 2 ] => 2 (11, 0,909)
                                                AgeGr in [ 4 ] [ Mode: 1 ] => 1 (7, 0,571)
                                          Region in [ 13 14 16 ] [ Mode: 1 ] => 1 (0)
                              Noins > 4 [ Mode: 2 ] (209)
                                    Discount = 1 [ Mode: 2 ] => 2 (117, 0,556)
```

```
            Discount = 0 [ Mode: 2 ] => 2 (92, 0,402)
City in [ 35 ] [ Mode: 2 ] (989,756)
      Paytool in [ "H" "K" ] [ Mode: 2 ] (966,756)
            ProductL2 in [ "CALT" ] [ Mode: 2 ] (112,284)
                  Addpr = 1 [ Mode: 2 ] (107,284)
                        PCI = 1 [ Mode: 2 ] (14)
                              AgeGr in [ 3 ] [ Mode: 2 ] => 2 (9, 0,556)
                              AgeGr in [ 4 5 ] [ Mode: 1 ] => 1 (5, 0,6)
                        PCI = 0 [ Mode: 2 ] => 2 (93,284, 0,517)
                  Addpr = 0 [ Mode: 1 ] => 1 (5, 0,6)
            ProductL2 in [ "CUST" ] [ Mode: 2 ] (854,472)
                  Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (762,377, 0,412)
                  Channel in [ "B" ] [ Mode: 2 ] (92,095)
                        Payment = T [ Mode: 2 ] (84)
                              Noins <= 2 [ Mode: 2 ] => 2 (29, 0,207)
                              Noins > 2 [ Mode: 2 ] (55)
                                    Sex in [ "" "C" "E" "K" ] [ Mode: 2 ] => 2
(55, 0,618)

                        Payment = P [ Mode: 1 ] => 1 (8,095, 0,753)
      Paytool in [ "N" ] [ Mode: 2 ] => 2 (7, 0,714)
      Paytool in [ "O" ] [ Mode: 2 ] (16)
            AgeGr in [ 3 5 ] [ Mode: 2 ] => 2 (8, 0,75)
            AgeGr in [ 4 ] [ Mode: 1 ] => 1 (8, 0,5)
City in [ 45 ] [ Mode: 2 ] (32,025)
      Payment = T [ Mode: 2 ] => 2 (24,021, 0,375)
      Payment = P [ Mode: 2 ] => 2 (8,003, 0,625)
City in [ 48 ] [ Mode: 2 ] (230,132)
      Region in [ 11 20 ] [ Mode: 1 ] => 1 (11,088, 0,543)
      Region in [ 12 14 16 18 ] [ Mode: 3 ] => 3 (0)
      Region in [ 13 17 ] [ Mode: 2 ] => 2 (10, 0,6)
      Region in [ 15 ] [ Mode: 2 ] (209,044)
            AgeGr in [ 3 4 5 ] [ Mode: 2 ] => 2 (209,044, 0,373)
City in [ 54 ] [ Mode: 2 ] (41,031)
      Paytool in [ "H" "K" ] [ Mode: 2 ] => 2 (37,031, 0,433)
      Paytool in [ "N" "O" ] [ Mode: 3 ] => 3 (4, 0,75)
```

```
            City in [ 67 ] [ Mode: 2 ] (25,012)
                  Region in [ 11 14 ] [ Mode: 1 ] => 1 (5,012, 0,799)
                  Region in [ 12 15 16 18 20 ] [ Mode: 3 ] => 3 (1, 1,0)
                  Region in [ 13 ] [ Mode: 2 ] (18)
                        AgeGr in [ 3 ] [ Mode: 2 ] => 2 (9, 0,667)
                        AgeGr in [ 4 ] [ Mode: 3 ] => 3 (8, 0,625)
                        AgeGr in [ 5 ] [ Mode: 1 ] => 1 (1, 1,0)
                  Region in [ 17 ] [ Mode: 2 ] => 2 (1, 1,0)
ProductL2 in [ "CORTA" "W" ] [ Mode: 2 ] (6.283,981)
      City in [ 1 9 18 22 33 38 42 47 61 63 64 67 74 77 ] [ Mode: 2 ] => 2 (302,24, 0,341)
      City in [ 6 ] [ Mode: 2 ] (472,376)
            Region in [ 11 ] [ Mode: 2 ] (17,075)
                  Addpr = 1 [ Mode: 1 ] => 1 (6, 0,833)
                  Addpr = 0 [ Mode: 2 ] => 2 (11,075, 0,451)
            Region in [ 12 15 18 ] [ Mode: 2 ] => 2 (6,075, 0,658)
            Region in [ 13 ] [ Mode: 2 ] (439,15)
                  Paytool in [ "H" ] [ Mode: 2 ] (46)
                        Discount = 1 [ Mode: 2 ] (31)
                              AgeGr in [ 3 ] [ Mode: 2 ] => 2 (11, 0,455)
                              AgeGr in [ 4 ] [ Mode: 2 ] (19)
                                    Sex in [ "E" ] [ Mode: 2 ] (10)
                                          PCI = 1 [ Mode: 1 ] => 1 (4, 0,5)
                                          PCI = 0 [ Mode: 2 ] => 2 (6, 0,667)
                                    Sex in [ "" "C" "K" ] [ Mode: 1 ] => 1 (9, 0,667)
                              AgeGr in [ 5 ] [ Mode: 1 ] => 1 (1, 1,0)
                        Discount = 0 [ Mode: 2 ] => 2 (15, 0,667)
                  Paytool in [ "K" "N" ] [ Mode: 2 ] => 2 (387,15, 0,418)
                  Paytool in [ "O" ] [ Mode: 1 ] => 1 (6, 0,5)
            Region in [ 14 16 17 20 ] [ Mode: 3 ] => 3 (10,075, 0,596)
      City in [ 7 ] [ Mode: 2 ] (176,14)
            ProductL2 in [ "CORTA" ] [ Mode: 2 ] (172,112)
                  Paytool in [ "H" ] [ Mode: 2 ] (15)
                        NCD = 1 [ Mode: 3 ] => 3 (4, 0,75)
                        NCD = 0 [ Mode: 1 ] => 1 (11, 0,545)
                  Paytool in [ "K" "N" ] [ Mode: 2 ] => 2 (152,112, 0,355)
```

```
                                    Paytool in [ "O" ] [ Mode: 1 ] => 1 (5, 0,6)
                              ProductL2 in [ "W" ] [ Mode: 1 ] => 1 (4,028, 1,0)
                        City in [ 2 3 4 5 8 11 12 13 14 15 21 24 25 27 28 29 30 32 36 37 40 43 44 49 51 56 57
58 60 62 65 66 69 70 71 72 73 76 78 79 80 ] [ Mode: 3 ] => 3 (46,037, 0,522)
                        City in [ 10 ] [ Mode: 2 ] (22,018)
                              Addpr = 1 [ Mode: 1 ] => 1 (7,004, 0,714)
                              Addpr = 0 [ Mode: 3 ] => 3 (15,014, 0,666)
                        City in [ 16 ] [ Mode: 2 ] (208,166)
                              NCD = 1 [ Mode: 2 ] (53)
                                    AgeGr in [ 3 5 ] [ Mode: 2 ] => 2 (25, 0,2)
                                    AgeGr in [ 4 ] [ Mode: 2 ] (28)
                                          Sex in [ "E" ] [ Mode: 2 ] (21)
                                                PCI = 1 [ Mode: 1 ] => 1 (4, 0,5)
                                                PCI = 0 [ Mode: 2 ] => 2 (17, 0,706)
                                          Sex in [ "" "C" "K" ] [ Mode: 3 ] => 3 (7, 0,857)
                              NCD = 0 [ Mode: 2 ] (155,166)
                                    Noins <= 3 [ Mode: 2 ] => 2 (98,132, 0,418)
                                    Noins > 3 [ Mode: 2 ] => 2 (57,033, 0,228)
                        City in [ 17 ] [ Mode: 2 ] (14,011)
                              NCD = 1 [ Mode: 3 ] => 3 (5, 0,6)
                              NCD = 0 [ Mode: 1 ] => 1 (9,011, 0,666)
                        City in [ 19 23 26 31 39 45 46 50 52 53 55 68 81 ] [ Mode: 1 ] => 1 (66,053, 0,666)
                        City in [ 20 ] [ Mode: 2 ] (25,02)
                              OccupGr in [ 0 1 ] [ Mode: 2 ] => 2 (21,02, 0,428)
                              OccupGr in [ 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ] [ Mode: 1 ] => 1 (4, 0,75)
                        City in [ 34 ] [ Mode: 2 ] (4.039,212)
                              Addpr = 1 [ Mode: 2 ] (957,642)
                                    Discount = 1 [ Mode: 2 ] => 2 (504,642, 0,194)
                                    Discount = 0 [ Mode: 2 ] (453)
                                          Payment = T [ Mode: 2 ] (363)
                                                Paytool in [ "H" ] [ Mode: 2 ] => 2 (37, 0,378)
                                                Paytool in [ "K" ] [ Mode: 2 ] (324)
                                                      OccupGr in [ 0 ] [ Mode: 2 ] (113)
                                                            Channel in [ "A" ] [ Mode: 2 ] (86)
                                                                  Noins <= 3 [ Mode: 2 ] (65)
```

```
                                                    PCI = 1 [ Mode: 1 ] => 1 (11, 0,727)
                                                    PCI = 0 [ Mode: 2 ] (54)
                                                          Region in [ 11 ] [ Mode: 2 ]
(29)
                                                                AgeGr in [ 3 ] [ Mode: 1 ]
=> 1 (16, 0,562)
                                                                AgeGr in [ 4 ] [ Mode: 2 ]
=> 2 (10, 0,6)
                                                                AgeGr in [ 5 ] [ Mode: 3 ]
=> 3 (3, 1,0)
                                                          Region in [ 12 15 20 ] [ Mode: 2
] => 2 (6, 0,5)
                                                          Region in [ 17 ] [ Mode: 2 ]
(17)
                                                                Sex in [ "E" ] [ Mode: 1 ]
=> 1 (7, 0,571)
                                                                Sex in [ "" "C" "K" ] [
Mode: 3 ] => 3 (10, 0,6)
                                                          Region in [ 18 ] [ Mode: 1 ] =>
1 (2, 1,0)
                                                          Region in [ 13 14 16 ] [ Mode: 3
] => 3 (0)
                                                    Noins > 3 [ Mode: 3 ] => 3 (21, 0,571)
                                              Channel in [ "B" ] [ Mode: 3 ] => 3 (9, 0,667)
                                              Channel in [ "D" ] [ Mode: 2 ] => 2 (18, 0,444)
                                        OccupGr in [ 1 ] [ Mode: 2 ] (82)
                                              Noins <= 4 [ Mode: 2 ] (77)
                                                    PCI = 1 [ Mode: 2 ] => 2 (11, 0,636)
                                                    PCI = 0 [ Mode: 2 ] (66)
                                                          Channel in [ "A" ] [ Mode: 2 ] => 2
(39, 0,308)
                                                          Channel in [ "B" ] [ Mode: 1 ] => 1
(11, 0,545)
                                                          Channel in [ "D" ] [ Mode: 3 ] => 3
(16, 0,562)
```

```
                                    Noins > 4 [ Mode: 3 ] => 3 (5, 0,8)
                              OccupGr in [ 2 ] [ Mode: 2 ] (27)
                                    AgeGr in [ 3 5 ] [ Mode: 3 ] => 3 (17, 0,588)
                                    AgeGr in [ 4 ] [ Mode: 2 ] => 2 (10, 0,5)
                              OccupGr in [ 3 ] [ Mode: 2 ] (23)
                                    Sex in [ "E" ] [ Mode: 2 ] (19)
                                          PCI = 1 [ Mode: 1 ] => 1 (5, 0,6)
                                          PCI = 0 [ Mode: 2 ] => 2 (14, 0,357)
                                    Sex in [ "" "C" "K" ] [ Mode: 1 ] => 1 (4, 1,0)
                              OccupGr in [ 4 ] [ Mode: 2 ] (26)
                                    Sex in [ "E" ] [ Mode: 2 ] (20)
                                          Noins <= 3 [ Mode: 2 ] => 2 (8, 0,75)
                                          Noins > 3 [ Mode: 1 ] => 1 (12, 0,583)
                                    Sex in [ "K" ] [ Mode: 3 ] => 3 (6, 0,667)
                                    Sex in [ "" "C" ] [ Mode: 1 ] => 1 (0)
                              OccupGr in [ 5 ] [ Mode: 1 ] (8)
                                    Noins <= 2 [ Mode: 3 ] => 3 (4, 0,75)
                                    Noins > 2 [ Mode: 1 ] => 1 (4, 1,0)
                              OccupGr in [ 6 10 12 14 15 ] [ Mode: 2 ] => 2 (17,
0,588)
                              OccupGr in [ 7 8 11 13 ] [ Mode: 3 ] => 3 (16, 0,688)
                              OccupGr in [ 9 ] [ Mode: 2 ] (12)
                                    Noins <= 2 [ Mode: 1 ] => 1 (4, 1,0)
                                    Noins > 2 [ Mode: 2 ] => 2 (8, 0,375)
                        Paytool in [ "O" ] [ Mode: 1 ] => 1 (2, 1,0)
                        Paytool in [ "N" ] [ Mode: 3 ] => 3 (0)
                  Payment = P [ Mode: 2 ] (90)
                        Channel in [ "A" "D" ] [ Mode: 3 ] => 3 (77, 0,506)
                        Channel in [ "B" ] [ Mode: 2 ] => 2 (13, 0,615)
Addpr = 0 [ Mode: 2 ] (3.081,57)
      Sex in [ "" ] [ Mode: 1 ] => 1 (2,571, 0,5)
      Sex in [ "E" ] [ Mode: 2 ] (1.914)
            OccupGr in [ 0 ] [ Mode: 2 ] (641)
                  Paytool in [ "H" "K" "N" ] [ Mode: 2 ] (624)
                        AgeGr in [ 3 4 ] [ Mode: 2 ] (582)
```

```
                                                
```

```
OccupGr in [ 4 ] [ Mode: 2 ] (189)
      Paytool in [ "H" ] [ Mode: 2 ] (15)
            Channel in [ "A" ] [ Mode: 3 ] => 3 (9, 0,556)
            Channel in [ "B" "D" ] [ Mode: 1 ] => 1 (6, 0,667)
      Paytool in [ "K" ] [ Mode: 2 ] (170)
            ProductL2 in [ "CORTA" ] [ Mode: 2 ] => 2 (157, 0,35)
            ProductL2 in [ "W" ] [ Mode: 3 ] => 3 (13, 0,615)
      Paytool in [ "N" ] [ Mode: 2 ] => 2 (4, 0,75)
      Paytool in [ "O" ] [ Mode: 1 ] => 1 (0)
OccupGr in [ 6 ] [ Mode: 2 ] (41)
      VIP = 1 [ Mode: 2 ] => 2 (5, 0,6)
      VIP = 0 [ Mode: 2 ] (36)
            Payment = T [ Mode: 2 ] (31)
                  Region in [ 11 18 20 ] [ Mode: 2 ] => 2 (22,
0,455)
                  Region in [ 17 ] [ Mode: 1 ] => 1 (9, 0,556)
                  Region in [ 12 13 14 15 16 ] [ Mode: 3 ] => 3 (0)
            Payment = P [ Mode: 3 ] => 3 (5, 0,8)
OccupGr in [ 8 ] [ Mode: 2 ] (93)
      Region in [ 11 ] [ Mode: 2 ] (48)
            Noins <= 4 [ Mode: 2 ] => 2 (44, 0,318)
            Noins > 4 [ Mode: 3 ] => 3 (4, 0,75)
      Region in [ 12 13 14 15 16 18 ] [ Mode: 3 ] => 3 (9, 0,778)
      Region in [ 17 ] [ Mode: 2 ] (28)
            AgeGr in [ 3 5 ] [ Mode: 2 ] (16)
                  NCD = 1 [ Mode: 3 ] => 3 (8, 0,625)
                  NCD = 0 [ Mode: 1 ] => 1 (8, 0,5)
            AgeGr in [ 4 ] [ Mode: 2 ] => 2 (12, 0,5)
      Region in [ 20 ] [ Mode: 1 ] => 1 (8, 0,625)
OccupGr in [ 9 ] [ Mode: 2 ] (44)
      Channel in [ "A" ] [ Mode: 1 ] => 1 (28, 0,5)
      Channel in [ "B" ] [ Mode: 2 ] => 2 (2, 1,0)
      Channel in [ "D" ] [ Mode: 3 ] => 3 (14, 0,571)
OccupGr in [ 10 ] [ Mode: 2 ] (19)
      NCD = 1 [ Mode: 1 ] => 1 (5, 0,6)
```

```
                        NCD = 0 [ Mode: 2 ] => 2 (14, 0,286)
                  OccupGr in [ 12 ] [ Mode: 3 ] => 3 (18, 0,556)
                  OccupGr in [ 13 ] [ Mode: 3 ] (16)
                        NCD = 1 [ Mode: 3 ] => 3 (6, 1,0)
                        NCD = 0 [ Mode: 1 ] => 1 (10, 0,7)
                  OccupGr in [ 14 ] [ Mode: 2 ] (25)
                        NCD = 1 [ Mode: 3 ] => 3 (5, 0,6)
                        NCD = 0 [ Mode: 2 ] (20)
                              PCI = 1 [ Mode: 2 ] => 2 (6, 0,333)
                              PCI = 0 [ Mode: 1 ] => 1 (14, 0,571)
            Sex in [ "K" ] [ Mode: 2 ] => 2 (1.164,999, 0,384)
            Sex in [ "C" ] [ Mode: 3 ] => 3 (0)
City in [ 35 ] [ Mode: 2 ] (680,541)
      Addpr = 1 [ Mode: 2 ] (144,108)
            Channel in [ "A" "D" ] [ Mode: 2 ] => 2 (125,108, 0,304)
            Channel in [ "B" ] [ Mode: 1 ] => 1 (19, 0,684)
      Addpr = 0 [ Mode: 2 ] => 2 (536,433, 0,347)
City in [ 41 ] [ Mode: 2 ] (53,025)
      NCD = 1 [ Mode: 2 ] => 2 (12, 0,667)
      NCD = 0 [ Mode: 2 ] (41,025)
            Channel in [ "A" ] [ Mode: 2 ] (26)
                  Region in [ 11 ] [ Mode: 2 ] (14)
                        Sex in [ "E" ] [ Mode: 2 ] => 2 (9, 0,778)
                        Sex in [ "" "C" "K" ] [ Mode: 3 ] => 3 (5, 0,6)
                  Region in [ 12 13 14 16 17 18 20 ] [ Mode: 1 ] => 1 (11, 0,727)
                  Region in [ 15 ] [ Mode: 3 ] => 3 (1, 1,0)
            Channel in [ "B" ] [ Mode: 3 ] => 3 (6, 0,833)
            Channel in [ "D" ] [ Mode: 3 ] (9,025)
                  Noins <= 2 [ Mode: 1 ] => 1 (4,025, 0,747)
                  Noins > 2 [ Mode: 3 ] => 3 (5, 0,8)
City in [ 48 ] [ Mode: 2 ] (114,091)
      Channel in [ "A" ] [ Mode: 2 ] => 2 (53,036, 0,245)
      Channel in [ "B" "D" ] [ Mode: 2 ] (61,055)
            NCD = 1 [ Mode: 2 ] => 2 (20, 0,5)
            NCD = 0 [ Mode: 3 ] => 3 (41,055, 0,609)
```

```
                        City in [ 54 ] [ Mode: 1 ] (34,027)
                              Payment = T [ Mode: 1 ] (30,022)
                                    Paytool in [ "H" ] [ Mode: 3 ] => 3 (4, 0,75)
                                    Paytool in [ "K" "N" "O" ] [ Mode: 1 ] => 1 (26,022, 0,769)
                              Payment = P [ Mode: 3 ] => 3 (4,005, 0,999)
                        City in [ 59 ] [ Mode: 1 ] (31,025)
                              Noins <= 2 [ Mode: 3 ] (9,02)
                                    AgeGr in [ 3 5 ] [ Mode: 3 ] => 3 (5,02, 0,998)
                                    AgeGr in [ 4 ] [ Mode: 1 ] => 1 (4, 0,75)
                              Noins > 2 [ Mode: 1 ] => 1 (22,005, 0,636)
ProductL2 in [ "S" ] [ Mode: 1 ] => 1 (4.745,07, 0,991)
```

APPENDIX C

The decision tree of model 3.

```
Loyal in [ "LC" "NC" ] [ Mode: 2 ] (60.148,325)
      Noins <= 1 [ Mode: 2 ] (33.822,157)
            Addpr = 1 [ Mode: 2 ] (5.139,031)
                  NCD = 1 [ Mode: 2 ] => 2 (756,002, 0,327)
                  NCD = 0 [ Mode: 2 ] (4.383,029)
                        ProductL2 in [ "CALT" "CORTA" ] [ Mode: 2 ] (1.322)
                              AgeGr in [ 2 3 ] [ Mode: 2 ] => 2 (792, 0,374)
                              AgeGr in [ 4 5 ] [ Mode: 2 ] (530)
                                    ProductL2 in [ "CALT" "CORTA" ] [ Mode: 2 ] => 2 (530, 0,387)
                              AgeGr in [ 1 ] [ Mode: 3 ] => 3 (0)
                        ProductL2 in [ "CUST" "W" ] [ Mode: 2 ] => 2 (3.061,029, 0,386)
            Addpr = 0 [ Mode: 2 ] (28.683,126)
                  Sex = K [ Mode: 2 ] (13.058,03)
                        NCD = 1 [ Mode: 2 ] (7.142,03)
                              AgeGr in [ 2 ] [ Mode: 2 ] (2.838)
                                    OccupGr in [ 0 1 3 4 5 6 7 8 9 10 11 12 13 14 ] [ Mode: 2 ] (2.690)
                                          PCI = 1 [ Mode: 2 ] => 2 (171, 0,304)
                                          PCI = 0 [ Mode: 2 ] (2.519)
                                                ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (140, 0,343)
                                                ProductL2 in [ "CORTA" "CUST" ] [ Mode: 2 ] (2.379)
                                                      Loyal in [ "LC" "NC" ] [ Mode: 2 ] => 2 (2.379, 0,308)
                                                ProductL2 in [ "W" ] [ Mode: 1 ] => 1 (0)
                                    OccupGr in [ 2 15 ] [ Mode: 3 ] => 3 (148, 0,52)
                              AgeGr in [ 3 4 5 ] [ Mode: 2 ] => 2 (4.304,03, 0,318)
                              AgeGr in [ 1 ] [ Mode: 1 ] => 1 (0)
                        NCD = 0 [ Mode: 2 ] (5.916)
                              ProductL2 in [ "CALT" ] [ Mode: 2 ] (358)
                                    AgeGr in [ 1 2 4 ] [ Mode: 2 ] => 2 (227, 0,485)
                                    AgeGr in [ 3 ] [ Mode: 2 ] (116)
                                          Loyal in [ "LC" ] [ Mode: 2 ] => 2 (67, 0,597)
                                          Loyal in [ "NC" ] [ Mode: 2 ] (49)
```

```
                                                      City in [ 6 34 ] [ Mode: 2 ] => 2 (47, 0,447)
                                                      City in [ 1 2 3 4 5 7 8 9 10 11 13 14 15 16 17 18 19 20 21 22
23 24 25 26 27 28 30 31 32 33 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 57 58 59 60 61 62 63
64 65 66 67 68 70 71 72 74 77 78 79 80 81 ] [ Mode: 1 ] => 1 (2, 0,5)
                                          AgeGr in [ 5 ] [ Mode: 1 ] => 1 (15, 0,533)
                                    ProductL2 in [ "CORTA" "CUST" "W" ] [ Mode: 2 ] => 2 (5.558, 0,368)
                              Sex = E [ Mode: 2 ] (15.625,096)
                                    NCD = 1 [ Mode: 1 ] => 1 (9.726,1, 0,457)
                                    NCD = 0 [ Mode: 2 ] (5.898,996)
                                          ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] (4.274,005)
                                                Loyal in [ "LC" ] [ Mode: 2 ] => 2 (2.971,004, 0,327)
                                                Loyal in [ "NC" ] [ Mode: 2 ] (1.303,001)
                                                      AgeGr in [ 2 3 4 ] [ Mode: 2 ] => 2 (1.296,001, 0,325)
                                                      AgeGr in [ 1 5 ] [ Mode: 1 ] => 1 (7, 0,571)
                                          ProductL2 in [ "CORTA" "W" ] [ Mode: 1 ] => 1 (1.624,991, 0,522)
                  Noins > 1 [ Mode: 2 ] (26.326,168)
                        AgeGr in [ 1 2 ] [ Mode: 3 ] => 3 (4.886,026, 0,525)
                        AgeGr in [ 3 4 5 ] [ Mode: 2 ] (21.440,142)
                              NCD = 1 [ Mode: 2 ] (8.846,002)
                                    City in [ 1 ] [ Mode: 2 ] (95,011)
                                          Addpr = 1 [ Mode: 1 ] => 1 (36, 0,528)
                                          Addpr = 0 [ Mode: 3 ] => 3 (59,011, 0,508)
                                    City in [ 3 20 27 31 38 39 42 44 59 61 68 71 74 ] [ Mode: 2 ] => 2 (187,021, 0,374)
                                    City in [ 6 ] [ Mode: 2 ] (727,082)
                                          AgeGr in [ 3 4 ] [ Mode: 2 ] (685,082)
                                                OccupGr in [ 0 ] [ Mode: 2 ] (286,082)
                                                      PCI = 1 [ Mode: 3 ] => 3 (40,082, 0,649)
                                                      PCI = 0 [ Mode: 2 ] (246)
                                                            AgeGr in [ 3 ] [ Mode: 2 ] (138)
                                                                  Sex = K [ Mode: 3 ] => 3 (68, 0,5)
                                                                  Sex = E [ Mode: 2 ] (70)
                                                                        ProductL2 in [ "CALT" "CORTA" "CUST" "W" ] [
Mode: 2 ] => 2 (70, 0,5)
                                                            AgeGr in [ 4 ] [ Mode: 2 ] => 2 (108, 0,435)
                                          OccupGr in [ 1 8 10 12 ] [ Mode: 2 ] => 2 (219, 0,425)
```

```
                  OccupGr in [ 2 ] [ Mode: 2 ] (31)
                        Addpr = 1 [ Mode: 2 ] => 2 (8, 0,875)
                        Addpr = 0 [ Mode: 2 ] (23)
                              Noins <= 2 [ Mode: 1 ] => 1 (11, 0,455)
                              Noins > 2 [ Mode: 2 ] (12)
                                    AgeGr in [ 3 ] [ Mode: 2 ] => 2 (7, 0,714)
                                    AgeGr in [ 4 ] [ Mode: 3 ] => 3 (5, 0,8)
                  OccupGr in [ 3 ] [ Mode: 2 ] (83)
                        Sex = K [ Mode: 2 ] => 2 (13, 0,769)
                        Sex = E [ Mode: 2 ] (70)
                              Channel in [ "A" ] [ Mode: 2 ] (37)
                                    Loyal in [ "LC" ] [ Mode: 1 ] => 1 (30, 0,5)
                                    Loyal in [ "NC" ] [ Mode: 3 ] => 3 (7, 0,571)
                              Channel in [ "B" "D" ] [ Mode: 2 ] => 2 (33, 0,636)
                  OccupGr in [ 4 ] [ Mode: 2 ] (29)
                        ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] (23)
                              Noins <= 3 [ Mode: 2 ] (16)
                                    AgeGr in [ 3 ] [ Mode: 2 ] (9)
                                          ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (5, 0,8)
                                          ProductL2 in [ "CUST" ] [ Mode: 1 ] => 1 (4, 0,5)
                                    AgeGr in [ 4 ] [ Mode: 1 ] => 1 (7, 0,714)
                              Noins > 3 [ Mode: 2 ] => 2 (7, 0,857)
                        ProductL2 in [ "CORTA" "W" ] [ Mode: 1 ] => 1 (6, 0,667)
                  OccupGr in [ 5 7 13 14 15 ] [ Mode: 3 ] => 3 (8, 0,75)
                  OccupGr in [ 6 ] [ Mode: 2 ] (16)
                        PCI = 1 [ Mode: 1 ] => 1 (4, 1,0)
                        PCI = 0 [ Mode: 2 ] => 2 (12, 0,75)
                  OccupGr in [ 9 ] [ Mode: 2 ] (11)
                        Sex = K [ Mode: 3 ] => 3 (5, 0,6)
                        Sex = E [ Mode: 1 ] => 1 (6, 0,5)
                  OccupGr in [ 11 ] [ Mode: 1 ] => 1 (2, 0,5)
            AgeGr in [ 5 ] [ Mode: 1 ] => 1 (42, 0,595)
      City in [ 7 ] [ Mode: 2 ] (268,03)
            AgeGr in [ 3 ] [ Mode: 2 ] (163)
                  PCI = 1 [ Mode: 3 ] => 3 (26, 0,5)
```

```
                                    PCI = 0 [ Mode: 2 ] (137)
                                          Noins <= 3 [ Mode: 2 ] => 2 (100, 0,35)
                                          Noins > 3 [ Mode: 2 ] => 2 (37, 0,541)
                                    AgeGr in [ 4 5 ] [ Mode: 2 ] => 2 (105,03, 0,381)
                              City in [ 9 ] [ Mode: 2 ] (75,008)
                                    AgeGr in [ 3 ] [ Mode: 2 ] => 2 (36, 0,389)
                                    AgeGr in [ 4 5 ] [ Mode: 1 ] => 1 (39,008, 0,487)
                              City in [ 10 17 18 21 32 46 47 50 51 52 63 64 78 81 ] [ Mode: 1 ] => 1 (116,013,
0,603)
                              City in [ 2 4 5 8 11 13 14 15 19 23 24 25 28 30 36 37 40 43 49 53 57 58 60 62 65 66
70 72 77 79 80 ] [ Mode: 3 ] => 3 (15,002, 0,667)
                              City in [ 16 ] [ Mode: 2 ] (280,032)
                                    ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (8, 0,625)
                                    ProductL2 in [ "CORTA" "CUST" ] [ Mode: 2 ] (272,032)
                                          Noins <= 4 [ Mode: 2 ] (258,032)
                                                OccupGr in [ 0 ] [ Mode: 2 ] (75,032)
                                                      Loyal in [ "LC" ] [ Mode: 2 ] (38,032)
                                                            Sex = K [ Mode: 3 ] => 3 (6, 0,833)
                                                            Sex = E [ Mode: 2 ] => 2 (32,032, 0,219)
                                                      Loyal in [ "NC" ] [ Mode: 2 ] (37)
                                                            Sex = K [ Mode: 2 ] => 2 (8, 0,5)
                                                            Sex = E [ Mode: 2 ] => 2 (29, 0,414)
                                                OccupGr in [ 1 ] [ Mode: 2 ] (76)
                                                      Channel in [ "A" ] [ Mode: 3 ] (23)
                                                            Loyal in [ "LC" ] [ Mode: 3 ] => 3 (12, 0,917)
                                                            Loyal in [ "NC" ] [ Mode: 2 ] => 2 (11, 0,364)
                                                      Channel in [ "B" "D" ] [ Mode: 2 ] => 2 (53, 0,34)
                                                OccupGr in [ 2 7 11 12 ] [ Mode: 2 ] => 2 (20, 0,55)
                                                OccupGr in [ 3 9 13 15 ] [ Mode: 1 ] => 1 (42, 0,619)
                                                OccupGr in [ 4 ] [ Mode: 2 ] (21)
                                                      Loyal in [ "LC" ] [ Mode: 2 ] => 2 (17, 0,588)
                                                      Loyal in [ "NC" ] [ Mode: 1 ] => 1 (4, 0,75)
                                                OccupGr in [ 5 10 14 ] [ Mode: 3 ] => 3 (3, 1,0)
                                                OccupGr in [ 6 ] [ Mode: 2 ] (10)
                                                      Noins <= 3 [ Mode: 3 ] => 3 (4, 0,75)
```

```
                        Noins > 3 [ Mode: 1 ] => 1 (6, 0,667)
                  OccupGr in [ 8 ] [ Mode: 3 ] (11)
                        PCI = 1 [ Mode: 3 ] => 3 (4, 1,0)
                        PCI = 0 [ Mode: 1 ] => 1 (7, 0,571)
            Noins > 4 [ Mode: 2 ] (14)
                  Loyal in [ "LC" ] [ Mode: 2 ] => 2 (7, 0,857)
                  Loyal in [ "NC" ] [ Mode: 1 ] => 1 (7, 0,571)
      ProductL2 in [ "W" ] [ Mode: 1 ] => 1 (0)
City in [ 22 ] [ Mode: 2 ] (11,001)
      Noins <= 2 [ Mode: 1 ] => 1 (5,001, 0,8)
      Noins > 2 [ Mode: 2 ] => 2 (6, 0,667)
City in [ 26 ] [ Mode: 2 ] (34,004)
      OccupGr in [ 0 3 10 12 ] [ Mode: 2 ] => 2 (12,004, 0,5)
      OccupGr in [ 1 ] [ Mode: 1 ] (19)
            Channel in [ "A" ] [ Mode: 3 ] => 3 (4, 0,5)
            Channel in [ "B" ] [ Mode: 1 ] => 1 (14, 0,786)
            Channel in [ "D" ] [ Mode: 2 ] => 2 (1, 1,0)
      OccupGr in [ 2 ] [ Mode: 3 ] => 3 (2, 1,0)
      OccupGr in [ 4 5 6 7 8 9 11 13 14 15 ] [ Mode: 1 ] => 1 (1, 1,0)
City in [ 33 ] [ Mode: 2 ] (93,011)
      OccupGr in [ 0 ] [ Mode: 2 ] (38,011)
            PCI = 1 [ Mode: 1 ] => 1 (6,011, 0,667)
            PCI = 0 [ Mode: 2 ] (32)
                  Addpr = 1 [ Mode: 2 ] (16)
                        AgeGr in [ 3 ] [ Mode: 1 ] => 1 (5, 0,6)
                        AgeGr in [ 4 ] [ Mode: 2 ] (10)
                              Noins <= 3 [ Mode: 1 ] => 1 (6, 0,667)
                              Noins > 3 [ Mode: 3 ] => 3 (4, 0,75)
                        AgeGr in [ 5 ] [ Mode: 3 ] => 3 (1, 1,0)
                  Addpr = 0 [ Mode: 3 ] (16)
                        Noins <= 2 [ Mode: 2 ] => 2 (10, 0,6)
                        Noins > 2 [ Mode: 3 ] => 3 (6, 0,833)
      OccupGr in [ 1 ] [ Mode: 2 ] (24)
            Addpr = 1 [ Mode: 1 ] => 1 (5, 0,6)
            Addpr = 0 [ Mode: 2 ] (19)
```

```
                        AgeGr in [ 3 ] [ Mode: 2 ] => 2 (9, 0,667)
                        AgeGr in [ 4 5 ] [ Mode: 3 ] => 3 (10, 0,7)
                  OccupGr in [ 2 4 5 6 ] [ Mode: 2 ] => 2 (14, 0,714)
                  OccupGr in [ 3 9 12 ] [ Mode: 1 ] => 1 (10, 0,6)
                  OccupGr in [ 7 8 10 11 13 14 15 ] [ Mode: 3 ] => 3 (7, 0,571)
            City in [ 34 ] [ Mode: 2 ] (5.289,599)
                  ProductL2 in [ "CALT" ] [ Mode: 2 ] (121)
                        AgeGr in [ 3 5 ] [ Mode: 3 ] => 3 (81, 0,494)
                        AgeGr in [ 4 ] [ Mode: 2 ] (40)
                              OccupGr in [ 0 2 3 5 7 9 10 11 12 14 ] [ Mode: 2 ] => 2 (19, 0,737)
                              OccupGr in [ 1 ] [ Mode: 2 ] (13)
                                    Noins <= 3 [ Mode: 3 ] => 3 (8, 0,625)
                                    Noins > 3 [ Mode: 2 ] => 2 (5, 1,0)
                              OccupGr in [ 4 6 8 ] [ Mode: 1 ] => 1 (5, 0,6)
                              OccupGr in [ 13 15 ] [ Mode: 3 ] => 3 (3, 1,0)
                  ProductL2 in [ "CORTA" "CUST" ] [ Mode: 2 ] (5.165,599)
                        AgeGr in [ 3 ] [ Mode: 2 ] (2.320,001)
                              Addpr = 1 [ Mode: 2 ] (990)
                                    ProductL2 in [ "CORTA" ] [ Mode: 1 ] => 1 (13, 0,769)
                                    ProductL2 in [ "CUST" ] [ Mode: 2 ] (977)
                                          OccupGr in [ 0 ] [ Mode: 2 ] (312)
                                                Channel in [ "A" "B" ] [ Mode: 2 ] (265)
                                                      Channel in [ "A" ] [ Mode: 2 ] (241)
                                                            Loyal in [ "LC" ] [ Mode: 2 ] (203)
                                                                  PCI = 1 [ Mode: 1 ] => 1 (32,
0,469)

                                                                  PCI = 0 [ Mode: 2 ] (171)
                                                                        Noins <= 3 [ Mode: 2 ]
(118)

                                                                              Noins <= 2 [ Mode: 2
] => 2 (54, 0,333)

                                                                              Noins > 2 [ Mode: 2
] => 2 (64, 0,266)

                                                                        Noins > 3 [ Mode: 2 ] => 2
(53, 0,528)
```

```
(38, 0,526)

0,404)

0,42)

0,385)

0,571)
```

```
                  Loyal in [ "NC" ] [ Mode: 2 ] => 2

            Channel in [ "B" ] [ Mode: 1 ] (24)
                  Sex = K [ Mode: 1 ] => 1 (12, 0,5)
                  Sex = E [ Mode: 3 ] => 3 (12, 0,583)
      Channel in [ "D" ] [ Mode: 2 ] (47)
            Sex = K [ Mode: 2 ] => 2 (24, 0,583)
            Sex = E [ Mode: 3 ] => 3 (23, 0,609)
OccupGr in [ 1 ] [ Mode: 2 ] (235)
      Loyal in [ "LC" "NC" ] [ Mode: 2 ] => 2 (235,

OccupGr in [ 2 ] [ Mode: 2 ] (112)
      Loyal in [ "LC" "NC" ] [ Mode: 2 ] => 2 (112,

OccupGr in [ 3 5 8 13 15 ] [ Mode: 2 ] => 2 (146, 0,39)
OccupGr in [ 4 ] [ Mode: 2 ] (78)
      Channel in [ "A" ] [ Mode: 2 ] => 2 (56, 0,464)
      Channel in [ "B" ] [ Mode: 3 ] => 3 (4, 0,5)
      Channel in [ "D" ] [ Mode: 2 ] (18)
            Noins <= 3 [ Mode: 2 ] => 2 (11, 0,636)
            Noins > 3 [ Mode: 1 ] => 1 (7, 0,714)
OccupGr in [ 6 ] [ Mode: 2 ] (24)
      Channel in [ "A" ] [ Mode: 3 ] => 3 (14, 0,571)
      Channel in [ "B" "D" ] [ Mode: 2 ] => 2 (10, 0,8)
OccupGr in [ 7 12 14 ] [ Mode: 1 ] => 1 (17, 0,529)
OccupGr in [ 9 ] [ Mode: 2 ] (33)
      Noins <= 2 [ Mode: 1 ] => 1 (13, 0,615)
      Noins > 2 [ Mode: 2 ] (20)
            Loyal in [ "LC" ] [ Mode: 2 ] => 2 (13,

            Loyal in [ "NC" ] [ Mode: 1 ] => 1 (7,

OccupGr in [ 10 ] [ Mode: 3 ] => 3 (7, 0,571)
OccupGr in [ 11 ] [ Mode: 2 ] (13)
      Noins <= 3 [ Mode: 3 ] => 3 (8, 0,625)
```

```
                        Noins > 3 [ Mode: 2 ] => 2 (5, 0,6)
      Addpr = 0 [ Mode: 2 ] => 2 (1.330,001, 0,32)
AgeGr in [ 4 ] [ Mode: 2 ] (2.538,598)
      OccupGr in [ 0 ] [ Mode: 2 ] (824,598)
            Noins <= 2 [ Mode: 2 ] => 2 (290,598, 0,32)
            Noins > 2 [ Mode: 2 ] (534)
                  Noins <= 4 [ Mode: 2 ] => 2 (514, 0,298)
                  Noins > 4 [ Mode: 2 ] (20)
                        PCI = 1 [ Mode: 1 ] => 1 (4, 1,0)
                        PCI = 0 [ Mode: 2 ] => 2 (16, 0,625)
      OccupGr in [ 1 ] [ Mode: 2 ] (670)
            Addpr = 1 [ Mode: 2 ] (264)
                  ProductL2 in [ "CORTA" ] [ Mode: 1 ] => 1 (6, 0,833)
                  ProductL2 in [ "CUST" ] [ Mode: 2 ] (258)
                        Channel in [ "A" ] [ Mode: 3 ] => 3 (147, 0,456)
                        Channel in [ "B" "D" ] [ Mode: 2 ] => 2 (111,
0,324)

            Addpr = 0 [ Mode: 2 ] (406)
                  Noins <= 4 [ Mode: 2 ] (386)
                        Channel in [ "A" ] [ Mode: 2 ] (219)
                              PCI = 1 [ Mode: 2 ] (57)
                                    ProductL2 in [ "CORTA" ] [ Mode: 2 ]

=> 2 (21, 0,381)
                                    ProductL2 in [ "CUST" ] [ Mode: 2 ]

(36)
                                          Noins <= 3 [ Mode: 1 ] => 1 (22,

0,545)
                                          Noins > 3 [ Mode: 2 ] => 2 (14,

0,5)
                              PCI = 0 [ Mode: 2 ] => 2 (162, 0,309)
                        Channel in [ "B" ] [ Mode: 2 ] (60)
                              Noins <= 3 [ Mode: 2 ] => 2 (35, 0,343)
                              Noins > 3 [ Mode: 2 ] => 2 (25, 0,44)
                        Channel in [ "D" ] [ Mode: 2 ] => 2 (107, 0,458)
                  Noins > 4 [ Mode: 2 ] (20)
```

```
                  PCI = 1 [ Mode: 1 ] => 1 (6, 0,667)
                  PCI = 0 [ Mode: 2 ] => 2 (14, 0,714)
OccupGr in [ 2 ] [ Mode: 2 ] (156)
      Noins <= 2 [ Mode: 2 ] => 2 (108, 0,324)
      Noins > 2 [ Mode: 2 ] (48)
            Addpr = 1 [ Mode: 2 ] => 2 (17, 0,471)
            Addpr = 0 [ Mode: 2 ] => 2 (31, 0,645)
OccupGr in [ 3 ] [ Mode: 2 ] (214)
      Noins <= 2 [ Mode: 1 ] => 1 (78, 0,5)
      Noins > 2 [ Mode: 2 ] (136)
            Noins <= 3 [ Mode: 2 ] (80)
                  Channel in [ "A" ] [ Mode: 2 ] (55)
                        Addpr = 1 [ Mode: 2 ] => 2 (24, 0,333)
                        Addpr = 0 [ Mode: 3 ] (31)
                              ProductL2 in [ "CORTA" ] [ Mode: 2 ]

                                    Loyal in [ "LC" ] [ Mode: 2 ] =>

                                    Loyal in [ "NC" ] [ Mode: 3 ] =>

                              ProductL2 in [ "CUST" ] [ Mode: 3 ] =>

                  Channel in [ "B" ] [ Mode: 1 ] => 1 (12, 0,583)
                  Channel in [ "D" ] [ Mode: 2 ] => 2 (13, 0,615)
            Noins > 3 [ Mode: 3 ] (56)
                  Loyal in [ "LC" ] [ Mode: 3 ] => 3 (50, 0,52)
                  Loyal in [ "NC" ] [ Mode: 1 ] => 1 (6, 0,667)
OccupGr in [ 4 ] [ Mode: 2 ] (271)
      Addpr = 1 [ Mode: 2 ] => 2 (113, 0,274)
      Addpr = 0 [ Mode: 2 ] => 2 (158, 0,329)
OccupGr in [ 5 ] [ Mode: 2 ] (49)
      Channel in [ "A" "D" ] [ Mode: 2 ] (42)
            ProductL2 in [ "CORTA" ] [ Mode: 3 ] => 3 (14, 0,643)
            ProductL2 in [ "CUST" ] [ Mode: 2 ] (28)
                  Sex = K [ Mode: 2 ] => 2 (13, 0,308)
```

```
(16)

2 (9, 0,667)

3 (7, 0,571)

3 (15, 0,667)
```

```
                                    Sex = E [ Mode: 1 ] => 1 (15, 0,533)
                              Channel in [ "B" ] [ Mode: 2 ] => 2 (7, 0,714)
                        OccupGr in [ 6 7 11 14 ] [ Mode: 2 ] => 2 (97, 0,412)
                        OccupGr in [ 8 ] [ Mode: 2 ] (112)
                              Sex = K [ Mode: 1 ] => 1 (5, 0,6)
                              Sex = E [ Mode: 2 ] => 2 (107, 0,346)
                        OccupGr in [ 9 15 ] [ Mode: 1 ] => 1 (70, 0,486)
                        OccupGr in [ 10 ] [ Mode: 2 ] (21)
                              Channel in [ "A" "B" ] [ Mode: 2 ] (16)
                                    Addpr = 1 [ Mode: 2 ] => 2 (9, 0,222)
                                    Addpr = 0 [ Mode: 2 ] => 2 (7, 0,714)
                              Channel in [ "D" ] [ Mode: 2 ] => 2 (5, 1,0)
                        OccupGr in [ 12 ] [ Mode: 1 ] (28)
                              Loyal in [ "LC" ] [ Mode: 1 ] (21)
                                    Noins <= 2 [ Mode: 1 ] => 1 (10, 0,7)
                                    Noins > 2 [ Mode: 3 ] => 3 (11, 0,636)
                              Loyal in [ "NC" ] [ Mode: 1 ] => 1 (7, 0,857)
                        OccupGr in [ 13 ] [ Mode: 3 ] => 3 (26, 0,5)
                  AgeGr in [ 5 ] [ Mode: 2 ] (307)
                        Noins <= 3 [ Mode: 2 ] => 2 (295, 0,264)
                        Noins > 3 [ Mode: 1 ] => 1 (12, 0,583)
            ProductL2 in [ "W" ] [ Mode: 1 ] => 1 (3, 0,667)
City in [ 35 ] [ Mode: 2 ] (1.225,14)
            ProductL2 in [ "CALT" ] [ Mode: 2 ] => 2 (35, 0,486)
            ProductL2 in [ "CORTA" "CUST" "W" ] [ Mode: 2 ] (1.190,14)
                  AgeGr in [ 3 4 ] [ Mode: 2 ] (1.123,14)
                        ProductL2 in [ "CORTA" ] [ Mode: 2 ] (292)
                              Noins <= 3 [ Mode: 2 ] (209)
                                    Sex = K [ Mode: 2 ] => 2 (83, 0,289)
                                    Sex = E [ Mode: 1 ] => 1 (126, 0,468)
                              Noins > 3 [ Mode: 2 ] => 2 (83, 0,253)
                        ProductL2 in [ "CUST" ] [ Mode: 2 ] => 2 (830,14, 0,301)
                        ProductL2 in [ "W" ] [ Mode: 1 ] => 1 (1, 1,0)
                  AgeGr in [ 5 ] [ Mode: 3 ] => 3 (67, 0,478)
City in [ 41 ] [ Mode: 2 ] (90,01)
```

```
      ProductL2 in [ "CALT" ] [ Mode: 1 ] => 1 (2, 0,5)
      ProductL2 in [ "CORTA" ] [ Mode: 2 ] (16)
            AgeGr in [ 3 ] [ Mode: 2 ] => 2 (8, 0,75)
            AgeGr in [ 4 5 ] [ Mode: 1 ] => 1 (8, 0,75)
      ProductL2 in [ "CUST" ] [ Mode: 2 ] (72,01)
            Loyal in [ "LC" ] [ Mode: 2 ] => 2 (46,01, 0,369)
            Loyal in [ "NC" ] [ Mode: 2 ] (26)
                  AgeGr in [ 3 ] [ Mode: 3 ] => 3 (13, 0,615)
                  AgeGr in [ 4 5 ] [ Mode: 1 ] => 1 (13, 0,615)
      ProductL2 in [ "W" ] [ Mode: 3 ] => 3 (0)
City in [ 45 ] [ Mode: 2 ] (15,002)
      Channel in [ "A" ] [ Mode: 3 ] => 3 (5,002, 0,6)
      Channel in [ "B" ] [ Mode: 2 ] => 2 (7, 0,143)
      Channel in [ "D" ] [ Mode: 1 ] => 1 (3, 1,0)
City in [ 48 ] [ Mode: 2 ] (230,026)
      AgeGr in [ 3 ] [ Mode: 2 ] => 2 (122, 0,352)
      AgeGr in [ 4 ] [ Mode: 1 ] (100,026)
            ProductL2 in [ "CALT" "CORTA" "W" ] [ Mode: 1 ] => 1 (26, 0,615)
            ProductL2 in [ "CUST" ] [ Mode: 2 ] (74,026)
                  Noins <= 3 [ Mode: 1 ] => 1 (46,026, 0,5)
                  Noins > 3 [ Mode: 3 ] => 3 (28, 0,571)
      AgeGr in [ 5 ] [ Mode: 1 ] => 1 (8, 0,625)
City in [ 54 ] [ Mode: 2 ] (67,008)
      OccupGr in [ 0 ] [ Mode: 2 ] (14,008)
            AgeGr in [ 3 ] [ Mode: 2 ] (9)
                  Channel in [ "A" ] [ Mode: 2 ] => 2 (5, 0,6)
                  Channel in [ "B" ] [ Mode: 1 ] => 1 (4, 0,5)
                  Channel in [ "D" ] [ Mode: 3 ] => 3 (0)
            AgeGr in [ 4 5 ] [ Mode: 2 ] => 2 (5,008, 0,998)
      OccupGr in [ 1 ] [ Mode: 2 ] (32)
            Noins <= 4 [ Mode: 1 ] => 1 (26, 0,5)
            Noins > 4 [ Mode: 3 ] => 3 (6, 0,667)
      OccupGr in [ 3 ] [ Mode: 3 ] => 3 (4, 0,75)
      OccupGr in [ 4 13 ] [ Mode: 2 ] => 2 (5, 0,4)
      OccupGr in [ 2 5 6 7 9 10 11 12 14 15 ] [ Mode: 1 ] => 1 (3, 1,0)
```

```
                              OccupGr in [ 8 ] [ Mode: 1 ] (9)
                                    Loyal in [ "LC" ] [ Mode: 2 ] => 2 (5, 0,4)
                                    Loyal in [ "NC" ] [ Mode: 1 ] => 1 (4, 1,0)
                        City in [ 55 ] [ Mode: 2 ] (10,001)
                              Noins <= 2 [ Mode: 1 ] => 1 (4,001, 0,75)
                              Noins > 2 [ Mode: 3 ] => 3 (6, 0,5)
                        City in [ 67 ] [ Mode: 2 ] (18,002)
                              Addpr = 1 [ Mode: 1 ] => 1 (8, 0,625)
                              Addpr = 0 [ Mode: 3 ] => 3 (10,002, 0,5)
                  NCD = 0 [ Mode: 2 ] (12.594,14)
                        ProductL2 in [ "CALT" "CUST" ] [ Mode: 2 ] => 2 (8.735,14, 0,426)
                        ProductL2 in [ "CORTA" "W" ] [ Mode: 2 ] (3.859)
                              AgeGr in [ 3 ] [ Mode: 2 ] (2.078,999)
                                    ProductL2 in [ "CORTA" ] [ Mode: 2 ] => 2 (2.029,999, 0,305)
                                    ProductL2 in [ "W" ] [ Mode: 1 ] (49)
                                          PCI = 1 [ Mode: 1 ] => 1 (15, 0,733)
                                          PCI = 0 [ Mode: 3 ] => 3 (34, 0,5)
                              AgeGr in [ 4 ] [ Mode: 2 ] (1.554,001)
                                    Noins <= 3 [ Mode: 2 ] => 2 (908,001, 0,319)
                                    Noins > 3 [ Mode: 2 ] (646)
                                          Loyal in [ "LC" "NC" ] [ Mode: 2 ] => 2 (646, 0,283)
                              AgeGr in [ 5 ] [ Mode: 2 ] (226)
                                    Sex = K [ Mode: 2 ] => 2 (13, 0,308)
                                    Sex = E [ Mode: 2 ] (213)
                                          ProductL2 in [ "CORTA" ] [ Mode: 2 ] => 2 (172, 0,244)
                                          ProductL2 in [ "W" ] [ Mode: 3 ] => 3 (41, 0,61)
Loyal in [ "N" ] [ Mode: 2 ] (14.414,169)
      ProductL2 in [ "CALT" ] [ Mode: 2 ] (3.686,999)
            AgeGr in [ 1 2 3 5 ] [ Mode: 2 ] (3.334,999)
                  Noins <= 1 [ Mode: 2 ] (2.092,998)
                        OccupGr in [ 0 1 2 3 4 5 6 7 8 9 10 11 12 14 ] [ Mode: 2 ] (2.020,998)
                              AgeGr in [ 1 ] [ Mode: 2 ] => 2 (18, 0,556)
                              AgeGr in [ 2 ] [ Mode: 2 ] (1.456,998)
                                    NCD = 1 [ Mode: 1 ] => 1 (6, 0,667)
                                    NCD = 0 [ Mode: 2 ] (1.450,998)
```

```
                                    Channel in [ "A" "B" "D" ] [ Mode: 2 ] => 2 (1.450,998, 0,356)
                              AgeGr in [ 3 ] [ Mode: 2 ] (543)
                                    Sex = K [ Mode: 2 ] => 2 (187, 0,444)
                                    Sex = E [ Mode: 2 ] => 2 (356, 0,303)
                              AgeGr in [ 5 ] [ Mode: 1 ] => 1 (3, 0,667)
                        OccupGr in [ 13 15 ] [ Mode: 1 ] (72)
                              Channel in [ "A" "B" ] [ Mode: 1 ] => 1 (51, 0,882)
                              Channel in [ "D" ] [ Mode: 2 ] => 2 (21, 0,381)
                  Noins > 1 [ Mode: 2 ] (1.242,001)
                        Addpr = 1 [ Mode: 2 ] (1.047,001)
                              NCD = 1 [ Mode: 3 ] => 3 (5, 0,8)
                              NCD = 0 [ Mode: 2 ] (1.042,001)
                                    Noins <= 2 [ Mode: 3 ] => 3 (498,001, 0,494)
                                    Noins > 2 [ Mode: 2 ] (544)
                                          AgeGr in [ 1 2 3 5 ] [ Mode: 2 ] => 2 (544, 0,472)
                        Addpr = 0 [ Mode: 2 ] => 2 (195, 0,518)
            AgeGr in [ 4 ] [ Mode: 2 ] (352)
                  Noins <= 3 [ Mode: 2 ] (286)
                        Sex = K [ Mode: 2 ] => 2 (108, 0,556)
                        Sex = E [ Mode: 1 ] => 1 (178, 0,517)
                  Noins > 3 [ Mode: 2 ] (66)
                        OccupGr in [ 0 1 5 6 8 14 ] [ Mode: 2 ] => 2 (53, 0,698)
                        OccupGr in [ 2 3 4 7 9 10 11 12 13 15 ] [ Mode: 1 ] => 1 (13, 0,538)
ProductL2 in [ "CORTA" "CUST" "W" ] [ Mode: 1 ] => 1 (10.727,17, 0,475)
```

RESUME

Name: Emre Tacenur

Birth: 16.05.1974, Eskişehir

Address: Prof.Hıfzı Özcan Cad. Kardelen Sitesi A Blok D:36
K.Bakkalköy/İSTANBUL

Phone: 0 (216) 572 43 24

e-mail: emretacenur@gmail.com

Marital Status: Married

Military Service: Completed (2000-2001)

Education

1997 - 1992 BSc. Industrial Engineering / Middle East Technical University

1992 - 1985 Eskişehir Anatolian High School

Experience

- 2001 Systems Research and Development Specialist

Business and Process Improvement, Business Systems Analysis, Project Management and Planning.

2000 - 1997 Purchasing Specialist / Arçelik

Supply Chain Management, Cost Analysis.

1997 - 1997 Production Planning and MIS Engineer (Part-time) / Aselsan

Reporting system design